

Junping Qiu · Rongying Zhao
Siluo Yang · Ke Dong

Informetrics

Theory, Methods and Applications



Springer

Informetrics

Junping Qiu · Rongying Zhao
Siluo Yang · Ke Dong

Informetrics

Theory, Methods and Applications



Springer

Junping Qiu
School of Information Management
Wuhan University
Wuhan, Hubei
China

Rongying Zhao
School of Information Management
Wuhan University
Wuhan, Hubei
China

Siluo Yang
School of Information Management
Wuhan University
Wuhan, Hubei
China

Ke Dong
School of Information Management
Wuhan University
Wuhan, Hubei
China

ISBN 978-981-10-4031-3
DOI 10.1007/978-981-10-4032-0

ISBN 978-981-10-4032-0 (eBook)

Library of Congress Control Number: 2017936334

© Springer Nature Singapore Pte Ltd. 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Since 1960s, three similar terms: Bibliometrics, Scientometrics, and Informetrics have appeared in the fields of library science, philology, information science, and science of science. These similar quantitative branches are called three-metrics. After decades of efforts on research and promotion, these disciplines all progressed at different degrees and became widely recognized by academia. Although these areas have different research objects and purposes, they have the same origin and share common principles, methods, and tools. Therefore, academia refers to these subjects as three-metrics. With development of science and technology and continuation of these three metrologies, convergence among them has developed, and International Society for Scientometrics and Informetrics (ISSI) was formed. Since 1990s, with rapid development and popularization of computer and network technologies and rise of knowledge economy and knowledge management, digitalization, networking, and knowledge have become remarkable characteristics of information society and knowledge economy era. Three-metrics is characterized by expanding breadth and depth of studies. Webometrics is based on network information and data, and Knowledgometrics is based on knowledge units; these subjects emerged in field of information management, prompting people to coin the term five-metrics. Five-metrics include literature, data, information (including network information), knowledge, and scientific activities. Subjects share many similarities but also have significant differences; they became important works on measurement research in information management field. Development of five-metrics reflects continuous innovation of quantitative research on information management and tracking of evolution with changing times and social background. Five-metrics also only involves legacy and development of bibliometrics and scientometrics.

Informetrics uses quantitative methods to describe and research phenomena, processes, and laws of information. This area is a new quantitative discipline of information science, and it is based on mathematics and statistics. Informetrics was initially presented as the German word “*Informetrie*,” which was proposed by German scholar Otto Nacke. Corresponding English term “*Informetrics*” soon appeared in subsequent literature works. Nacke expanded concept of informetrics on first Seminar on Informetrics (including Scientometrics) in Frankfurt in September

1980. German and English terms also appeared in Chinese journals in 1981. Informetrics did not only spread rapidly in English-speaking countries but was also recognized by International Federation for Information and Documentation (FID), marking the rise of a new branch of discipline. As early as 1980, FID established informetric communications (FID/IM). In 1987, Belgium held the first International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval. The well-known information scientist, Brookes, suggested at the meeting that term informetrics should be added to name of second International Academic Conference, which would be held in Canada in 1989; participants were generally supportive. However, conference was not renamed until June 1995 on Fifth International Conference on Scientometrics and Informetrics held in Chicago, USA. Informetrics was replaced with Bibliometrics in conference name. At present, the conference is known as ISSI. Given that “Informetrics” has been used in the titles of numerous proceedings published by international academic conferences since 1987, a number of well-known foreign information scientists regard 1987 as time when informetrics was recognized formally by international information academia. Chinese academic community responded accordingly to Informetrie (German) and Informetrics (English) and the disciplines they represented and introduced. As early as 1981, related papers were published. My monograph < Bibliometrics(Chinese) >, published in 1988, not only discussed in detail relationship among three-metrics but also proposed systematic framework of informetrics.

Our team has been teaching and researching on bibliometrics, informetrics, and scientometrics at Wuhan University since early 1980s. We led in offering Bibliometrics course in Chinese colleges in 1983 and compiled Chinese teaching material under the same name. Materials were published officially by Scientific and Technical Documentation Press (Beijing) in 1988 after being featured in mimeograph in 1983, letterpress in 1985, and few years of teaching. This book was the first to comprehensively structure content system of bibliometrics from the perspective of theory, method, and application, and it was praised and welcomed by academic community. This textbook is used in more than 10 colleges and universities, and its citation rate is among the best. Yang Peiting, a famous information scientist, said, “This is undoubtedly a positive contribution to the study and teaching of Information Science in China, and this can be said to be a ground-breaking research.” Afterward, our team carried out studies on three-metrics, which significantly influence people both in local areas and abroad. With rapid development of information technology and information science and with popularization of information resources, electronic, digital, and network, information resources are becoming more popular. Information resources greatly influenced and resulted in profound changes in development of human society, economy, science and technology, culture, and other fields. Under new social environment and technical conditions, new developments transpired in bibliometrics research. Facing this new situation, trends, and topic, our team led development of informetric and webo-metric research in China and published series of research papers with “Informetrics” and “Webometrics” on their titles in Information Studies: Theory and Application in 2000–2001; these publications had great repercussion and high

rate of citation in academia, locally and abroad. These papers became classic series of articles in research of informetrics and webometrics. *<Informetrics (Chinese)>* by Qiu Junping was published by Wuhan University Press in January 2007. This book was crystallization of long-term teaching and research on three-metrics and reflected development characteristics of three-metrics in information age. The material-oriented teaching and research on library science, information science and information management, and other related disciplines are included in “Ministry of Education for the 21st century curriculum materials” and “information management college and university core course textbook.” The book was selected as part of national quality courses and national Twelve Five planning materials. In recent years, we focused on trends in metrology research, undertook series of research projects, such as national social science major and the National Science Fund Project, and published series of research results. On this basis, under Science Press, we published “Metrology Research Series in Information Science,” which included *<Scientometrics>*, *<Knowledgometrics>*, *<Webometrics>*, and other related monographs in Chinese version.

With development of social economy and science and technology in China, research on information metrology rapidly progressed. China hosted ISSI meeting in Beijing (2003) and Wuhan (2017) and many other relevant meetings related to informetrics. Our country and other nations, such as the United States and some European countries, developed exchanges and cooperation. We published numerous related works and set up corresponding university courses and direction of graduate education. Many professional students and scholars go abroad to pursue degrees. Research institutions on informetrics also emerged; some of these organizations include National Professional Association and Chinese Society for Scientometrics and Informetrics. In China, informetrics adheres to dual development principle of internationalization and localization, with both closely following pace of foreign countries and having their own characteristics. Under the guidance of “bringing in and going out” strategy, Chinese scholars played increasingly important role in international professional organizations and extensively absorbed achievements of foreign professional treatises. These academics published more academic papers in foreign informetric professional journals with increasing influence. However, owing to the influence of policy orientation and language, international publications inadequately released relevant works of Chinese professionals. To introduce informetric research and to teach contents with Chinese characteristics, we published *<Informetrics—Theory, Methods and Applications>* in Springer-Verlag. We believe that publication of this book will provide basis for foreign countries to understand informetric research in China and will promote further development in research and practice of informetrics.

We always believed in close linkages and differences among bibliometrics, informetrics, and scientometrics. These disciplines have more interconnections, cross-connection, and overlapping than differences. Some foreigners regarded such areas as synonymous or advocated to use different names of subjects in different situations. We used the title “*Informetrics—Theory, Methods and Applications*” with following considerations in mind: first consideration is wider scope of

informetrics, which may include bibliometrics and scientometrics and also appeared in the name of ISSI meeting. Second consideration is that informetrics is based on research and is also recognized as branch of academic discipline. Third, with in-depth application of computer network technology, rise of Web 2.0 and big data technology, and popularization of open access and digital publishing, popularity of social networking and We-Media and rapid development in 4G mobile services and e-research profoundly influenced all aspects of information communication and technology innovation and provided necessary conditions and possibilities for information metrology and rare opportunity for development of informetrics. Fourth consideration is to arouse interest of people and to promote further research and development of informetrics. Future research and development should focus on informetrics. In this book, three-metrics is inevitably involved as basis of bibliometrics, and focus of discussion is metrology problem of literature information; this problem is influenced by present research situation and facts. To facilitate narrative, we also used “Informetrics” in the book.

The book consists of 11 chapters. Main content can be summarized into theory, method, and application. Research on theory of informetrics is found in Chaps. 1–7. Research on informetrics method spans Chaps. 8–9. Chapters 10–11 discuss application of informetrics. Law applications are also discussed in some chapters. This book retains some of typical application examples in < Bibliometrics > because they are classic cases and can still explain the problem; representative new cases cannot replace them. Though content is not updated, novelty of the book should not be affected. During compiling, we attempted to construct disciplinary system of informetrics from angles of theory, method, and application; attention was provided to combination of theory and practice, inheritance, and innovation; traditional statistical tools were combined with new information technology methods; no effort was exerted to ensure clear thinking, reasonable structure, comprehensive explanation, rich content, novel idea, and detailed material for this book. Material should not only reflect and absorb latest development on three-metrics both at home and abroad but also add our research results to make study of included disciplines more scientific, innovative, systemic, and practical. The book is suitable as teaching material in information management and information system, management science, information resource management, e-commerce, information science, library science, archives science, publishing science, science of science and management of S.&T., and evaluation and prediction of science in colleges. This work also serves as learning reference for majority of information workers, knowledge workers, researchers, evaluators, and managers.

Qiu Junping chaired revision of the book. The following people participated in revision and translation work: Qiu Junping, Zhao Rongying, Yang Siluo, Dong Ke, Tan Chunhui, Ma Ruimin, Ding Jingda, Song Yanhui, Zhang xinyuan, Yang Jinli, and Yuan Qingli. Finally, Qiu Junping and Yang Siluo made some additions, deletions or modifications, and completed English proofreading and drafting works. This book is legacy and innovation of < Informetrics > (Chinese version), < Bibliometrics > (Chinese version), < Scientometrics > (Chinese version), and

<Knowledgometrics> (Chinese version), and is completed based on <Informetrics> (Chinese version) with modification, supplementation, updates, and expansion.

Given that chapters of this book were written separately by different authors, mistakes may inevitably exist. We sincerely ask readers for criticisms and corrections.

Wuhan, China

Junping Qiu

Contents

1	Introduction	1
1.1	Origin and Development of Informetrics	1
1.1.1	Origin of Informetrics	1
1.1.2	Background of Informetrics	3
1.1.3	Development of Informetrics	4
1.2	Concept and System Structure of Informetrics	12
1.2.1	Aim and Significance of Informetric Research	12
1.2.2	Research Object of Informetrics	14
1.2.3	Concept of Informetrics	14
1.2.4	Content Structure of Informetrics	16
1.3	Tools and Methods for Informetric Research	17
1.3.1	Data Sources for Informetrics	17
1.3.2	Tools and Application Software for Informetrics	19
1.3.3	Method Systems for Informetrics	23
1.4	Informetrics and Related Disciplines	24
1.4.1	Related Disciplines of Informetrics	24
1.4.2	Relation Among Informetrics, Mathematics, and Statistics	25
1.4.3	Relation Between Informetrics and Bibliometrics	27
1.4.4	Relation Between Informetrics and Scientometrics	28
1.4.5	Relation Between Informetrics and Webmetrics	29
1.4.6	Relation Between Informetrics and Scientific Evaluation	29

2 Literature Information Growth Law	31
2.1 Characteristics of Literature Information Flow and Meaning of Growth Law	31
2.1.1 Characteristics of Literature Information Flow	31
2.1.2 Influence of Literature Information Growth and Countermeasures	32
2.1.3 Research on and Significance of Literature Information Growth Laws	34
2.2 Growth of Science Knowledge and Scientific Literature	35
2.2.1 Growth of Scientific Knowledge	35
2.2.2 Relationship Between the Growth of Scientific Knowledge and the Growth of Scientific Literature	38
2.3 Exponential Law of Literature Information	39
2.3.1 Indicators and Methods of Literature Information Measure	39
2.3.2 Literature Information Index Growth Model	40
2.3.3 Analysis of the Literature Index Growth Law	42
2.4 Law of Literature Information Logic Growth	44
2.4.1 Literature Information Logic Growth Model	44
2.4.2 Analysis of the Law of Literature Information Logic Growth	45
2.4.3 Modification of the Model for Literature Information Logical Growth	47
2.5 Other Mathematical Models for Literature Information Growth	48
2.5.1 Linear Growth Model	48
2.5.2 Hierarchical Sliding Exponential Model	49
2.5.3 Transcendental Function Model	51
2.5.4 Шестопал–Бурман Growth Model	52
2.6 Analysis of the Mechanism of Literature Information Growth	53
2.6.1 Reason for Literature Information Growth	53
2.6.2 Explanation for the Literature Information Growth Law	56
2.7 Applications of Literature Information Growth Laws	58
2.7.1 Applications to Science of Science	58
2.7.2 Applications to Information Research	58
2.7.3 Application to Literature Information Management	59

Contents	xiii
3 Literature Information Obsolescence Law	61
3.1 The Concept and Measure of Literature Information Obsolescence	61
3.1.1 The Concept of Literature Obsolete and Intelligence Obsolete	62
3.1.2 Measure for Literature Obsolescence	65
3.2 Research Methods of Literature Obsolescence	68
3.2.1 Statistical Data Analysis of Literature Management	68
3.2.2 Citation Analysis	69
3.2.3 Mathematical Methods	71
3.2.4 Comprehensive Analysis Method	71
3.3 The Mathematical Model and Index of Literature Information Obsolescence	72
3.3.1 Classical Mathematical Model and Obsolescence Index	72
3.3.2 Grey Dynamic Model (GM) and Obsolescence Index	80
3.4 The Mechanism of Research and Analysis of Literature Information Obsolescence	83
3.4.1 Several Types of Literature Obsolescence	83
3.4.2 Several Circumstances of Literature Information Obsolescence	84
3.4.3 Factors Affecting the Literature Information Obsolescence	85
3.5 The Application of Literature Information Obsolescence Law	87
3.5.1 The Application in Document Information Management	87
3.5.2 The Application in the Study of Science of Science and Technology	88
4 Concentration and Scattering Distribution of Literature Information: Bradford's Law	89
4.1 Background of Bradford's Law	89
4.1.1 Founder: Bradford	89
4.1.2 Background of Bradford Law's	90
4.2 Formation of Bradford's Law	92
4.2.1 Proposal of Bradford's Law	92
4.2.2 Establishment of Bradford's Law	95
4.3 Basic Content of Bradford's Law	95
4.3.1 Elaboration of Bradford's Law	95
4.3.2 Consistency of Theory with the Practice of Bradford's Law	99

4.4	Development of Bradford's Law	101
4.4.1	Development Process of Bradford's Law	101
4.4.2	Vickery's Inference on Bradford's Law	103
4.4.3	Leimkuhler's Contribution to Bradford's Law	106
4.4.4	Brookes' Description of Bradford's Law	108
4.4.5	Unified Equation of Смольков	110
4.4.6	Theory and Development Trend of Bradford's Law	112
4.5	Applications of Bradford's Law	113
4.5.1	Basic Method for the Application of Bradford's Law	114
4.5.2	Main Region of the Application of Bradford's Law	115
4.5.3	Conditions and Limitations of the Application of Bradford's Law	119
5	Word Frequency Distribution of Literature Information:	
	Zipf's Law	121
5.1	Theoretical Basis of Zipf's Law: Principle of Least Effort	121
5.1.1	Principle of Least Effort	122
5.1.2	Principle of Least Effort and the Word Frequency Distribution Law	122
5.2	Formation and Establishment of Zipf's Law	123
5.2.1	Appearance of Frequency Dictionary	123
5.2.2	Estoup's Found	124
5.2.3	Condon's Formula	124
5.2.4	Zipf's Research and the Establishment of Zipf's Law	126
5.3	Basic Content of Zipf's Law	127
5.3.1	Textual Representation of Zipf's Law	127
5.3.2	Image Description of Zipf's Law	128
5.3.3	General Mathematical Form of Zipf's Law	130
5.3.4	Applicability of Zipf's Law	130
5.4	Development of Zipf's Law	131
5.4.1	Joos's Double-Parameter Formula	131
5.4.2	Three-Parameter Formula of Mandelbrot	132
5.4.3	Low-Frequency Word Distribution: Zipf's Second Law	133
5.5	Applications of Zipf's Law	135
5.5.1	Application to Literature Indexing and Thesaurus	136
5.5.2	Application to Information Retrieval	138
5.5.3	Application to Science Evaluation	139

6 Author Distribution of Literature Information: Lotka's Law	145
6.1 Background of Lotka's Law	146
6.1.1 Founder of Lotka's Law: Lotka	146
6.1.2 Background of Lotka's Law	146
6.2 Formation and Basic Content of Lotka's Law	147
6.2.1 Formation of Lotka's Law	147
6.2.2 Content of Lotka's Law	152
6.2.3 Generalized Lotka's Law	153
6.3 Development of Lotka's Law	154
6.3.1 Verification of Lotka's Law	154
6.3.2 Contributions of Fracci	158
6.3.3 Development of Lotka's Law in China	161
6.3.4 Research on Collaborators	168
6.4 Price's Law and the Distribution of Other Authors	174
6.4.1 Price's Law	174
6.4.2 Distribution of Other Authors	176
6.5 Application of Lotka's Law	178
6.5.1 Function of Lotka's Law	178
6.5.2 Problems that Should Be Noticed During Application	183
7 Statistical Analysis Method for Literature Information	185
7.1 Significance and General Concept of Literature Information Statistics	185
7.1.1 Literature on Information Statistics and Its Significance	185
7.1.2 General Concept of Literature Information Statistics	187
7.2 Principles and Indexes of Literature Information Statistics	190
7.2.1 Principle Requirement for Literature Information Statistics	190
7.2.2 Index System of Literature Information Statistics	191
7.2.3 Statistical Indicators of Information Resource Management	193
7.3 Types and Basic Steps of Literature Information Statistics	194
7.3.1 Main Types of Literature Information Statistics	194
7.3.2 Basic Steps of Literature Information Statistical Analysis	196
7.4 Application of Literature Information Statistical Analysis	198
7.4.1 Application to Information Resource Management	199
7.4.2 Application to Information Users and Literature Information Utilization Research	200
7.4.3 Application to Literature Information Law Research	200
7.4.4 Application to Discipline Development Law Research	201

7.5	Mathematical Statistical Method and Its Application	203
7.5.1	Summary of Mathematical Statistical Method	203
7.5.2	Applications of the Mathematical Statistical Method	204
8	Methods of Citation Analysis	207
8.1	Basic Concepts and Methods of Citation Analysis	207
8.1.1	Basic Concepts of Citation Analysis	207
8.1.2	Citation Behavior and Motives	209
8.1.3	Basic Types and Steps of Citation Analysis	210
8.2	Citation Analysis of the Main Tools.	212
8.2.1	SCI	212
8.2.2	Essential Science Indicators (ESI)	224
8.2.3	Main Tools of Domestic Citation Analysis	245
8.3	Distribution Law of Citation and Key Indicators of Analysis	247
8.3.1	Citation Structure and Its Significance.	248
8.3.2	Distribution Law of Citation Quantity	248
8.3.3	Garfield's Law of Citation Concentration	251
8.3.4	Analysis of the Main Index of Citation Measures	252
8.3.5	Self-citation Analysis of Scientific Literature.	257
8.4	Citation Analysis of Scientific Journals.	259
8.4.1	Decentralization and Centralization Law of Periodical Literature	260
8.4.2	Main Indices of Evaluating Journals	262
8.4.3	Journal Citation Reports (JCR)	270
8.5	Citation Network and Cluster Analysis.	282
8.5.1	Concept of Bibliographic Coupling and Co-citation	283
8.5.2	Coupling Analysis	288
8.5.3	Co-citation Analysis	292
8.5.4	Citation Cluster Analysis.	295
8.6	Application of the Citation Analysis Method	301
8.6.1	Application of the Citation Analysis Method.	302
8.6.2	Application Example of the Citation Analysis Method	306
8.6.3	Limitation of the Citation Analysis Method.	308
9	Methods of Computer-Aided Informetrics Analysis	311
9.1	Significance of Computer-Aided Informetrics	311
9.1.1	Improve the Efficiency of Informetrics Analysis	311
9.1.2	Improve the Reliability of Informetric Analysis.	312
9.1.3	Promote the Accuracy of Informetrics Research	313
9.1.4	Expand the Research Area of Informetrics	313

9.2	Feasibility of Computer-Aided Informetrics Analysis	314
9.2.1	Computer and Network Technology as Fundamentals	314
9.2.2	Development of Literature Digitalization as a Prerequisite	315
9.2.3	Theoretical Foundation Laid by the Development of Informetrics	316
9.2.4	Research Development from Abroad Provides Experience	316
9.3	Theoretical Basis of Computer-Aided Informetrics	317
9.3.1	Major Types of Computer-Aided Informetrics Analysis	317
9.3.2	Structure and Function of Computer-Assisted Information Metering Analysis	318
9.3.3	Steps of Computer-Assisted Informetrics Analysis	319
9.4	Construction of a Citation Database and Data Mining Analysis	320
9.4.1	Citation Analysis Database	320
9.4.2	Citation Analysis System Design	321
9.4.3	Case Study on the Design of the Online Edition of Chinese Social Science Citation Index (CSSCI)	324
9.4.4	Mining Analysis Methods of Citation Data	326
9.5	Application of Computer-Assisted Information Metering Analytical Method	331
9.5.1	Application in Scientific Research	331
9.5.2	Application in the Information Source Field	332
9.5.3	Application in Competitiveness Analysis	333
9.6	Development Direction of Computer-Assisted Information Metering Analysis	333
9.6.1	Development in Breadth and Depth	333
9.6.2	Development in Practicality	334
9.6.3	Development in Integration	334
9.6.4	Development in Modelling	335
9.6.5	Development in Intelligentization	335
10	Application of Informetrics in Information Resource Management and Research	337
10.1	Informetrics and the Determination of Core Journals	337
10.1.1	Theoretical Basis and Formation Mechanism of Core Journals	338
10.1.2	Concept of the Core Periodical and the Important Meaning of the Measurement	339
10.1.3	Measurement Method—The Method System of Informetrics	340

10.2	Informetrics and Documentation Information Collection and Management	348
10.2.1	Determine the Best Program for Periodical Collection	348
10.2.2	Select the Best Means to Collect Literature	349
10.2.3	Using Bradford's Law as a Literature Purchasing Strategy	350
10.2.4	Using Literature Obsolescence Law to Guide Book Weeding Out	352
10.2.5	Best Allocation for Literature Purchasing Funds	353
10.2.6	Calculation Method of Book Shelf Placeholder	355
10.2.7	Evaluation of Literature Collection Work	356
10.3	Informetrics and Information Retrieval	357
10.3.1	Determination of the Integrity of Search Tools	357
10.3.2	Bradford's Law for Information Retrieval	359
10.4	Informetrics and User Research	361
10.4.1	User Distribution in Line with the Law of Bradford	362
10.4.2	Guide Users in Using Journals	362
10.4.3	Guide the Reader in Buying and Reading the Best Books	363
10.5	Concentration, Dispersion Laws, and Examples of Document Information Flow	364
10.5.1	Research Methods	364
10.5.2	Research Results	367
10.6	Examples of and Research on the Law of Literature Information Utilization	367
10.6.1	Research Methods of the Law of Literature Information Use	369
10.6.2	Study of the Law of Scientific Researchers Using Literature Information	374
11	Application of Informetrics in Science and Technology Management and Forecasting	377
11.1	Informetrics and Science of Science	377
11.1.1	Basic Principles	377
11.1.2	Research Contents	378
11.2	Informetrics and Talent Evaluation	386
11.2.1	Talent Evaluation Theory of Informetrics	386
11.2.2	Talent Evaluation Methods	387
11.2.3	Selection of Outstanding Scientists	388
11.2.4	Forecasting of Future Winners	391

11.3	Informetrics and Regional and Institutional Research Evaluation	393
11.3.1	Scientific and Quantitative Study of Braun	393
11.3.2	Schubert and Other Scientists' Measurement Research	395
11.3.3	Evaluation of Scientific Research Institutions	397
11.3.4	Evaluation of Scientific Research in China	399
11.4	Informetrics and Science and Technology Forecasting	406
11.4.1	Informetrics and Science and Technology Forecasting	406
11.4.2	Conducting Technical Evaluation and Prediction by Using Patent Documents	415
11.4.3	Using Informetrics to Predict the Development Trend of a Discipline.	421
11.4.4	Prospects of Product Development and Application Using Informetrics.	428
Bibliography	431

Chapter 1

Introduction

1.1 Origin and Development of Informetrics

1.1.1 *Origin of Informetrics*

Informetrics, which originates from the German word “informetrie,” was first proposed by the German scholar, Otto Nacke. The corresponding English term “informetrics” appeared in later literature. Some scholars believe that this English term first appeared in the title of an annual research project published by the National Science Foundation in 1980, whereas others assume that its English translation came from non-English speaking countries, namely, the Japanese magazine “Information Management” and the “Information Science Abstracts” from the Soviet Union. In September 1980, Professor Nacke introduced the term “informetrics” during the first informetrics (including scientometrics) seminar that convened in Frankfurt, Germany. In 1981, both German and English terms appeared in domestic periodicals and were translated as “informetrics.” However, arguments were raised. A number of articles have pointed out that this term should be subjected to particular scrutiny. Related studies have determined that the suffix “metrics” follows a vowel (such as “a” or “o” in bibliometrics and scientometrics) in about 10 subject names that contain the element “metrics.” Nevertheless, this suffix follows a consonant “r” in the term “informetrics,” which appears at odds with the general word formation rule in English. The credibility of the term, which has been translated by non-English speakers, remains to be verified. Nonetheless, the term has already gained prevalence in English-speaking countries, and it has been recognized by the International Federation of Documentation (FID). The doubtful attitude of some scholars toward the word has been eventually eliminated, and this emerging subdiscipline has begun to flourish.

In the early 1980s, FID/IM established the Committee of Metrology (FID/IM), which had a standing body located in the Indian National Scientific Documentation Center and with T.N. Nagy as the president. With the objective of strengthening

education and research in informetrics, FID/IM formulated key points for the work scheme and planned to conduct important academic activities to promote the development of the aforementioned subject. In 1982, the former Czechoslovakia held a conference on the teaching plans of informetrics. Two years later, a few basic theoretical issues were elucidated in two important papers of B.C. Brooks, which explicitly advocated for the vigorous development of informational metrology. With its plans evaluated and demonstrated, FID/IM published “Newsletter Informetrics” in India in 1985. In the First International Symposium on Bibliometrics and Information Retrieval Theory held in 1987 in Belgium, the proposition of Brooks to add “informetrics” to the name of the second session of the international conference scheduled to be held in 1989 in Canada won universal recognition and support from attending scholars. The views of Brooks were accepted in the third and fourth international conferences held in India and Germany in 1993 and 1994, respectively. In June 1995, the academic conference held in Chicago was renamed as the “Fifth International Conference on Scientometrics and Informetrics.” Although bibliometrics was discussed in the conference, the term was not included in the title of the conference. The organization that held biennial international seminars also had its name changed to “International Society for Scientometrics and Informetrics” (ISSI). All the aforementioned changes have demonstrated the recognition of informetrics by the international academic circles and its rising prominence.

The year 1987 was regarded as the one during which “informetrics” was officially recognized by the international information community and by several foreign information scientists because the term appeared on the title of papers related to or published by international academic conferences since 1987. Treatises with the term “informetrics” on their titles were published, as evidenced by the “Introduction to Informetrics” written by L. Egghe and R. Rousseau in 1990 and the paper “Introduction to Informetrics” by Canadian information scientist, S.T. Tague. After 1987, several Western information service corporations even changed their names to the trendy “Informetrics Limited” when the term became prevalent in North America and Western Europe.

Academies in China have also reacted promptly and given sufficient attention to informetrie (in Germany), informetrics (in English), and the subjects represented by the term. For example, related papers were published early in 1981, and the official publication of “bibliometrics” in 1988 did not only elucidate scientometrics, informetrics, and bibliometrics (tri-metrics), but also propose the framework for the contents of informetrics. Informetrics has acquired a consistent translation result from scholars as a third-grade branch discipline in response to information science. However, a relevant department in China altered its translation of “informetrics” in 1992, which became the origin of informetrics in China.

1.1.2 Background of Informetrics

Similar to other disciplines, the introduction to informetrics has not occurred as a certainty but has formed and developed under a certain background and scientific environment.

First, informetrics has extended and evolved based on traditional bibliometrics and scientometrics. The early stage of informetrics coincided with the rise of bibliometrics and scientometrics when related studies were active. Several early information scientists who focused on quantitative studies regarded bibliometrics and scientometrics as their fields of study; thus, people assumed that these fields combined specific methods from information science. When some information scientists joined in bibliometric research, considerable progress was achieved on useful extensions and studies on the research scope, methods, models, applications, and other aspects of bibliometrics. In the early 1980s, Brooks studied Bradford's law under common social circumstances and replaced the terms "periodical" and "paper" with "source" and "item," respectively, to make them widely acceptable. He also extended bibliometrics into a calculus of social science, such that it would play a broader role in society. However, after all the extensions scholars made on the research scope of and method for bibliometrics, they determined that informetrics was not a subject that entirely belonged to bibliometrics, but one with wider areas of measurement and quantitative research, as well as its own unique research method. They finally realized a simple but important fact, one that would be impossible to notice earlier, i.e., the number of and metering method for information considerably exceeded that of the literature. This significant finding was proposed in a series of papers published by Brooks, who clearly advocated for the development of informational metrology. At that time, bibliometric research mainly served the research purposes and needs of library science. Library scientists wanted bibliometrics to be their exclusive field of study. This situation prompted information scientists to establish their own area of quantification research that corresponded to information science, thereby promoting the formation and development of informetrics.

Second, informetrics is the inevitable outcome of the development of information science. Quantification research has always been a significant direction and an inevitable tendency in the development of information science because of several reasons. Information science will inevitably evolve from a qualitative stage to a quantitative stage given the general discipline development rules. Only through strengthened quantitative research can information science become highly scientific and accurate, thereby establishing and promoting its status in the entire science system. As Brooks stated, "information science will remain a heap of unconnected techniques and never become a subject of science until quantitative studies are conducted." This important academic idea has received increasing recognition from a growing number of scholars. Others have also actively participated in studying this aspect. With extensive research findings being published, informetrics, as a quantitative subdiscipline of information science, is gaining increasing momentum.

Therefore, informetrics is an inevitable product of the quantitative development of information science.

Third, the field of informetrics comprises a number of backbones and discipline leaders. Information scientists, with their solid foundation of knowledge in mathematics and physics as well as familiarity with quantitative research methods, have the advantage and talent that can guarantee further development of informetrics. Early scholars, such as Bradford, Lotka, and Price, and later ones, including Brooks and Garfield, are all specialists who have devoted themselves to research with a solid store of knowledge and well-used methods. They have played the role of leader researchers. For example, as an outstanding representative, Brooks actively advocated and attach great importance to the quantification of information science, he thought information science is the study of the essence, as same as measurement, of information and knowledge, and also creatively put forward the “ranking technology” and “logarithmic perspective principle” as a way of quantitative information science. In 1988, Brooks proposed that we should replace bibliometrics to informetrics, and the reason is that bibliometrics is only confined to the bibographic metrology, and is not suitable for modern electronic measurement of the carrier of literature. This has played a role on the formation and development of informetrics.

1.1.3 Development of Informetrics

The development course of informetrics is described as follows: statistical bibliography → bibliometrics → scientometrics → informetrics. The earliest informetric research started at the beginning of the 20th century when the famous philologists F.T. Cole and N.B. Eales conducted literature statistical research in 1917. In 1992, English library scientist E.W. Hulme used the term “statistical bibliography” for the first time in his book “The Relation between Statistical Bibliography and the Development of Modern Civilization.” This term refers to a new subdiscipline under bibliometrics that determines the nature of library materials through statistical methods. In 1969, the proposal of the renowned English information scientist Alan Pritchard that the term “bibliometrics” could be used to replace “statistical bibliography” received universal acknowledgment from library science and information science scholars. The emergence of this term officially marked the birth of bibliometrics. Similar to the early history of bibliometrics, the history of scientometrics dates back to the beginning of the 20th century when European and Russian scholars conducted statistical analyses of bibliographic citations. In the same year when the term “bibliometrics” was coined, scholars from the former Soviet Union, V.V. Nalimov and Z.M. Mulchenko, introduced the term “scientometrics” as a scientific quantitative approach to studying and analyzing information. In 1978, the magazine “Scientometrics” founded by Hungarian scholar Tibor Braun provided an academic exchange platform for international scientometric scholars, thereby promoting the development of scientometrics. The early

history of informetrics is incorporated into the history of bibliometrics and scientometrics given that informetrics is a legacy and extension of the latter two. The term “informetrics” proposed by the German scholar Otto Nacke in 1979 failed to gain universal recognition from library science and information science scholars. However, Western information scientists have exerted continuous efforts to establish the important position of informetrics, as evidenced by the founding of the Committee on Informetrics by FID through the persuasion of information scientists in 1980 and the long-term project on teaching and studying informetrics; the publication of the informal magazine “*Informetrics Newsletter*” in India; the first international academic conference, namely, the International Symposium on Bibliometrics and Information Retrieval Theory held in Belgium and initiated by ISSI, and the subsequent conference proceedings on informetrics, which had attracted considerable attention from the bibliometric and informetric circles. Informetrics has invariably been designated as the core theme of the aforementioned conference since 1987, and the name of the conference has been officially changed to “International Conference of the International Society for Scientometrics and Informetrics” since 1995 when informetrics has gained wide acceptance among scholars with its gradually increasing influence.

ISSI has played a pivotal role in the development of informetrics. Since 1987, it has held biennial international conferences on scientometrics and informetrics, with 15 successful consecutive conferences. Different themes have been selected, and related papers have been widely collected and discussed during each conference, thereby positively affecting the development of informetrics. The title, time, venue, and theme of each conference are listed in Table 1.1.

The accelerated development of information science and information technology, coupled with the digitalization of information resources and the increasing popularity of the Internet, has significantly affected and profoundly transformed all aspects of our society, economy, technology, and culture. Under the new social context and technological condition, the study and development of informetrics have presented different directions and trends. In the current study, we mainly focus on the following aspects.

(1) From bibliometrics to informetrics

In our article entitled “The Progress and Development Direction of Domestic Bibliometrics,” we indicated that similar to the relationship between literature and information, bibliography and information science and bibliometrics and informetrics are also inextricably intertwined and mutually complementary. Bibliometrics is the foundation of informetrics, and informetrics is the development direction of bibliometrics. The book entitled “Bibliometrics” states that the progress achieved in bibliometrics also contributes to and promotes the development of informetrics. We must incorporate the study and research of informetrics into our agenda in due course and exert assiduous efforts to promote the development of informetrics. The article “Progress in the Quantitative Research on Domestic Information Science” divides papers on the quantitative research of information

Table 1.1 Outline of the 15 International Symposiums on informetrics and scientometrics

Number	Title	Date	Place	Topics
First	International Symposium on Bibliometrics and Information Retrieval Theory	25–28 August, 1987	Diepenbeek, Belgium	1. In-depth discussion on basic laws 2. Application of citation analysis
Second	International Symposium for Bibliometrics, Scientometrics, and Informetrics	5–7 July, 1989	London, Canada	1. Tri-metric (scientometrics, informetrics, and bibliometrics) scope definition 2. Rule generalization
Third	International Symposium on Informetrics (Indian Statistical Institute)	9–12 August, 1991	Bangalore, India	1. Application of statistical methods to informetrics 2. Application of mathematical methods
Fourth	International Symposium for Bibliometrics, Scientometrics, and Informetrics	11–15 September, 1993	Berlin, Germany	1. Relationship among tri-metric (scientometrics, informetrics, and bibliometrics) studies 2. Application of citation analysis
Fifth	International Symposium for Scientometrics and Informetrics	7–10 June, 1995	Illinois, United States	1. Discussion on periodical evaluation 2. Extension to basic laws
Sixth	(same as above)	16–19 June, 1997	Jerusalem, Israel	1. Application of citation analysis 2. Studies on the aging and dispersion laws of literature 3. Data compression 4. R&D management
Seventh	(same as above)	5–8 June, 1999	Colima, Mexico	1. Academic journal evaluation 2. Content analysis 3. Citation analysis and mathematical model 4. Law distribution and demonstration
Eighth	(same as above)	16–20 June, 2001	Sydney, Australia	1. Laws and their distribution in the field of science 1 2. Mathematical model for information measurement 3. Citation motivation and scientific evaluation

(continued)

Table 1.1 (continued)

Number	Title	Date	Place	Topics
				4. Knowledge map and visualization 5. Analysis and forecast of scientific and technological policies 6. Library management
Ninth	(same as above)	25–29 August, 2003	Beijing, China	1. Mathematical modeling of information measurement 2. Scientific evaluation and university-ranking methodology 3. Citation analysis and database 4. Quantitative analysis of scientific and technological innovations (patent) 5. Network information retrieval research
Tenth	(same as above)	24–28 July, 2005	Stockholm, Sweden	1. History of scientometrics 2. Citation motivation research 3. Knowledge map 4. Webmetrics 5. Science policy analysis and forecast
Eleventh	(same as above)	25–27 June, 2007	Madrid, Spain	Refer to http://issi2007.cindoc.csic.es/ ^a
Twelfth	(same as above)	14–17 July, 2009	Rio de Janeiro, Brazil	Refer to http://www.issi2009.org/php/index.php
Thirteenth	(same as above)	4–7 July, 2011	Durban, South Africa	Refer to http://www.issi2011.uzulu.ac.za/index.php
Fourteenth	(same as above)	15–19 July, 2013	Vienna, Austria	Refer to http://www.issi2013.org/
Fifteenth	(same as above)	29 June–4 July, 2015	Istanbul, Turkey	Refer to http://www.issi2015.org/en/default.asp

^aWe have not listed the diverse topics

science into four categories: ① bibliometrics and its application, ② information retrieval theory, ③ theoretical study of information science, and ④ information economics and information result evaluation. Among which, the category

“bibliometrics and its application” accounts for a relatively large proportion of 46.6%. Bibliometrics constitutes an important aspect in the quantitative research of information science and is currently progressing toward informetrics.

In terms of metrological units, informetrics has gone beyond the metrological analysis of bibliometric units, such as articles, volumes, and books, but has probed further into the literature to conduct metrological analysis of contents and related information therein, such as titles, subject terms, keywords, word frequency, knowledge items, citation information, author, publisher, date, language, and format. In early 1980, Sen Long, who worked for the Japan Science and Technology Information Center, successfully predicted the structure and development prospects of polymer material products by conducting statistical analysis of the keyword occurrences of terms such as “plastic,” “rubber,” and “fiber.” We once conducted a statistical analysis of the number of content topics of a large number of related literature in the article “The Quantitative Analysis of the Research Topic Trends of Domestic Library and Information Science” and provided a quantitative revelation of the development process, research priorities, popular themes, and trends of domestic library and information science. The electronic publication developed recently by Professor Chen Guangzuo introduced informetrics and knowledge item clustering functions, thereby opening new areas of applications and creating a new development approach for bibliometrics. Any knowledge element or even every word of the text of an electronic publication, as a full-text database, can be retrieved and statistically analyzed. In this case, the metrological unit of bibliometrics can evolve from an independent piece of literature to the knowledge elements or even a single word in the literature, thereby making in-depth informetric analysis possible. This evolution is an important progress that shows that bibliometrics has developed into informetrics, and this trend will continue further.

(2) Computer-aided research and application of informetrics

A large number of studies on computer-assisted informetric analysis have achieved considerable success and increasingly widespread applications since the 1990s. Informetric study requires data support of a certain scale. In addition, a systematic and standardized system of data source and channels to obtain original data should be established; modern methods and tools, such as computers, should be used to conduct data processing and analysis. Foreign academic circles place considerable importance on efforts in this respect. In the early 1960s, the United States began to prepare the “Science Citation Index” (SCI). The publication and distribution of this huge index provide a powerful and multifunctional tool to study informetrics and, to a certain extent, a large amount of data are indispensable to conduct citation analysis, thereby effectively promoting the full-scale quantitative study of informetrics and information science. This case is also observed domestically. We have long recognized that if modern technical means, such as computers, are not used to solve the problem of informetric tools, then domestic study on informetrics will probably never reach a new stage of development. Hence, we emphasized in some related literature that conducting studies on the modernization of informetric tools

and instruments was an urgent task that should merit our attention. We established three proposals to conduct the study: ① introduce and develop SCI, ② compile our own “Chinese SCI” (CSCI), and ③ conduct research on computer-assisted informetric analysis.

We have performed bold explorations in these aspects and achieved considerable progress in recent years. The Chinese Institute of Science and Technology established the Chinese Science and Technology Papers and Citation Database (CSTPC) in 1987 and conducted a multi-index statistical analysis on a number of papers cited by Chinese scholars. The results, which were released annually, profoundly influenced the society, and thus, effectively promoted the popularization and development of informetrics. The Documentation and Information Center of the Chinese Academy compiled and published CSCI in 1995 after years of efforts. The production and release of its CD-ROM version followed shortly in 1998. In 2000, the China Social Science Research Evaluation Center of Nanjing University and the Hong Kong University of Science and Technology jointly developed the Chinese Social SCI (CSSCI) to compensate for the shortage of data sources in humanities and social sciences. They provided a large-scale tool for retrieving data that could be extensively applied to quantitative studies on informetrics.

The paper entitled “The New Trend of the Study on Bibliometrics—Computer-aided Bibliometric Study” was the first to discuss the quantitative analysis of computer-assisted informetrics and its approaches; it proposed that computer-aided bibliometric studies should be used to make bibliometric research more standard and modern. This paper also discussed system designs and approaches to conducting computer-aided quantitative analysis of literature information as follows. ① A database system that is mainly used for the informetric analysis of literature information is established. We can design various types of bibliometric information system in terms of the different objectives and requirements of studies starting from the principles of bibliometrics. ② Existing information retrieval systems should be utilized and improved to adapt to the informetric analysis of literature information. Related recordings of data should be increased according to the characteristics and requirements of informetric studies based on the original information retrieval system, thereby expanding the scope of statistics to be capable of conducting quantitative analyses. ③ Copy technology is used to establish a database, particularly for the informetric analysis of literature information. Software programs, such as SCI-NATE, that are capable of performing copy technology to establish databases, are already available. The paper “Bibliometric Methods and Computer-aided Bibliometric Research” discussed new methods for bibliometric research, particularly computer-aided bibliometric research methods from a methodological point of view. The authors of the paper designed and developed software on computer-aided bibliometric analysis, reestablished a set of recorded data, and conducted statistical analysis on a variety of data through Pro*C based on theoretical analysis with the support of the Oracle database system. They also performed statistical analysis on a series of multifaceted literature information, which was all based on 16,000 physics articles published by Chinese scholars in the British? “Physics Abstracts” (PA) in 1992–1994, using self-programmed software.

All the aforementioned efforts do not only practiced and verified methods of computer-aided bibliometric research, but also offered a quantitative revelation of the development characteristics, key areas, and power distributions in the field of physics in China, thereby drawing useful results. In summary, the establishment and perfection of computer-aided bibliometric analysis methods indicate that the domestic informetric analysis method system has basically taken shape and will continue to improve in the future.

(3) Study of webometrics

Two new terminologies, namely, “webometrics” and “cybermetrics,” have emerged in related web pages and literature in recent years. However, the measurement object is online information or information controlled by the computer, rather than the “Net” or the “computer” itself. Therefore, we can paraphrase the terms as “online informetrics.” Current literature information indicates that the term “webmetrics” was first proposed by Almind in 1997. In the same year, Armand et al. first coined the term “webmetrics” in the paper entitled “Informetric Analysis on the World Wide Web—A Discussion on the Method of Webometrics” and proposed that various informetric methods could be used to conduct quantitative analysis on the World Wide Web. For another similar term, i.e., “cybermetrics,” electronic journals or academic forums on the Internet have been named after this term. These forums and journals are mainly organized and published by the Spanish Scientific Information and Documentation Center (CINDOC).

Different interpretations of the concept of “webometrics” are available abroad. Some scholars have defined it as a discipline that aims to conduct statistical analysis of online literature; others consider it a study on the mutual references of online data. Furthermore, some researchers approach the problem from the perspective of cyberspace computing and application software, believing that webometrics is a discipline about computer software design, which is not the case. If the research objects, methods, contents, and objectives of webometrics are considered, then we believe that it is an emerging subdiscipline that aims to offer a quantitative description and statistical analysis of the organization, storage, distribution, transmission, mutual citation, development, and utilization of online information through quantitative methods, such as mathematics and statistics. This subdiscipline is designed to present the quantitative characteristics and inherent laws of online information. As an interdisciplinary subject, webometrics incorporates network technology, network management, information resource management, and informetrics. This subject also offers a new direction and an important research domain for informetrics, thereby gaining extensive application prospects. Its fundamental purpose is to provide the necessary quantitative basis for the orderly and rational distribution of online information, the optimized allocation and effective use of information resources, and the standardization of network management. Therefore, the organizational management and information management of networks can be significantly improved, thereby optimizing their economic and social benefits.

Webometrics has rapidly formed and developed under the current scientific background and technical conditions. First, the burgeoning online information and literature do not only provide the necessary foundation and conditions for the emergence of webometrics, but also generate a pressing practical need for it, thereby promoting the formation and development of this discipline. Second, the statistical analysis and research findings of online literature information form the foundation of the discipline and accumulate related experience. In the 63rd International Federation of Library Associations and Institutions (IFLA) conference held in 1997, three papers that focused on the statistical analysis of online information were presented. Among which, the paper “Feature of the Accessible Information on the World Wide Web” by T.O. Edward of the United States Online Computer Library Center discussed statistical indicators, statistical types, and other issues regarding online information. The other two articles discussed statistical issues regarding online information service in libraries. In the 65th IFLA conference held in Thailand in 1999, a number of papers also focused on such issues, thereby demonstrating progress to a certain extent. Third, the development of informetrics has generated certain practical needs. With the increasing amount of online information, the research object and scope of informetrics are bound to expand to the online world, which is an objective requirement and inevitable trend for the development of a discipline. Fourth, network management should be strengthened and improved. With the increasing popularity of networks, strengthening network management has become a top priority, and the implementation of quantitative management is one of the major strategies to do so. Research findings on webometrics are bound to provide theoretical guidance and quantitative basis for quantitative and scientific network management; in turn, the practical need for quantitative network management will promote the comprehensive development of webometrics.

We believe that the study object of webometrics should be understood in a broad sense. In the current study, the term “network” does not only refer to the Internet, but also to other types of networks, such as local area networks. The metering target of online information mainly involves three levels or components: ① the direct measurement of online information, including digital information, text message, and multimedia information, which incorporate text, images, and sound, such as information in the unit of byte and the measurement of information flow; ② the measurement of online literature and its information, as well as other related information, such as online electronic journals, theses, books, reports, the literature distribution structure, discipline theme, keywords, author information, and publishing information that involves primary literature as well as secondary and tertiary sources of measurement issues; ③ the measurement problems of an information network structural unit, such as the information growth of network sites, distribution of subjects, information transmission, and mutual citations and links among sites. Thus, webometrics covers a wide range of problems.

Similar to bibliometric and informetric systems, a webometric system consists of its theories, methods, and applications. Among which, theory is the basis, method is the means, and application is the goal. These three aspects complement one another,

and neither of them should be neglected. Theory mainly focuses on the fundamental issues that must be addressed in webometrics, such as an independent subject, new concepts, new indicators, and new laws, including the concentration and dispersion laws of online information, law about the author, growth and aging rules of word frequency, citation rules, multimedia information, and the theoretical explanation and mathematical models for these laws. Method mainly focuses on the principle, applicability, and operating procedures of the application of various quantitative methods, such as literature information statistical analysis, mathematical model analysis, citation analysis, bibliographic analysis, and systematic analysis. Necessary amendments to improve and perfect these methods are also included. With regard to application, the main task of webometrics is to study its applications in multidisciplinary and multi-industry sectors, such as library information systems, information resource management, network management, science, scientific evaluation, technology management, and forecasting theory. Therefore, the value of webometrics can be fully utilized and contributions to the development of technology, economy, and society are made.

1.2 Concept and System Structure of Informetrics

1.2.1 Aim and Significance of *Informetric Research*

The aims of informetric research are to introduce the concept of quantity and quantitative methods as well as to further present the structure and law of quantity change in information elements (including documents, data, object, information, and events), thereby theoretically improving the science and accuracy of information science and other disciplines related to information management as well as developing them in a quantitative stage. Moreover, research on informetrics provides a quantitative basis for improving information systems and realizing high efficiency of scientific management, thereby helping information communication systems to always operate at their best status and resolving basic contradictions in information service through providing the best information service. Consequently, information management can better serve the development of science and technology, the economy, and the society when getting over an information crisis.

In general, the greatest significance of informetric research is as follows: it continues to sum up various experiences and laws from a theoretical perspective and transforms information “work” into information “science,” thereby making its theory highly extensive and profound. Simultaneously, such research verifies and corrects old experiences and laws under new information conditions and explores their new applicability. All these efforts make information science considerably scientific and enable it to provide theoretical guidelines for practical work.

The significance of this thesis lies in its application, whereas those of the theory of and method for informetrics in different fields are as follows:

- (1) Application to improve and enrich information science theory research. On the basis of information work, informetrics demonstrates principles and methods in information collection, information arrangement, information analysis, and information utilization, thereby creating conditions for scientific and standardized management as well as significantly improving the scientific nature of basic information science theory.
- (2) Application to library management. The use of informetrics in library work can help identify key periodicals; determine the utilization rates of books and reference materials; provide scientific basis for the rational purchasing and storage of books; ensure the preservation period to optimize collections; evaluate the integrity of the bibliography, abstract, and index to compile a retrieval language and to improve secondary literature service; assess the circulation of book resources and refusing ratio; select the best number of copies and establish a rational layout for a library database; and allocate funds rationally through the statistical analysis of book circulation rates after clearing the economic proportion of order quantity and copy quantity, which will be helpful in providing readers with targeted service when understanding their reading tendency and habits to obtain information resources.
- (3) Application to information analysis and forecast. We can analyze and estimate the development trend of a certain subject or technology by using the basic principles and methods of informetrics and by processing and sorting survey data. For example, the generation, development, differentiation, and mutual penetration of a certain subject can be estimated according to the relationship among its growth, the number and content construction of relevant literature, and their mutual references. On this basis, the research results of overviews and special field surveys can be obtained, and the subject distribution of information sources and subject characters can be easily determined. The analysis of document statistics is also conducive to determining the distribution of documents regarding a certain subject to understand the nature of a discipline. The analysis of literature citation frequency will help in understanding the influence of a subject and the significance of subjects in certain countries. Research on literature obsolescence speed contributes to determining the corresponding speed of subject development and can provide hard evidence for science and technology management. Such parameter can also be used to study the history of scientific development mainly from citation time distribution (historical map) and the net-like relationship among citations.
- (4) Application to information retrieval. In information work, the use of informetrics can demonstrate the characteristics of information units, such as growth rate, aging coefficient, valued year of life, and dispersion rules; it will be beneficial for reasonably determining the retrospective retrieval of periodicals during the automatic retrieval process, the time of data scraping, cumulative index journals, and the value and availability of related information. In terms of contribution to scientific research, informetric analysis can help researchers obtain a large number of information sources, understand the background value of literature, identify its references and referenced behavior, track the

development trends and directions of science and technology in a timely manner, and select and determine scientific research subjects. Research on informetrics also contributes considerably to determining the attraction and accuracy of a certain topic, research results, and the effect of some research methods. The use of the co-citation analysis method has effectively improved the retrieval percentage of precision and recall by clustering citations with considerable citation strength.

- (5) Application to scientific evaluation. Informetric data are useful in evaluating the quality of scientific papers, scientific achievements, scientific research and its efficiency, and author reputation. They also provide important basis for the planning and management of scientific research. As a current crucial social issue, the scientific evaluation of talents demands an objective indicator based on scientific achievements that are generally presented on papers. In this case, informetric data play a significant role in talent studies. Hence, university evaluation and scientific research evaluation are crucial.
- (6) Application to other social disciplines. Informetrics is also concerned with applications in the fields of history, sociology, and economics.

1.2.2 Research Object of Informetrics

Related studies are generally included in the literature of bibliometrics and scientometrics. In the mid-1990s, Buckland provided a comprehensive explanation of information. He concluded that information could have three connotations: information as a process (process information), information as knowledge (knowledge information), and information as things. Information as things consists of data, text, documents, objects, and events. Thus, the research object of informetrics has apparently more contents than those of bibliometrics and scientometrics and focuses more on the quantity of information as things. With its content, including messages, data, events, objects, text, and documents, the research object covers both formal and informal communication things. From the perspective of the generalized informetrics, the quantity of process information and knowledge information should also be included. Tagne-Sutcliffe believed that the definition and measurement of information, which were partly included in information theory, also belonged to the contents of informetrics. Therefore, informetrics comprises a wide range of research objects.

1.2.3 Concept of Informetrics

Informetrics was first proposed by Nacke to outline mathematical applications in all the fields of information science. Afterward, informetrics has been defined as a discipline that describes and studies the phenomenon, process, and law of

information through a quantitative method. This discipline is a new quantitative branch of information science that is based on the combination of mathematics, statistics, and information science.

Informetrics should be understood from the “broad sense” and from the “narrow sense.” The former mainly discusses the measurement of generalized information based on general information theory, which is widely used. Information, material, and energy are the three basic elements of the objective world. Among these, information, which functions as the essential nature of objects, is a system in which we communicate with the outside world through our sense organs to eliminate uncertainties in a system. As noted by Shannon in his paper “A Mathematical Theory of Communication,” the information received by a system will help eliminate uncertainties. Thus, we conclude that similar to material and energy, information is measurable. The key work should be on understanding the quantitative principle and method as well as on determining relative standard. Information is generally measured via the uncertainty degree of a system in informatic circles. Three American scientists, Shannon, Wiener, and Fischer, introduced the basic concept of information as a unit at nearly the same time and arrived at the same conclusion. In particular, Wiener extended the concept of information and proposed the definition and calculation formula for information quantity in his papers “Cybernetics” and “the Human Use of Human Beings,” published in 1948 and 1950, respectively. Fischer studied information measurement problems from the perspective of classical statistical theory. Shannon demonstrated his achievements in detail in his renowned works “A Mathematical Theory of Communication” in 1948 and “Communication in the Presence of Noise” in 1949. He established unified communication theory by considering the abstracted common characteristics of different signals in types of information systems after omitting specific content as a random event, describing information from the perspective of quantity, and conducting research on grammar and transport. The theory helped solve the problems in using different channels for the same and different information transmissions through a single channel. He also presented mathematical formulas for information content and defined them as the amount of eliminated uncertainty; thus, information theory was founded as an independent discipline.

We generally regard information theory as a discipline that aims to study measurement, transmission, and changing laws using mathematical statistical methods. Various inevitable limitations exist in resolving the semantics, utility, and fuzziness of information for the broad sense of informetrics, given that information theory by Shannon mainly focuses on statistical information in the context of the communication field. The good news is that the rapid development of information technology and the expanding research in the field of information have led to the emerging trend of returning to information unity in quantity and quality, syntax and semantics, and transmission and use for its research. For example, Carnap proposed semantic information problems in 1964. Chad published his papers “Fuzzy Sets” in 1965 and “Communication: Fuzzy Algorithm” in 1968, in which he proposed that fuzzy math could be used for information processing, and information theory established based on fuzzy sets can directly reflect semantic information. Belize and

Gaiasu first proposed the unified measurement of information quantity and quality in 1968 by considering the quantity and quality of information, namely, information utility. Sharma generalized the notion of “quantity–quality” to “generalized effective information” in 1978. All of the aforementioned studies and developments will help overcome the limitations of Shannon’s information theory (narrow sense of information theory), thereby leading to and promoting the formation and generalization of information theory. The achievement and progress of generalized information theory will consequently provide necessary theoretical support and specific measurement methods to promote the continuous development of generalized informetrics.

The so-called “special informetrics” is general informetrics that is mainly used to study the measurement of information (or documents). The main content includes analyzing and dealing with the contradictions of the information process by applying mathematics and statistics, studying the dynamic characteristics from a quantitative point of view, and determining the inherent law.

During informetric study, a set of scientific concepts based on quantity, unified indicators, and information quantity units, as well as new quantitative approaches suitable for the nature of the information, must be established to describe the information phenomenon, process, and rules. We, the Chinese scholars, did not achieve progress in this field, and problems also remain unsolved in foreign academic circles. Thus, a number of errors will inevitably occur with the use of only literature measurement at the physical level or only literature information measurement based on literature content and an indirect measurement method at the grammar level. In conclusion, the final resolution for information measurement still depends on the breakthrough in research achievement in generalized information theory and information technology, coupled with arduous efforts from scholars both at home and abroad.

1.2.4 Content Structure of Informetrics

In the narrow sense of informetrics, the structure consists of theory, method, and application, including the following seven aspects:

- (1) Discussion on several basic questions, such as the mathematical description of the information concept, the relationship of the research object, content, and field with relative subjects, coupled with its formation and development
- (2) Fundamental measurement of information by establishing a set of indicators, such as information content, along with a discussion on measurement concepts, such as bit, content unit, information entropy, information field, and information potential
- (3) Study of basic laws, including Bradford’s law, Zipf’s law, and Lotka’s law
- (4) Study of information flow models, such as modeling and evaluating literature growth, obsolescence, dispersion, and citation distribution

- (5) Discussion on a series of quantitative methods, such as frequency ranking, principle of logarithmic perspective, application of fuzzy mathematics, information theory, set theory, and quantitative evaluation of the information utilization rate
- (6) Study from the aspect of the automated implementation of methods and tools, particularly, computer implementation of clustering, correlation analysis, citation database, measurement management information system, and word frequency statistic
- (7) Application to the fields of library and information services, information resource management information retrieval, information analysis and prediction, scientific study, and scientific evaluation.

1.3 Tools and Methods for Informetric Research

1.3.1 *Data Sources for Informetrics*

Data sources of informetrics vary with time. The traditional “10 great information sources” and new ones, such as disc data and network data, are presented in this paper.

- (1) Science and technology book. It refers to a source of information that systematically discusses or outlines special knowledge or subject, including monographs, papers, textbooks, encyclopedias, dictionaries, and handbooks. Its introduction with respect to new achievements, ideas, and technologies is generally not as fast as the presentation of a technology periodical because of its longer publishing cycle.
- (2) Technology periodical. Also called serial, it is a regular or irregular continuous publication of the literature carrier, typically with the same name. Its volume number is set according to the time series, which is divided into a number of periods. Its shorter publishing cycle makes rapid publishing possible, and achievements and level in various fields are reflected well because it has plenty of varieties. The use of a technology periodical as reference is of considerable value in obtaining first-hand information, grasping the progress of a situation, and thinking extensively. The types covered are as follows: ① academic and technical, as the core components of technology periodicals, such as acta, journals, annuals, bulletins, transactions, proceedings, reviews, and progress/advances; ② bulletins, including communications and letters; ③ news journals, such as news and newsletters; ④ data journals, including data and events.
- (3) Technology report. It is a type of literature that reports or records the results or progress of an investigation. It has different forms, such as reports, technical notes, memorandums, papers, bulletins, technical translations, and special publications, as well as names, including primary report, progress

report, interim report, and final report, with regard to their progressing conditions. Famous examples include ASTIA Document report, NASA report, and PB report.

- (4) Conference literature. It refers to documents from academic conferences, which reflect the development trend of science and technology. It is characterized by having a short interval in presenting the latest achievements and its content is less mature than that of periodicals. Conferences can generally be international, nationwide, and regional conferences. Conferences and conference literature comprise common types, namely, conference, meeting, symposium, proceeding, paper, and transaction, such as the proceeding of the “Cold Spring Harbor Symposia on Quantitative Biology” and the special proceeding from the Federation of European Biochemical Societies.
- (5) Patent documents. In countries where a patent system is implemented, all local or foreign individuals and enterprises can fill out an application for unique inventions and submit it to the local or foreign patent office. The office will then go through the application, and individuals and enterprises will possess the right under protection for the inventions for a certain number of years if they pass. Such legally protected technology is what we call a patent. Patents are closely related to industrial activities with extensive practicality. They can be classified into invention patent, utility patent, and design patent according to their technical levels and applications. The patent law of the People’s Republic of China came into effect on April 1, 1985.
- (6) Standard literature. Standardization is mainly reflected in three aspects: (1) product standardization, which refers to a product whose quality complies with the technical requirements; (2) product specification and seriation, which indicate that products should be ranked in size and developed into a series to fulfill a wide range of requirements with minimal variations; and (3) component generalization, i.e., the components of the same type of model, particularly the vulnerable parts, should realize maximum general interchangeability. The standard literature is generally the standard achievements given via an authorized go-ahead, fixed with the form of documents or the basic unit (physical constant), and presented in documentation form.
- (7) Degree thesis. It consists of papers presented to assess the bachelor’s, masteral, and doctoral degrees of undergraduates and postgraduate students. It is characterized by its originality, academic nature, and reference value, presenting the original research results under review. Theses are generally not published separately, and papers with superior quality are presented in the professional periodicals that they are related to.
- (8) Product information. It refers to a product sample, such as product specifications. A useful product information is rich in content, including product specifications, features, patent number, and a variety of useful information about its production.

- (9) Technical file. It refers to technical documents that possess certain engineering objects, which are formed in the production and construction processes and in the technical activities conducted by the technology department.
- (10) Science and technology newspaper. The science and technology literature published in a newspaper mainly refers to news reports about technology. Newspapers report a large amount of information in a timely manner, and thus, they enable readers to acquire important scientific and technological news. Relevant technology columns in newspapers function as windows for us to remain up-to-date with the latest trends both local and abroad.
- (11) CD data. A large amount of data are organized and stored in CDs. Many CD citation index databases are established, such as CSCI, CSSCI, and CSTPC in China and SCI-E and Engineering Index (EI) abroad. All of the aforementioned CD data are significant for us to perform information retrieval and scientific research evaluation.
- (12) Network data. An increasing amount of data are stored and presented on the Internet with the development of networks. Search engines, such as Alta Vista and Google, work as powerful tools to make timely updating and convenient data searching possible. Such data can be classified into digital, text, sound, graphic, and multimedia data according to their content manifestation. Wanfang Data, VIP Data, Tsinghua University's CNKI, CSSCI, CSTPC, and SCI also provide data guarantee and service for measurement and evaluation.

1.3.2 Tools and Application Software for Informetrics

Studies in informetrics mainly rely on domestic and foreign indexing tools, whereas application software mainly focuses on mathematical and statistical tools. Such tools and software are introduced as follows to capture the attention of scholars and to make full use of them in later studies.

(1) Informetric tools

① SCI

SCI is a citation index originally produced by the Institute for Scientific Information (ISI) with the collection of the author, title, source journals, summary, and key words for literature. It helps evaluate the academic value of a literature from the citation perspective and establish a reference network of research subjects rapidly and easily. Since its first publication in 1961, SCI has developed from print only to a multidisciplinary and comprehensive retrieval system with electronic, networked, and integrated features. SCI is divided into SCI, with journals coming from over 3800 types of printed and compact disc editions (SCI CDE), and SCI-E, a larger

version that covers over 6000 notable and significant journals. The index is available online through the international online service platform and the Internet. Moreover, over 150 disciplines are included in SCI, which mainly focuses on agriculture, biology and environmental science, engineering technology and applied science, medicine and life sciences, physics and chemistry, and behavioral sciences.

② Social Sciences Citation Index (SSCI)

SSCI is an interdisciplinary citation index product developed by ISI from SCI. It conducts statistical analysis on the quantity of social science papers collected from different countries and areas. This citation database covers 1700 of the leading social science journals worldwide across over 55 disciplines, including anthropology, law, economy, history, geography, and psychology. The types of literature collection include research papers, reviews, discussions, editorials, autobiographies, and letters.

③ Arts and Humanities Citation Index (A&HCI)

A&HCI is a comprehensive arts and humanities journal database with indexing for over 1100 journals produced by the ISI, and disciplines covered include language, literature, philosophy, Asian studies, history, and art.

④ Essential Science Indicators (ESI)

ESI is a basic analysis and evaluation tool for measuring scientific performance and tracking development trends. It has been funded by ISI since 2001. ESI also serves as an informetric analysis database that covers over 10 million bibliographic records of 8500 academic journals worldwide collected by SCI based on ISI, SSCI, and A&HCI. It is available through ISI Web of Science as an important part of the ISI network integration service platform. ESI conducts statistical analysis and sorting of countries, research institutions, journals, papers, and scientists from 22 areas of expertise from the citation analysis perspective by considering the indicators of indexed papers, citations of papers, and the average citation per paper. Users can clearly understand the influence and development in a given subject areas of scientists, research institutions (universities), countries (cities), and academic journals ranked in certain positions the from database to identify key scientific discoveries, conduct performance assessment studies, and master science development trends. ESI is also available for analyzing international academic literature systematically and precisely. As part of ISI Web of Knowledge, ESI provides a dynamic, integrated, web-based research and analysis environment for scientific research.

⑤ EI

EI, founded in 1884, is a renowned comprehensive search tool that specializes in engineering and technical science published by the American Engineering Information, Inc. With a selection of over 2000 types of engineering and technical journals worldwide, the recorded literature covers nearly all areas of engineering and technology, including power, electrotechnics, electronics, automatic control,

mining, metal technology, machinery manufacturing, construction, and water conservation. EI is characterized by comprehensiveness, extensive sources of information, geographic coverage, numerous and high-quality reports, and authority. Over 3500 types of scientific and technical journals and 1000 types of conference proceedings, papers, academic presentations, science and technology books, almanacs and standards, and other publications worldwide are involved in EI, with annual reports of 500,000. Publication editions include EI CompendexWeb, EI Compendex, EI Microfilm, EI Compendex, and printed versions (including annual and monthly editions).

⑥ Index to Scientific and Technical Proceedings (ISTP)

ISTP was founded in 1978 and edited and published by the American Society of Scientific Information. Its conference proceedings cover life sciences, physical and chemical sciences, agriculture, biological and environmental sciences, engineering, and applied science, with its literature accounting for 35%.

⑦ CCSI

CCSI, developed by the National Science Library of the Chinese Academy of Sciences, covers 669 types of the most important scientific journals published in China across various fields, including mathematics, physics, chemistry, aeronautics, geography, biology, agriculture, medicine, and engineering technology. Approximately 2,220,000 Chinese citations are involved in the annual report literature (approximately 710,000). The current publication includes a print version and a CD version.

⑧ CSTPC

CSTPC, developed by the Institute of Scientific and Technical Information of China (ISTIC), covers over 1600 types of Chinese science and technology core journals across various disciplines of natural sciences. The database combines document retrieval and statistical analysis, as well as contains functions, such as finding important scientific papers published in China, determining the statistical analysis and ranking results of Chinese scientific papers over the years, and clarifying the details of publications from various regions, departments, units, authors, disciplines, as well as of fund-published papers.

⑨ CSSCI

CSSCI, produced by Nanjing University and the Hong Kong University of Science and Technology, provides a wealth of data across various disciplines in the field of social science. It is available through the subsystem of statistical analysis, such as author-issued statistics, agency-issued statistics, regional statistics, issued documents of discipline distribution statistics, books, journal-cited statistics, publisher-cited statistics, author-cited statistics, and paper-cited statistics. Each statistic can be performed separately according to discipline.

⑩ Chinese journal full-text database and its citation report

This database is currently the largest continuously updating Chinese periodical full-text database worldwide, with an accumulated 8,000,000 full-text documents, over 15 million titles, 9 sub-albums, and 126 thematic literature databases. It owns a collection of 6100 full-text core journals and professional journals, covering Technology A (basic sciences), Technology B (energy and material chemistry), Polytechnic C (industrial technology), agriculture, medicine and health, history and philosophy, economics, politics and law, education and social sciences, electronic technology, and information science.

(2) Informetric application software

① Matrix Laboratory (MATLAB)

With Mathworks developing a set of calculations, graphical visualization, and editing functions into a language that is powerful, easy to operate, and easy to expand, MATLAB has become one of the outstanding internationally recognized mathematical application software. Its powerful system consists of its core contents (language system, development environment, graphics system, math library, and application programming interfaces) and auxiliary toolbox (symbolic computation, image processing, optimization, statistics, and control toolbox). We are currently using a new version of MATLAB6.

② Statistical Analysis System (SAS)

SAS was originally developed by two graduate students at the North Carolina State University, and the company was founded in 1976. SAS exhibits comprehensive capabilities that include data access, data management, and data analysis, and is recognized as the standard software for data statistics and analysis in the international arena. SAS is a software system that features a module-combined structure with over 30 functional modules. Written in the assembly language, SAS typically requires programming that is more suitable for professional statistics personnel but difficult to use for nonprofessionals. SAS 9.2 is the newest version.

③ Statistical Package for the Social Sciences (SPSS)

SPSS, SAS, and Biomedical Programs are the most influential major statistical software packages worldwide. Although named SPSS to emphasize its applications in social science (random phenomena may occur in social science research, and thus, studies require statistics and probability theorem), it also plays a significant role in various fields of natural sciences, being applied to a wide range of subjects, including economics, biology, education, psychology, medicine, sports, industry, agriculture, forestry, commerce, and finance. SPSS includes functions such as completion of data entry, editing, statistical analysis, reporting, and graphic production, with 136 functions under 11 types. SPSS also provides a simple description and complex statistical multivariate statistical analysis methods, such as exploratory data analysis, statistical description, table analysis, 2D correlation, rank correlation, partial correlation, ANOVA, nonparametric test, multiple regression,

survival analysis, covariance analysis, discriminant analysis, factor analysis, cluster analysis, nonlinear regression, and logistic regression. SPSS is preferred by non-statistical professionals over SAS. The latest version is SPSS 22.0.

1.3.3 Method Systems for Informetrics

Informetrics comprises a range of methods, such as statistical analysis, mathematical model analysis, citation analysis, word frequency analysis, co-occurrence analysis, cluster analysis, and computer-aided analysis.

(1) Statistical analysis

Statistical analysis consists of two parts. The first part comprises a number of professional statistical terminologies, such as “collection volume,” “liquidity,” “citation amount,” “difference coefficient,” and “word frequency analysis. The second part comprises index statistics, including random sampling, sampling distribution, parameter estimation, hypothesis testing, regression analysis, variance analysis, and cluster analysis. In summary, mathematical statistics perform research on two aspects: (1) research about sampling methodologies, i.e., how to draw up a sample from a population, including the number and appropriate drawing methods; and (2) research on how to conduct a reasonable analysis of results (sample data) and to make scientific inferences. The birth and development of mathematical statistics are partly attributed to the limited ability of people to deal with too much data at one time, and the method is suitable for handling random statistical objects. The specific content will be discussed in Chap. 8.

(2) Mathematical modeling approach

The mathematical modeling approach, which is a common effective analysis method, is one of the most basic and important methods for informetrics given its quantitative nature and reliability in quantitative analysis. A mathematical model is built from the structure and behavior of a system described by a mathematical language. Apart from simulating the structure and movement of the system from a macroscopic point of view, the mathematical modeling method consists of refined and accurate characteristics because it mainly describes system structure, contact, and movement law in a microscopic and quantitative manner. A mathematical model is divided according to different criteria and methods, such as in terms of expression form and the categories of analytic models and images, model equations, and graph mode. The basic steps in creating a mathematical model include ① target determination, ② raw data collection, ③ theoretical model establishment, ④ parameter determination, ⑤ theoretical model verification, and ⑥ forecasting and decision making. Mathematics, as an important basic discipline, will play an increasingly important role in the future development of informetrics, as elucidated in Chap. 9.

(3) Citation analysis

Citation analysis is a bibliometric analysis method that aims to present quantitative characteristics and inherent laws after conducting a specific analysis of citations and the cited phenomenon of objects, such as scientific journals, papers, and authors, using mathematical and statistical methods as well as various logic methods, such as comparison, induction, abstraction, and generalization. For decades, citation analysis has developed into a wide range of applications, while constantly enriching its technical methods. Three study types are currently involved: ① studies based on quantity; ② research related to network or link relationship among citations, such as co-occurrence analysis; and ③ research on topic relevance reflected from citations. Many other types of citation analysis can also be derived from other features of citations, such as language, country, time, and author. In conclusion, as a unique approach for informetrics, citation analysis has been receiving considerable attention, has been applied widely, and is expected to develop extensively, particularly in information retrieval and information prediction, as indicated in Chap. 10.

(4) Computer-aided informetric analysis method

With the popularization and application of information technology and the increasing popularity of computers, research methods for informetrics have gradually transformed from manual statistical analysis into a computer-aided stage with a clear trend toward automation. Since the 1990s, the increasing number of studies on the computer-assisted analysis of informetrics have made considerable achievements and applications. Computer-aided information measurement research based on theoretical analysis has emphasized software design and development work to achieve a reestablished library that covers recorded data and statistical analysis on multiple data. The establishment and improvement of computer-aided measurement and analysis methods indicate that the method system has been basically formed, as evidenced by the successful development and publication of a series of databases, such as the CD-ROM or network version of the China Science and Technology Papers and Citation Database, the Chinese Science Citation Database, and the CSSI Database, which provide necessary tools and conditions for auto-statistics and literature information analysis, thereby significantly facilitating the automation development trend. Chap. 11 presents the details.

1.4 Informetrics and Related Disciplines

1.4.1 *Related Disciplines of Informetrics*

Science-of-science research indicates that science itself is interrelated, and each discipline is more or less related to others within the field of science in accordance with the principle of unity. Such interconnection and penetration tend to develop deeply and commonly at present.

Informetrics is inevitably connected to many other disciplines. First, informetrics is a novel frontier science among literature science, information science, and library science according to the nature of this discipline. Second, informetrics is related to both natural science and social science, with the process affected socially and psychologically by objective factors, such as science and technology, and by human-controlled subjective factors. In addition, informetrics is becoming increasingly close to other subjects because, on the one hand, informetrics promotes its development by utilizing the knowledge, achievements, and methods of other disciplines, and on the other hand, certain theories and methods of informetrics are also required in other disciplines. Relevant disciplines with close connection include library science, literature science, information science, mathematics, statistics, bibliometrics, scientometrics, computer science, science of science, prognostics, science and technology management, and science evaluation.

1.4.2 Relation Among Informetrics, Mathematics, and Statistics

Informetrics has a close relation to mathematics. From the perspective of its development process, informetrics is formed and developed based on a combination of mathematical methods and applications in the information research field. During the period of modern science, literature science and information science have increasingly adopted mathematical methods in researching and solving problems under the influence of the quantitative trend of scientific research. From the statistics of 589 papers prepared by O. Nacke, the existing 22 branches of mathematics have been applied to 45 works in the field of information science. Table 1.2 presents the different mathematics and information science fields involved in bibliometrics, which indicates that despite the wide range of involved mathematics and information science fields, mathematical methods are applied to nearly every aspect of information work. Informetrics is developed based on such a wide combination and cross connection of mathematics to information science and library science.

From the viewpoint of research content and methods, informetric study is inseparable from mathematical tools. Mathematical tools are prerequisite in the study of quantitative relationship between the various factors of information and its carriers. With regard to methods, as the most powerful tool for quantitative analysis, mathematics also fulfills the requirement. The development of modern mathematics, particularly the emergence of new branches of fuzzy mathematics and stochastic mathematics, provides a practical possibility for applying mathematics to the field of information. One of the main trends in informetric development is to use and combine informetrics with mathematics to extend its depth and breadth given that the application of mathematical methods has become an important symbol in measuring the development level of informetrics. That is, informetrics will not exist

Table 1.2 Mathematics and information science fields involved in bibliometrics

Mathematics	Statistics	Integral calculus
	Analytical statistics	Differential equation
	Experimental statistics	Vector analysis
	Multivariate statistics	Matrix theory
	Measure theory	Graph theory
	Combinatorial	Linear regular
	Probability (theory)	Cluster analysis
	Set theory and mathematical logic	Queuing theory
	Information theory	Game theory
	Algebra	Simulation method
	Differential calculus	Fuzzy mathematics
Information science	Introduction to information science	Putaway and filing
	Information concept	Information retrieval
	Information flow	Searching
	Information channels	Lending
	Information processing	Replication
	Transmit information	Translation
	Work object	Printing and publication
	General literature	Work plan working programming
	Periodical/journal	System design
	Special issue	Space design
	Work system	Financial planning
	Library	Management method
	Literature library/archive	Cost–benefit analysis
	Information system	Quality analysis
	Working process	Disadvantage analysis
	Registration	Measured value
	Selection	Similarity (degree)
	Duplicate processing	Correlation degree
	Classification	Aging degree
	Numbering	Literature growth rate
	Collection	Library loss index
	Working means	Information demand index
	Index	Analysis method
	Directory/catalog	Bibliometric method
	Classification (method)	Scientific metrology method
	Retrieving dictionary/thesaurus	Citation analysis
	Processing and circulation circulation	Index analytical method
	Cataloging	

without the application of mathematics in the field of information, and in turn, the rise and development of informetrics provide a new arena for mathematics, with its unique research material and achievement-enriching mathematical databases.

1.4.3 Relation Between Informetrics and Bibliometrics

Informetrics and bibliometrics are evidently distinct in spite of their similar research contents given the aforementioned concept and content structure of informetrics.

First, bibliometrics mainly concentrates on documents and information carrier (sometimes also contains literature information), whereas informetrics focuses on the measurement of information itself. With reference to the threeworld theory by Popper, British information scientist Brooks sorted pure technology literature as world 1 (the world of physical objects and events) and technology knowledge as world 3 (the world of objective knowledge). Documents measured as articles, volumes, and books certainly belong to world 1, whereas, knowledge and information, as intellectual products, belong to world 3. However, people typically measure by literature instead because of the inseparable relation between literature and information as well as the lack of rational standards and methods for information measurement. For example, quantitative research on the laws of knowledge growth and efficiency evaluation of information retrieval system are based on literature quantity.

Second, informetrics is more extensive than bibliometrics given that other forms of information apart from document information exist, as evidenced by the appearance of “zeroth information,” which refers to information without carrier or produced recently through an informal communication process in the information field. A comprehensive study of literature information measurement is necessary and its laws should be determined, which will provide a good foundation for quantitative research in a broad sense. Therefore, the progress of bibliometric study is also a contribution and promotion of informetrics.

Third, with regard to tasks and methods, suitable quantitative methods for informetrics are difficult to determine. Under this context, we have to find new ones in accordance with the nature of information. Information is a social phenomenon related to the understanding and judgment processes. It is subject to the effect of human behavior, such as subjectivity and personality, which overtakes mathematical methods and tools that are generally adopted in the physics world, and thus, makes the related study difficult. Various problems, including finding the perfect methods and approaches, have not yet been involved and resolved.

During informetric study, we should establish a series of scientific concepts based on “quantity,” new quantitative approaches, and an extensive research field to describe the information process and its rules. Informetrics will then gradually develop into an integrated and quantitative discipline. As indicated by British information scientist C. W. Hanson, science information is a synonym for literature

for many people, and the term “bibliometrics” can still be used largely based on the content and achievement of informetrics.

The quantification of information science has been an inevitable trend in subject development, and we need to include the study and research of informetrics in our agenda in due course and make assiduous efforts to promote the development of informetrics.

1.4.4 Relation Between Informetrics and Scientometrics

Scientometrics is a new field in science, and one of its particularly active and potential branches. The academic term “наукометрия” in Russian, which is defined as a subject that studies the science of science using quantitative methods, or rather, the summation of all the quantitative methods used to study the scientific development process, was first proposed by famous Soviet scientist F. Narimov and colleagues in 1969.

Informetrics and scientometrics have connections and differences. Scientometrics aims to explore the inherent law of scientific development from the quantitative perspective, and ultimately, promote the progress of science and technology. Informetrics aims to determine the law of scientific literature itself, facilitate the scientific nature of information management, and determine the laws of science and technology by studying information law. Evidently, both fields aim to explore the internal relations among scientific activities and consider the characteristics and laws of scientific development as among the basic tasks in their research.

The two fields are similar in terms of research object and method. In particular, research on scientific documents and their quantity are involved on both sides, and they both conduct analyses using quantitative mathematics and statistical methods. Informetrics and scientometrics cross with each other in a large field with regard to their research content.

However, significant differences still exist between them. First, informetrics has a larger coverage than scientometrics. Scientometrics is limited in scientific documents, quantity, messages, events, things, and objects of informal communication. For informetrics, formal information is also included in the narrow sense, and process information and knowledge information are even involved in the broad sense. Second, they position themselves differently because informetric research serves the demand of information science research, whereas scientometrics aims to study and verify the quantity law of production, transmission, and utilization of science. Third, they have distinct application fields. Informetrics plays a pivotal role in evaluating literature, personnel, discipline, unit, and data (including network), whereas scientometrics exhibits importance in technology-related policy making.

The preceding analysis emphasizes that scientometrics will inevitably develop toward informetrics given its relatively simple content and purpose, which have already been involved in informetric studies.

1.4.5 Relation Between Informetrics and Webmetrics

Webometrics or cybermetrics, which has developed rapidly under the current specific scientific background and technical conditions, is an interdisciplinary subject formed by a combination of network technology, network management, information resource management, and informetrics; it is also a new direction and important research field in informetrics. Webmetrics has been widely used in the international academic community since its proposal in 1997 and has received concern from social sectors. With the rapid development of information science and information technology and the popularization of the Internet, information resources have evolved from tangible written materials to electronic and digital forms on networks, which have rapidly increased the interchange activities of netinfo. Consequently, informetrics has further developed into webmetrics given that the original metric indicators are no longer applicable to netinfo measurement. The revolutionary reform of scientometrics, bibliometrics, informetrics, and technometrics in the new information network era essentially introduced webmetrics. From the words of Professor Egghe of Belgium, “the new Internet virtual world has posed a challenge to informetrics.” Hence, well-known laws of scientometrics and informetrics have been redefined as “classical laws” in the new generation. We need to adapt ourselves to the new research field of exploring the new distribution laws and characteristics of network information resources, which will be a complicated and systematic but meaningful project. In summary, webmetrics is the legacy and further continuation of informetrics; it is a newly developed subdiscipline that is adapting to a new network environment.

1.4.6 Relation Between Informetrics and Scientific Evaluation

In the 1990s, scientific evaluation was not only an important issue of the scientific system itself, but also a crucial subject for the government and the society. Increasing research has focused on which method to use, how to construct a set of complete scientific evaluation system, how to develop a normative system, and how to establish a scientific evaluation institution competently and independently. We believe that scientific evaluation should be understood from two levels. Scientific evaluation refers to evaluations in science in the narrow sense. In the broad sense, it is considered conducting evaluation using scientific methods, i.e., scientizing it. From the current practice, scientific evaluation has extended from the evaluation of a specific subject in the fields of natural science and social science to a larger field of organization, product, service, industry and industry-like university, enterprise competitiveness, journal, information resource, management information system, and site evaluations. The broad sense mainly includes the following types: (1) evaluation of scientific publications or literature evaluation, including the

evaluation of theses, journals, books, and patents; (2) institutional evaluation, including the evaluation of scientific research institutes and universities; (3) scientific research evaluation, including the evaluation of scientific research (in terms of its science, advancement, and applicability), scientific projects, scientific achievements, input and output ratios, and the efficiency of scientific research; and (4) subject evaluation, including the evaluation of the development stage, status, level, prospects, structure, and correlation of a subject.

Various methods for scientific evaluation have formed a derivation, with qualitative evaluation based on peer review and quantitative evaluation. Informetric theory and method are widely applied to the quantitative science evaluation process. Peer review and metrological analysis methods are always interconnected to a certain extent rather than completely independent.

Chapter 2

Literature Information Growth Law

2.1 Characteristics of Literature Information Flow and Meaning of Growth Law

2.1.1 Characteristics of Literature Information Flow

In informetrics, the flow of information in a document is typically called literature information flow. Literature is the most basal carrier of intelligence. Document information flow is a collection of scientific literature with the characteristics of a series of themes. Literature information flow is sometimes called literature flow, for short.

Literature information flow has many features that are divided into two aspects, namely, static and dynamic characteristics. Scientific literature in the spatial distribution of properties within a certain period is called the static characteristic of literature information flow, such as concentration–discrete distribution, literature distribution according to the author, word distribution in the literature, citation distribution, and topic distribution regularity.

The continuation of scientific literature over time and the nature of growth and aging denote the dynamic characteristic of literature information flow. Scientific literature both grows and ages, i.e., aging in growth and growth in aging. Growth is the main trend of literature information. A famous former Soviet intelligence expert, А.И.Михайлов, said that “At present, the growth of published articles, aging and discrete, is rightly deemed the most fundamental law of development of the scientific literature.”

The growth of scientific literature mainly refers to the number. The number increases with the growth of the number of times. In modern science development, the basic situation of literature growth performance for scientific literature is increasing at an annual rate of 6% to 8%. Approximately every 10 years, the number of scientific literature will be doubled. The literature published over the past

20 years is more than 2000. The time in which 1 million chemical abstracts (CAs) will be published in the United States is continuously shortening, i.e.,

First 1 million 32 years (1907–1938),
Second 1 million 18 years,
Third 1 million 8 years,
Fourth 1 million 4.75 years,
Fifth 1 million 3.3 years...
1 million at present, only 2 years.

Such information is a powerful indication of scientific literature surge. Scientific journals in the world comprise 100000 types, nearly 1 million types of books are circulated, and printed literature amounts to over 10 million in 1 year. This “information explosion” is a growing trend.

2.1.2 Influence of Literature Information Growth and Countermeasures

The rapid growth of scientific literature is an objective social phenomenon. The effects of and countermeasures for this phenomenon are important subjects to be explored.

(1) Surge of literature information

The rapid growth of scientific literature causes many problems and significantly influences the development of scientific research and literature collection, management, and utilization. The increasing scientific literature leads to overcrowded book libraries and intelligence agencies, accompanied by lack of funds and human resources. The collection and proper storage of such literature are difficult. Many intelligence service measures are also difficult to implement, thereby directly affecting intelligence efficiency and development. The working time of science and technology personnel is considerably increased. When they scan the literature, they are unable to check and read all the required information documents. A scientist can only read 5% of the professional literature. Literature quantity is significantly large, and thus, recall and precision of the required information are difficult for the personnel. The language barrier of reading literature is also increasing, thereby causing many difficulties in scientific research as well as extensive repetition and waste of scientific research. Experts estimate that nearly 40% of domestic research projects conducted by scientific research departments have foreign counterparts. This repetition does not only cause considerable waste, but also seriously affects scientific research efficiency and science development.

The corresponding loss is incalculable. A rough estimate made by a Soviet intelligence member indicated that if science and technology intelligence can be effectively used, then we will save 60% of the funding, on average, and shorten

research time for 2 years to 3 years. The considerable drop in “information explosion” intelligence literature utilization causes information loss by up to 20–80%.

(2) Countermeasures for literature information explosion

People apply basic countermeasures to adapt to the scientific literature information explosion situation via the following two aspects.

In theory, the growing number of scientific literature is the main cause of “information explosion” to strengthen the study of literature law. Therefore, actively implementing laws of growth and aging of the literature, which theoretically proves its inherent regularity, is one of the basic countermeasures to overcome the intelligence crisis. Research can be a reliable basis for the management and optimization of scientific literature and can guide the present stage of scientific literature, thereby making it more adaptable to the current literature surge. On the basis of new knowledge, countermeasures against literature growth have been proposed from the perspective of literature quality concept. For example, in the “International Federation of Literature, the 40th Annual Meeting” held in Copenhagen, Denmark in August 1980, American literature scientist H.R. Brinberg introduced the “more with less” intelligence collection principle, instead of the previous “the more the better” or 100% of the collection method. The proposal received the common response of the book intelligence community. Simon emphasized “content analysis” to establish related contact. Professional literature will not achieve unlimited expansion because of differentiation in subjects. Therefore, increasing the degree of specialization can reduce the burden that comes with literature growth to scientific and technical personnel.

With regard to technology, modern advanced technology and equipment, such as computer, are used to process and utilize document information. “As a result of the sharp increase in the scientific literature, library and intelligence agencies are facing great practical difficulties, and we should use machines to properly deal with the literature as soon as possible,” as stated by the founder of cybernetics, Wiener. Under this idea, the test and research for computer information retrieval began in the 1950s. Many countries in the world are presently using new highly efficient document-processing technologies, computer information retrieval systems, and information management integration systems, all working with automation. These advancements have all achieved evident effects. The specific measures include using new information carriers (magnetic tape, miniature flat piece, disks, and network), carrying out machine translation, setting up a computer literature information database, modifying online information retrieval, developing a network for literature information and knowledge, and establishing information and knowledge services.

The current situation requires China to take countermeasures. First, the formulation of long-term reasonable planning and development strategy, the creation of necessary conditions for forward modernization of intelligence, such as constructing and improving the literature intelligence agencies at all levels of various

types, and the establishment of a national intelligence network are crucial. Second, the retrieval procedure for domestic literature of science and technology systems should be improved as soon as possible, the law of document information should be explored extensively, and the information management of scientific and quantitative data should be accelerated. Finally, we should pay attention to taking practical measures to improve current intelligence work. A scientific quantitative method is used to select book journal literature; optimize collection; accelerate the construction, popularization, and application of characteristic database and information network; and provide readers with high-quality services.

2.1.3 Research on and Significance of Literature Information Growth Laws

(1) Significance of literature information growth rule research

The problem of scientific literature growth law has been a concern of the intelligence community for a long time and one of the main research subjects of metrology.

Research on the laws of scientific literature growth is of considerably important both in theory and in practice. The determination of the relationship between the sum of scientific literature changes and time can approximately present some characteristics and laws of science development. The intelligence model, which indicates the scientific prediction of a number of changes in accordance with the relevant literature, is widely used in intelligence analysis. We can also predict literature growth tendency based on the study of the rule of scientific literature growth, thereby providing a decision-making basis for the development of science and intelligence work in the future. Research on the rules of the growth of the scientific literature is not only an important theoretical problem of bibliometrics, but can also directly work for intelligence services, and therefore, help deal with the increasingly serious intelligence crisis. A Soviet intelligence expert, И.Б. Маршакова, indicated that bibliometric research involves using information, including showing both the entire field of science with its individual inherent parameters and regularity. Such research is essential to optimize management when determining and implementing a research plan and is necessary in intelligence works.

(2) Status quo of literature information growth rule research

People have been exploring the law of scientific literature growth for a long time. As early as the beginning of the 20th century, several scholars began to study the total number of scientific literature. In the 1940s, the law of scientific literature growth attracted the attention of many researchers as a theoretical problem. Scholars have then made a series of research achievements in this field and

proposed various theoretical models to describe the law of literature growth. When D. Price introduced the rule of scientific literature exponential growth, a breakthrough has been made in this issue. Considerable statistical research and hundreds of papers are being circulated in Europe and the United States, and a number of monographs are being published in the science and information science fields in the Soviet Union. In China, the study of literature growth law remains limited and urgently requires strengthening. In a world scale, however, the problems in scientific literature growth regularity are among the most active subjects in the field of informetrics.

Research on the laws of relevant scientific literature growth mainly concentrates on two aspects. The first aspect is theoretical research, which focuses on establishing accurate mathematical models and theoretical explanations to further determine the growth regularity of scientific literature. The second aspect is using literature growth law to guide actual intelligence and information management and using a literature quantitative index to measure knowledge and present an application in the law of science development research. A famous intelligence figure, А.И.Михайлов, stated that a number of regulars had been gradually revealed, which marked the inner link between scientific publications and scientific development as well as the quantitative relationship in the number of articles published and the scientific growth index.

2.2 Growth of Science Knowledge and Scientific Literature

The growth of scientific knowledge is closely linked to the growth of scientific literature and its law. Therefore, when discussing scientific literature growth law, we should introduce scientific growth regularity of utility at the beginning. This introduction is necessary and helpful to fully understand literature growth law and mathematical models and correctly understand the law of literature growth theory.

2.2.1 *Growth of Scientific Knowledge*

After World War 2, science and technology have been rapidly developed and experienced a profound revolution. Science has developed into the “big science” era, namely, the period of modern science. The development of modern science is swift. One of the main performances is that the knowledge of human beings is rapidly increasing. Western countries show that human knowledge has quadrupled from the beginning of era of human being to the 1960s Table 2.1.

The double cycle of science learning becomes significantly shorter. The increasing and rapidly changing knowledge of humans has markedly expand

Table 2.1 Situation of increasing scientific knowledge

First double	Beginning of era–1750	1750 years
Second double	1750–1900	150 years
Third double	1900–1950	50 years
Fourth double	1950–1960	10 years

science at a rapid pace. Scientific results considerably influence the society and have exceeded the total results in the past 2000 years.

Over 100 years ago, Engels said that “the development of science is proportional to the legacy of utility that the previous generation leaves.” This view demonstrates the law of index development. Modern history also confirms that the numerous indicators in the field of science are according to the law of index growth. The mathematical language description is as follows:

$$W = \alpha e^{\beta t}, \quad (2.1)$$

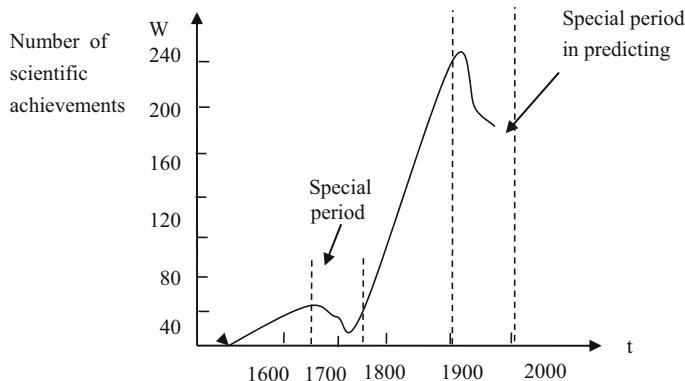
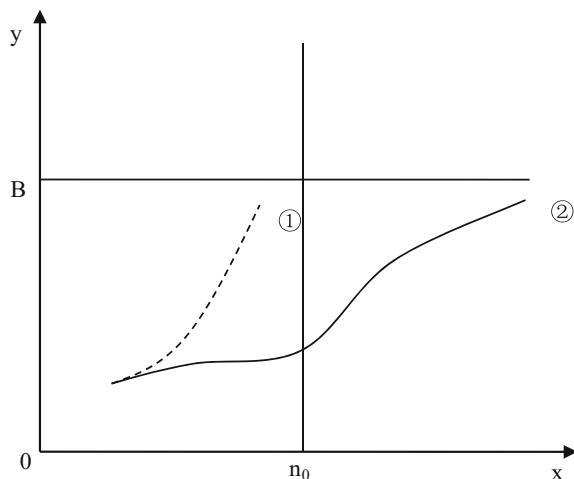
where W is the scientific index, alpha and beta are arbitrary constants, and t is time. Formula (2.1) is called the law of science and utility index growth. Some people also call it the law of science accelerating development.

The exponential law of science development was an important finding in the study of bibliometrics and science in the 1940s. The conclusion was based on the growth of SCI. D. Price researched and found the exponential growth trend of scientific journals. F. Ryder studied the growth of science books in the United States via statistics. In his book “the Science since Babylon,” he described the speed of scientific development and obtained the result that scientific development is increasing exponentially by considering science magazine and academic papers as important measures of utility.

The law of scientific development index growth has resulted in academic debates. One of the most important questions is that if the index of scientific indicators keeps increasing, then the number of scientists will be more than the total population of the world. In fact, this view is wrong because any scientific law is established approximatively under certain conditions. Exponential law is a law within a certain historical period. The “times” of scientific development are likely to be damaged at another time.

Figure 2.1 clearly shows that the rule of the scientific development index is broken twice in the special period since 1550; the first time was in 1670–1740, and the second started in 1940 and continues until today. Evidently, the index law was disrupted in a certain historical period but would be set up in another historical period

The index law is constantly set up and destroyed because the changes in the timeline are under the interaction of the accumulation model and specification for change. The former is characterized as the accumulation of utility, i.e., the so-called index development law. The latter is characterized by a qualitative leap, i.e., the so-called scientific revolution. The accumulation between normative and change is both opposing and unified. In the period during which standard accumulation

**Fig. 2.1** Scientific development index curve**Fig. 2.2** Scientific development logic curve

works, scientific development follows the exponential law. When change specification works, the exponential law does not indicate anything. Consequently, the destruction of the exponential law is not a disastrous event, but one of the important signs of the arrival of the scientific revolution.

Another question about exponential law originates from the “saturation phenomenon” of scientific development, which is also called the “S”-shaped law of development. A few scholars abroad have held the point of view that any type of growth according to the index must balance at a certain point; otherwise, we will fall into fallacy. They have portrayed the entire history of science as a large “S” shape and supposed that index curve development would inevitably turn into a logical curve (Fig. 2.2). The contemporary period is on the balance point (point n),

and it will be close to the saturation limit in another 30 to 45 years. Scientific growth is expected to stop by that time.

We believe that this statement is partial. The “S”-shaped phenomenon of scientific development is a historical fact, but is not always evident on the timeline. As shown in Fig. 2.1, the “S”-shaped phenomenon appeared at around 1670. Simultaneously, the index law occurred at “extraordinary times”. In the 18th century, scientific development presented a steeper index curve. In fact, the “saturation phenomenon” in science development is not surprising; this phenomenon is the nature of and a common regularity in human society. After a period of “saturation,” a new period characterized by the considerable acceleration of exponential growth will occur. The step exponential law undergoes “acceleration–saturation–broadening.” Science exhibits the relationship between growth of scientific knowledge and the growth of the scientific literature. The step exponential law does not only exhibit dialectical unification between quality and quantity in scientific development, but also suggest that the development of science and technology is endlessly advancing.

2.2.2 Relationship Between the Growth of Scientific Knowledge and the Growth of Scientific Literature

Scientific literature provides an objective record of scientific knowledge. The rapid development of science and technology will significantly increase scientific utility. All types of scientific knowledge should be recorded, preserved, and accelerated in the form of literature.

The number of scientific literature is one of the important measures of science and utility because changes in the number of scientific literature directly reflect scientific utility changes. As the main carrier of scientific information, changes in the number of scientific literature are also important symbols of science development. The number of science and technology books, academic publications about science, and scientific paper topics, such as the number of proportional literature volume, is frequently used as scientific indicators to reflect the scientific law of development in the study of science. The identification of various characteristics and laws of scientific development by analyzing the total number of relevant literature is the history of science, and scientific methods are frequently used in such study.

The growth trends of scientific literature and science knowledge are generally synchronous. Their growth rules also exhibit similarities to a considerable extent. Studies on the laws of scientific literature growth and science knowledge growth frequently cross and promote one another. The identification of scientific literature growth rules can provide the basis for scientific research on the laws of utility growth, which also helps deepen the understanding of literature growth rules. For example, D. Price proposed the theory that famous scientific knowledge grew

according to the index law of growth theory. One of the main bases of this theory is that scientific literature is based on the index law of growth. Thus, a close connection exists between the growth of scientific literature and that of science learning.

2.3 Exponential Law of Literature Information

2.3.1 *Indicators and Methods of Literature Information Measure*

The indicators and methods of literature information measure should be determined in literature quantitative research. Literature indicators can be classified into absolute value indicators and relative indicators. An absolute value index indicates the document number, such as the number of books, periodicals, and publications; a relative index indicates the proportion of different parts, such as the proportion of each type of literature and the proportion of each language literature.

Research on the laws of literature growth is generally based on literature accumulated data, especially in the research on the growth law of literature in a particular field or subject during a certain period, because the number of published literature and the cumulative number of literature always increase annually, which is likely a fixed rule. The results typically exhibit a rule curve, which can be described by an accurate function, and thus, are advantageous to the quantitative analysis of literature. The number of published literature each year is easily affected by various complicated social factors, and thus, is always fluctuating. Whether the number is similar to a set of rules is difficult to determine. The curve of the result has no rule and is difficult to describe in a certain function. All the aforementioned problems result in difficulties in the quantitative analysis of literature.

Two approaches can be adopted to measure the growth of scientific literature. The first approach is based on the total accumulated quantity of published literature each year, whereas the second approach is based on newly published literature quantity each year. The former focuses on utilizing the growth of the amount of literature, whereas the latter evaluates how much research literature is increased or decreased every year based on the number of new literature. Research on the rules of scientific literature growth is mostly from cumulative data but does not exclude the non-accumulated data of the research methods. The growth of scientific literature is shown by the increasing number of new scientific publications, and therefore, is typically measured by the number of published books each year. The growth curve generated using the measurement method can clearly reflect the change tendency of new literature every year. Although this irregular curve is difficult to describe using the quantitative description model in most cases, this approach is still frequently used to illustrate the problem in the study of intelligence analysis.

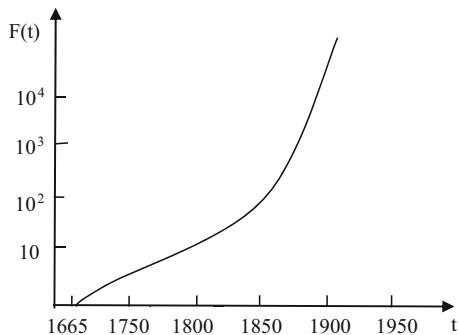
The results obtained using different methods and indicators to study the growth law of scientific literature vary. The examination of specific disciplines in the total number of journals published over a given period presents the law of exponential growth. However, as A.I. Михайлов indicated, “in judging the progress of scientific literature growth by the number of publications in each year, it is not even geometry but merely arithmetic.”

2.3.2 Literature Information Index Growth Model

As early as 1944, Fremont Ryder, a librarian at the Wesleyan University, studied the library collection rate of representative universities in the United States. He found that the main university library collection increased at the average of 16 years based on a huge amount of statistics. Thereafter, Derek de Solla Price, a famous scientist and intelligence expert, popularized and applied this discovery to all fields of scientific knowledge and conducted a series of research. In 1949, he found that stacks of philosophical transactions that stacked against the wall could become a perfect index curve. In the following year, he published the first research paper on exponential growth. In 1959, he presented a lecture series on scientific index growth at Yale University and officially published it in 1961 with the title “Science since Babylon.” In this classic scientific work, Price emphasized that the first scientific magazine in the world was “The Philosophical Transactions of the Royal Society of London,” which was published in 1665. Three or four similar magazines were then published by several European national academies of sciences. Science magazines published in 1700 were less than 10 worldwide. This number increased to 100 in 1800, 1000 in 1850, and 10000 in 1900. At present, scientific magazines in the world have over 100000 types. Accordingly, the number of science magazines has increased by 10 times every 50 years since 1750. Price also observed the same trend when he studied the growth of CA, biological abstracts, and science abstracts in recent decades. On the basis of these studies, Price identified the rule that science magazines increased exponentially. He obtained the same result when he statistically researched the characteristics of the increase in the number of journal articles using PA and 30 other abstracts. Price then concluded that literature in any normal growing field of science exponentially increased to double nearly every 10 or 15 years; the growth was approximately 5–7% a year.

Price’s comprehensive analysis of a large amount of statistical data considers the number of scientific literature as the longitudinal axis and historical s as the horizontal axis, thereby pointing out scientific literature in different s locations on the coordinate point by point, with each point connected by a smooth curve. This analysis approximately represents the rule of scientific literature growth over time and is known as the famous Price curve (Fig. 2.3).

Fig. 2.3 Exponential growth curve of scientific literature



Through the analysis of the curve, Price first observed the exponential function relationship between scientific literature growth and time. If the amount of literature at time t is expressed in $F(t)$, then the law of index can be represented as

$$F(t) = ae^{bt} \quad (a > 0, b > 0), \quad (2.2)$$

where

- t time, in years;
- e natural logarithm of the bottom ($e = 2.718\dots$, can take approximately 3);
- b time constant, sustained growth, i.e., a cumulative increase in literature within a year and the ratio of the cumulative total last year

If r denotes the percentage of literature volume growth in one year, then $r = 100(b)$ or approximately $r = 100b$, $b = r\%$.

$$F(100) = 10,000e^{0.1(100)} = 220,264,660 \text{ (piece)}$$

At an initial moment, e.g., in a certain amount of scientific literature for $a = 10000$, the growth rate is 10%; 10 years later, the literature volume will be $F(10) = 10,000e^{0.1(10)} = 27183$;

100 years later, the amount of literature will be

$$F(100) = 10,000e^{0.1(100)} = 220,264,660.$$

The time d required for the literature to double is typically used as the quantitative indicator of literature growth evaluation. Formula (2.2) indicates that the formula for computing time is $d = n^2/b$. In the example, the time required for literature volume to double is $d = 6.93$ years.

Literature growth varies in different disciplines. The volume of some academic literature doubles every few years, whereas that of others requires over 10 years. For example, the multiplication period of chemical literature is 8 to 9 years. By contrast, the literature of certain areas and emerging cutting-edge disciplines, such

as atomic energy and environmental science, requires 2 to 3 years to double. The SCI growth rule is also relative to the cumulative number of annual literature.

2.3.3 Analysis of the Literature Index Growth Law

Correctness of the law of literature exponential growth

From a mathematical point of view, the price index growth formula, i.e., Formula (2.2), is an analytical function, and a derivative is obtained on the interval (0, up). If the first derivative on (2, 2) is taken, then we obtain the growth curve as

$$dF(t)/dt = abe^{bt} = bF(t). \quad (2.3)$$

The relative growth rate is

$$dF(t)/dt/F(t) = b. \quad (2.4)$$

Notably, $a > 0$, $b > 0$;

hence, $dF(t)/dt > 0$, $(0, \infty)$.

The mathematical point of view is as follows: exponential function $F(t) = abe^{bt}$ is a monotone increasing function on interval $(0, \infty)$. This function represents increases in scientific literature content over time.

Formulas (2.3) and (2.4) show that the growth rate of scientific literature is the index function that is decided only by the achieved level ($F(t)$), and the relative growth rate is a constant. $F(t)$ represents the total number of literature in the past or in the near term, regardless of periodical publication, literature aging, and other factors. This conclusion agrees with the statistical results of the past history of literature.

From the point of view of statistics, the SCI growth rule correctly reflects the actual situation of literature growth. For example, a statistical analysis of the growth situation of world books in 1952–1982 shows that book types approximately double every 20 years. This analysis is a good fitting exponential growth model for the actual situation. Book growth in 1952–1982 conforms to the law of exponential growth. As Price emphasized, “the existence of the exponential curve is apparently universal and long term.” Therefore, the law of scientific literature exponential growth demonstrates a high degree of accuracy and is recognized by the public.

Limitations of the literature index growth law

The limitations of the literature index growth law are mainly manifested in the following two aspects.

Scientific literature does not always present exponential growth. A relationship exists between Price's index growth model and research literature in terms of discipline and time. Numerous studies have shown that extensive research subjects require a long time to follow the index law. The counting of beginning time significantly influences literature growth. The growth obtained from statistics is

larger than the actual growth rate. Thus, not all subjects in any period of literature increase as an exponential growth pattern. In fact, a flat trend of exponential growth curve in literature steepness occurs.

The index law cannot predict the future trend of literature growth. We attempt to investigate curve (2, 3) to determine what causes the laws of scientific literature absolute growth to change over time as follows:

$$\begin{aligned}\Delta F(t) &= ae^{b(t+1)} - ae^{bt} \\ &= a(e^b - 1)e^{bt}.\end{aligned}$$

When $t \rightarrow \infty$, $\Delta F(t) \rightarrow \infty$.

As time passes by, the increment in scientific literature tends toward infinity. This finding is evidently unrealistic. Thus, based on trend extrapolation in the research field of prediction, the Price curve is used to predict the amount of scientific literature at a certain point in the future. When the forecast period is longer than 10 years, reliable results are difficult to obtain. Although many factors influence literature growth, such growth is mainly the result of scientific research and the growth of scientific and technological personnel. If scientific literature has always been in accordance with exponential growth, then even if every person in the world becomes scientific research personnel and scientific research funds account for 100% of the national output, the requirements for scientific literature infinite growth will remain difficult to fulfill. The scientific literature law of exponential growth has exhibited the limitation that it cannot predict future scientific literature trend. Reason for the limitations in the literature exponential growth law.

In summary, the law of science literature index growth correctly reflects the growth situation of scientific literature in the past but cannot predict the future scientific literature trend.

The growth of scientific literature is a complicated social phenomenon and process. The limitations of the literature index growth law are attributed to the influences of many complex factors on the growth of scientific literature and restrictions, as specified below.

Theoretical research on scientific literature growth remains limited. Various factors that influence literature growth are difficult to fully consider and provide an accurate analysis and quantitative description.

In establishing the exponential growth regularity of scientific literature, Price did not consider the increasing aging of literature.

When Price accumulated the total number of scientific journals in one year, he did not rule out unprinted journals. A. И. Михалове highlighted that every three types of science and technology journals appear while journal closure occurs. The influence of this continuous publication closure should not be ignored. Therefore, the price index growth model has certain errors.

2.4 Law of Literature Information Logic Growth

2.4.1 Literature Information Logic Growth Model

On the basis of research on the SCI growth law, scholars from many countries have conducted considerable research and have proposed several theories and mathematical models to search for a perfect literature growth model. This research trend has created the largest effect of “literature information logistic growth model.”

In 1963, Price discussed the scientific literature exponential growth law and the logic growth law for scientific research personnel in his classic book “Little Science, Big Science.” He argued that exponential law would eventually become a logical type.

В.Налимов, a scientist from the former Soviet Union, researched the scientific literature growth law and found that literature growth comprised stages, with each stage having a different growth model. For example, the use of the exponential model to describe the literature in the field of system research in the Soviet Union between 1957 and 1974 can be divided into three periods. The doubling time of each period literature differs.

After extensive research, В.Налимов and Г.ВлӘдуц determined that scientific literature initially passed through a rapid growth process and then gradually slowed down. The growth process of the growth rate index changed into a logical curve growth process. Налимов and ВлӘдуц also considered that material conditions, economic sources, and author intelligence influenced the growth rate of scientific literature. On the basis of a specific literature research, they proposed the famous literature logical curve theory and model as follows:

$$F(t) = k / (1 + ae^{-kt}) \quad (b > 0); \quad (2.5)$$

$F(t)$ literature cumulant during t years;

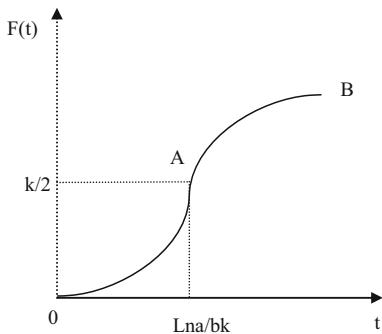
k literature cumulant, $t \rightarrow \infty$ is the maximum literature cumulant;

a, b parameter

If the second derivative is taken with Formula (2.5) on t and the second derivative is zero, then the inflection point A of curve (2.5) can be obtained for the coordinates $(\ln a/kb, k/2)$. This curve is symmetrical around the inflection point and is called the logic symmetry curve.

Figure 2.4 shows that literature growth will be limited ($y < k$). The relative growth rate $\left(\frac{dy}{dt}\right)/y = b(k - y)$ is a linear function of y . When $t < \ln a/kb$, the growth rate is increasing; when $t > \ln a/kb$, the growth rate is decreasing. The growth rate evidently slows down when y is nearly approaching k . The aforementioned process respectively corresponds to the curves of OA and AB.

Fig. 2.4 Scientific literature logic growth curve



When $y < k$, an approximate growth curve can be obtained from (2, 5), i.e., $dy/dt = kby$; the relative growth rate of the curve is $(dy/dt)/y = kb = \text{constant}$. The initial stages of scientific literature growth are in accordance with the law of exponential growth. However, the index value changes with the change in time t , and relative speed is not always the same. Consequently, the exponential growth momentum cannot be maintained. When literature volume increases to half of the maximum, its growth rate becomes small and tends toward a limited value of $y = k$.

The logic curve of scientific literature is relative to the accumulation of the present number, which is a conclusion obtained based on statistics and research in a certain field of knowledge or the same type of literature for a long period.

2.4.2 Analysis of the Law of Literature Information Logic Growth

(1) Correctness of the law of literature information logic growth

Friedrich von Engels indicated that the movement of the universe was confined within a limited circle of infinite development in the book “Dialectics of Nature.” The logic rule of scientific development is similar to this thought. This finding shows that science is experiencing a process of time continuation. This process is a prophase slow development in the early stage, acceleration development in the middle stage, and speed reduction until the late saturated development. Saturated development does not indicate the end of science development, but instead, it denotes that science development has achieved a dynamic and balanced system. The science development process corresponds to the logic growth process of scientific literature. Therefore, the literature logic growth rule exhibits considerable necessity and rationality in terms of philosophy or science.

In fact, the scientific literature growth logic curve has succeeded in describing the literature growth rule. The logic curve perfectly describes the growth law in terms of the growth of the scientific literature for a specific subject area. The growth

of the literature on coal gasification counted by D.J. Frame et al. from 1965 to 1975 perfectly conformed to the logical growth curve model. Several stages of the literature development on mast cells have compellingly proven the correctness of the logic curve growth rule.

From the perspective of scientific history, all the subjects in the field of science are currently in the birth, development, and relatively mature historical stages. Statistics show that when subjects are in the birth and development stages, scientific literature exhibits exponential growth. However, the lifespan of scientific literature is relatively short. The growth rate is inversely proportional to the lifespan of literature. With the deepening of subject research and coming into a relatively mature period, the growth of scientific literature cannot maintain the original index rate, the growth rate decreases, and the curve flattens out. Thus, the lifespan of literature becomes relatively longer. However, the decrease in literature growth rate does not indicate a stagnation in scientific development. This scenario is still considered a relatively mature stage after major research progress has been made in a particular knowledge field. It can also imply that knowledge areas are facing a new breakthrough and will produce updated subfields.

The descriptions of the constants (i.e., a , b , and k) in the literature growth logic curve of different disciplines generally differ. If we can separately perform a comprehensive statistical analysis of the literature growth of each specific discipline and draw a logic growth curve, then these approaches can play a certain practical guiding role on stage evaluation, future development prediction, related literature collection, and intelligence service. Therefore, research on the law of scientific literature logic growth considers the theory of information science, has scientific value, and can serve as a guide in practical applications.

(2) Limitations of the literature information logic growth law

Scientific literature logic growth theory, which is relative to Price's growth theory, is a significant improvement. Logic curves can be utilized to describe the growth of past literature and predict future literature growth. However, several literature statistical studies have not reached a logical conclusion, such as the curve. The logic curve has limitations, i.e., as indicated in Formula (2.5), when $t \rightarrow \infty$, $y \rightarrow k$, $dy/dt = b(y-k) \rightarrow 0$. This scenario shows that when the scientific development is at a certain stage, the growth rate of scientific literature tends to be zero, and the amount of insurmountable scientific literature reaches its maximum value. This scenario further implies that after a certain stage wherein no new scientific literature is developed, scientific literature will fade away from the field of human science communication as a means of transferring information. Therefore, some scholars believe that the development of science will reach its limit someday and then will be eventually saturated. Scientific progress will become stagnant or even suffocate. Price asked questions regarding this issue in his book "Little Science, Big Science," which was also at the extreme side. The quantity growth of literature on science and technology is slowing down, but this situation does not imply that the speed of scientific development will drop. In addition to scientific literature that reflects

human awareness, knowledge, and support level in the form of the further development of scientific research, we can have new and better communication technology intelligence methods. This scenario makes it possible for us to supplement or gradually replace existing traditional forms of scientific literature.

(3) Reason for the limitations of literature logical growth

The difficulties in utilizing exponential and logistic curves (i.e., semi-exponential curve) to predict the future amount of scientific literature has two main reasons. First, an exponential growth model has several limitations as previously mentioned. Second, the predicted scientific literature on these two growth curves is developed based on the predictions of trend extrapolation. However, scientific literature is a complex system of scientific communication subsystems, and its growth is affected by many aspects of laws and constraints. The adoption of system theory to its systematic analysis can obtain more accurate results. This approach should be the starting point of our modified index and logistic curves to explore a new scientific literature growth model.

2.4.3 Modification of the Model for Literature Information Logical Growth

The logic of the scientific literature growth curve overcomes the defects of index curve “divergence”, but the limitations of “bounded of the index curve” remain, which must be further improved. The modified literature in the logistic growth model must fully consider the influence of different factors, which is the correct thinking approach.

First, the basic processes of growth, mechanism, and trend in scientific literature must be clarified. The growth of scientific literature will generally experience two large transformations in the past and future of a long historical period. The first transformation is from exponential growth to a logical type of growth, and the second is from logical to linear growth. The transformation can either be gradual or sudden. Science will have different means and tools for communication with the development of science and technology, but scientific literature can remain as a means of communication for a long period. A particular balance can also be achieved. That is, the absolute growth of literature each year will experience ups and downs, but deriving the statistical average for the long term shows that the annual growth in literature is the same. The amount of scientific literature basically increases with uniform velocity at any time.

Second, scientific literature must be regarded as a communication system in its entirety in studies. A scientific literature system is a subsystem of scientific information communication. An information exchange system of the interaction between

each subsystem is generally nonlinear, and the scientific literature growth model is a solution set for nonlinear and complicated partial differential equations.

Furthermore, we must identify the main decisive factor and disregard other relatively minor factors. The rapid growth of research, science, and technology personnel is evidently a factor that affects scientific literature development in two of the most powerful sectors.

Finally, we must determine the specific goals and requirements of a model. New mathematical models must overcome index curve “divergence” and the “limited” logic curve of difficulty, as well as describe a better approximation of the literature growth rule. When $t \rightarrow \infty$, the growth rate of scientific literature dy/dt is a constant, which is unlike the logic curve that tends toward zero. Literature also does not tend to be a maximum total k , but acts as a function of time. The annual growth in scientific literature also tends to be stable when $t \rightarrow \infty$, which suggests that scientific literature exists simultaneously with other scientific communication media. The scientific communication system forms a relatively stable “ecological balance” with the social, economic, and technological development levels. This scenario is in line with the actual process and development trend in scientific literature, which can provide a relatively satisfactory explanation for the literature growth rule.

From the preceding principle, a new and more appropriate description for the literature growth regularity of a mathematical model can be constructed based on the logic curve, thereby increasing secondary modifications and considering a constant term after reasoning and transformation.

2.5 Other Mathematical Models for Literature Information Growth

2.5.1 Linear Growth Model

(1) Contents

Linear growth of scientific literature:

$$F(t) = bt + a \quad (2.6)$$

where

- F(t) t : amount of literature growth per year,
- b literature growth gate,
- a amount of literature when $t = 0$

(2) Analysis

In 1963, Price indicated that the index law could likely be damaged, and the literature of exponential growth could not continue forever. In fact, some aspects in literature do not regard the growth model for the exponential curve as a logical curve. However, they also present a linear growth model as stated in “Little Science, Big Science”. A.I.Mikhaylov stated that the members of the Eastern Europe Cooperative Organization within the scope of science books and journals, as well as the number of patent specifications, all exhibited a straight line law of growth. Relevant statistics show that several books and pamphlets published worldwide from 1960 to 1972 also exhibit a linear law of growth.

The scientific literature linear growth model does not only apply to describing certain knowledge or types of literature growth. The development of science and literature in the future will be more inclined toward the linear model. Rescher indicated that linear growth would be attained in the future through past exponential growth.

2.5.2 *Hierarchical Sliding Exponential Model*

(1) Contents

Nicholas Rescher, an American scientific historian and information scientist, believed that the world could not bear the input index of literature, and publication growth would become linear during the years when resources were lower. In his book “Scientific Progress,” Rescher indicated that the growth rates of publications with various quality levels differed with the increase in publication number related to their quality. Hence, the following hierarchical sliding exponential model was proposed to describe the scientific literature growth rule.

Rescher introduced the index λ , which represented the quality of literature, and $0 < \lambda < 1$. The meaning of λ is detailed as follows:

- $\lambda = 1$: common literature (i.e., it represents all literature types),
- $\lambda = 3/4$: literature with meanings,
- $\lambda = 1/2$: important literature,
- $\lambda = 1/4$: highly important literature,
- $\lambda = 0$: most important literature.

The number of literature and literature features vary in different quality levels. In particular, if the amount of literature is $F(t)$ at time t , then the number becomes $[F(t)]^\lambda$ at the λ level. The λ equations are presented as follows:

$$F(t)_{\lambda=1} = ae^{bt},$$

$$F(t)_{\lambda=3/4} = (ae^{bt})^{3/4},$$

$$F(t)_{\lambda=1/2} = (ae^{bt})^{1/2}, \quad (2.7)$$

$$F(t)_{\lambda=1/4} = (ae^{bt})^{1/4},$$

$$F(t)_{\lambda=0} = 1 na + bt. \quad (2.8)$$

For example, in the case of a million literature samples, the number of literature in each level is as follows based on the preceding equations:

$$\lambda = 1:1,000,000;$$

$$\lambda = 3/4:31,623;$$

$$\lambda = 1/2:1000;$$

$$\lambda = 1/4:32;$$

$$\lambda = 0:14.$$

Hence, the following conditions can be shown.

- ① The first level ($\lambda = 0$), or the number of most important literature, is extremely few, which occurs only as 0.0014% of the total.
- ② The number of important literature ($\lambda = \frac{1}{2}$) is the square root of the total number.
- ③ When $0 < \lambda < 1$, the number of literature is still increasing at all levels based on the index law only when people attempt to improve their quality. Growth speed gradually slows down with the increase in literature importance degree. When $\lambda = 0$, the rules are completely broken. Only a constant increase during each period is observed. Thus, the number of literature at that time will increase linearly.

At time t , the number of primary literature is $F(t)_{\lambda=0} = \ln F(t) = \ln ae^{bt} = \ln a + bt$. When the annual growth rate is assumed as 10%, important documents increase every 10 years. Such growth speed is relatively slow.

In this model, if the double time of total number of documents is d , then the quality of the λ level for the λ literature doubles the amount of time decided by the press type as follows:

$$d' = 1n2/b\lambda = d/\lambda. \quad (2.9)$$

For example, the literature volume doubling time is 6.93 years in the third quarter, and thus, the time that corresponds to each λ level literature volume doubled (d) is as follows:

$$\begin{aligned}\lambda &= 1:6.93 \text{ years}, \\ \lambda &= 3/4:9.84 \text{ years}, \\ \lambda &= 1/2:13.86 \text{ years}, \\ \lambda &= 1/4:27.72 \text{ years}, \\ \lambda &= 0:10 \text{ years to increase the } \lambda \text{ primary literature.}\end{aligned}$$

This scenario shows that the growth speeds of different literature quality levels differ. Scholars call this phenomenon the “principle of quality check,” wherein literature growth is affected by quality.

(2) Analysis of the slide grading index model

The le hill model shows that the growth rates of different literature quality levels vary. The more important the literature type is, the slower its growth rate. A few high-quality papers are always accompanied by many general papers, which is a logical condition. The le hill model can also be considered an attempt to study the growth rule from the internal process of growth in scientific literature. However, using specific data to verify the correctness of the model is difficult. For example, Kenneth Mei considered the new ideas and results of his thesis as the first level of the first important documents. Mei identified the determinant and conducted a statistical analysis on the aspects of a paper before 1923 according to six types. However, his results cannot prove that the le hill model is correct.

2.5.3 Transcendental Function Model

Former Soviet Union intelligence scientists P.A. Гиляревский and Шремдер stated that scattered journal articles must be considered factors of the increase in scientific journal articles. When obtaining a journal rank in a discipline or knowledge field according to Bradford's distribution, the growth in periodical publications varies at different grades. Thus, these scientists proposed a new model that highlighted growth in number, called the transcendental equation or Ji growth model.

This model also attempts to delve into the internal structure of literature growth to determine the different distributions of the literature growth law. However, this model is only a special hypothesis theory model that studies the increase in number of journal articles.

2.5.4 Шестопал–Бурман *Growth Model*

(1) Content of the Шестопал–Бурман growth model

The growth in scientific literature is a significantly complicated process. Not only the absolute number of new literature differs every year, but the relative growth rate also varies. Price, Rescher, and other researchers had long believed that the increase in literature should be less increasing". In particular, the relative growth rate increases with the increase in time t , which then decreases with increasing total N or literature. Soviet intelligence experts B.M. Шестопал and П.Н. Бурман proposed a new type of literature growth model in 1978 from the perspective of research in the quantitative evaluation of document information flow growth.

For a convenient discussion, the literature growth equation is first rewritten in the following form:

$$dN/dt = qN \quad (q > 0), \quad (2.10)$$

where

N total number of literature accumulated;

q relative growth rate, which can be represented by t as $q(t)$. Thus,

$$dN/dt = q(t)N.$$

Furthermore,

$$N = N_0 \cdot e^{\int q(t)dt}$$

or

$$\ln N = \ln N_0 + \int q(t)dt. \quad (2.11)$$

This equation represents the research literature information flow growth model.

(2) Analysis of the Шестопал–Бурман growth model

The Шестопал–Бурман growth model from the literature growth rate q is a variable set used to study the growth law of literature. This model is a significant correction of past research. Formula (2.11) includes the literature index growth model, as well as the linear growth model and the index of the sliding model. Therefore, the proposed model is a significant progress of the science literature growth rule. Models cannot apply but give up, including the logical growth curve model and the general model. Formula (2.11) is clearly not comprehensive.

Шестопал and other researchers used the direct graphic model, whose methods could be applied to determine the expression of q . This approach results in a larger error, and thus, applying the least squares method is better.

2.6 Analysis of the Mechanism of Literature Information Growth

We have discussed the six models for describing science literature growth laws in the previous sections. Some of the increases in scientific literature follow the exponential rule, which is a complicated process. Some exhibit regularity with a logistic curve, whereas others are linear. Different models indicate a common point, i.e., science literature increases with time, and the only difference is the velocity of increase. This section discusses the mechanism of science literature growth, which explains why the number of science literature increases and why the increase follows different models.

2.6.1 Reason for Literature Information Growth

Scientific literature growth occurs for many reasons. The basic reason is the rapid development of science and technology, which has also significantly boosted its disciplines, research groups, and funding. Therefore, scientific research is continuously generating new discoveries that can result in the continual expansion of science literature. The progress of science and technology can lead to an increase in the number of publications, which is a common occurrence. For example, the emergence of new technologies in laser and variety improvements has resulted in a surge in literature, from approximately 20 papers in 1960 to 1200 papers in 1965. In particular, the main reasons for the increase in science literature are discussed as follows.

(1) Surge in research finance and number of researchers

Scientific development is a basic indicator of the numbers of scientific research, scientific and technological personnel, and science and technology literature. The increase in the third aspect is largely the result of the increase in the first two aspects. Price's theory states that the exponential input of the entire science system (i.e., finance, human resources, and material resources) has led to the exponential output of science literature. The main reason for the rapid increase in science literature is that research financing and number of researchers are also rapidly increasing. The increase in scientific research investment and number of researchers will undoubtedly lead to an increase in scientific research achievements, as well as record and reflect the increasing number of scientific results. World research financing doubled every year from the 1900s to the 1970s. The number of

researchers increases tenfold every 50 years. For example, 1000 researchers were documented in 1800. The number increased to 10000 in 1850, 100000 in 1900, 1 million in 1950, and 10 million in 2000. These increases are attributed to the increased funding and the increased number of scientific research personnel, which have doubled scientific literature every 15 years.

Research financing and number of researchers have been increasing exponentially; however, the orders of these exponents exhibit differences. In 1966, Price determined the relationship between these three factors, i.e., the growth of the cost of scientific research is squarely proportional to the number of researchers. Nevertheless, the number of scientific products only increases proportionally with the square root of the number of researchers. When the number of scientists is n , their financing is n^2 , but the number of scientific literature is \sqrt{n} . For example, when the number of scientists increases thrice, then research financing will increase 9 times, whereas the number of literature will increase 1.7 times. This conclusion can approximately demonstrate the relationship among science literature, research financing, and number of researchers (Tables 2.2, 2.3 and 2.4).

(2) Expansion and detailing of the professional scope

Over 2300 disciplines currently exist worldwide, and 10 billion publications have been documented as of 2005.

(3) Mutual infiltration among disciplines

No closed discipline exists in modern science, such that innovations and inventions can affect other disciplines, research methods, or can directly be applied to other disciplines.

(4) Internationalization of science and technology

Important inventions or discoveries can more easily become worldwide trends and will be immediately available for further study. Thus, people must collect all of the literature and information in the world for transfer and utilization.

(5) Cooperative and collective research

At present, hundreds and thousands of people are being mobilized, and billions of research projects are still increasing. The multinational cooperation development project that focuses on the continual research of cooperative and collectivization degrees is strengthened.

Table 2.2 Increase in scientific research personnel worldwide

Time	Number of scientists
1800	1000
1850	10 000
1900	100 000
1950	1 000 000
1970	3 200 000
2010	11 260 000

Table 2.3 Funding growth in the United States

Year	1920	1930	1940	1950	1955
Project					
Cost (millions of US dollars)	80	160	377	2870	6270
GDP ratio (%)	0.1	0.2	0.4	1.0	1.6
Year	1960	1965	1970	1975(estimation)	
Project					
Cost (millions of US dollars)	13 730	20 430	26 566	35 600	
GDP ratio (%)	2.7	3.0	2.7		

Table 2.4 Funding growth in the Soviet Union

Year	1950	1955	1960	1965	1970	1971	1972
Project							
Cost (1 billion rubles)	1.0	-	3.9	6.9	11.7	13.0	14.4
GDP ratio (%)	1.8	-	2.7	3.56	4.03	4.27	4.60
Year	1973	1974	1975	1976	1977	1978	
Project							
Cost (1 billion rubles)	15.7	16.5	17.5	17.7	18.2	19.1	
GDP ratio (%)	4.65	4.67	-	4.0	-	-	

(6) Shortened research cycle and boost in production and translation

Statistics showed that the commercialization cycle of inventions before World War 2 was 20 to 30 years, which had been shortened since then to only 2 to 3 years, and eventually to 1 over 10 years.

(7) Improvement in communication, publishing technologies, and information science

With the development of science and technology, as well as more advanced communication, literature databases, online services, and ongoing research project information science are expanding. All of the aforementioned factors have accelerated the exchange in scientific literature and have promoted the rapid increase in scientific literature volume.

Scientific literature is comprised of the requirements of society and the growth of a country's policies on the development of scientific and cultural undertakings.

2.6.2 *Explanation for the Literature Information Growth Law*

We considered that the laws on the increase in the number of scientific literature are determined by the objective process of scientific development. The actual process of scientific development is restricted by two factors: the intrinsic of science law and the environments of science.

(1) Scientific development laws influence on the increase in literature information laws

The science discipline teaches that the changes in number of scientific literature denote an important symbol of the development of science itself. Thus, we developed our model to explain the growth of scientific literature analysis according to scientific development.

The famous science historian and scholar Thomas S. Kuhn proposed the science development pattern in his book “The Structure of Scientific Revolutions.” Kuhn believed that scientific revolution would always occurs, and that its development process would occur from “original science” to “normal science” and the transition from “normal science” to another “normal science” of the process. Entire processes of scientific development are repeated in cycles under the impetus of scientific revolution.

During the original science period, many scientists from different academic schools were writing articles to participate in scientific discussions, thereby increasing the number of schools of thought. Although the amount of scientific literature was not large at that time, it was still increasing rapidly. Its growth rate was akin to a constant, and scientific literature followed an index law of growth during this period. New ideas and achievements of the most important papers of the period were recorded.

When normal science occurred, general science was in a mature and stable development period. New ideas and achievements were widely adopted in education, popularized, and promoted during this period. Hence, the number of science literature and its incensement were large, but the increase rate had declined. The total number of literature would become an extremum. The growth in literature transformed from the exponent model into the logical curve model. The increase in logic curve tended toward the extremum, which showed that the discipline or knowledge area was developing into a new crisis after long-term stable development. When a crisis occurred, the scope of the subjects would be changed or divided to produce new disciplines. Science literature would transition once more to an exponentially increasing stage along with science revolution. Therefore, if the statistics of the literature for a fixed number of years corresponded to a knowledge area from original science to conventional science or from conventional science to update the entire period of normal science, then the logical growth curve model would be followed.

Kuhn's theory of scientific revolution states that the logistic growth model should be the ideal model of the scientific literature growth law. Logistic curves with exponential curves are similar to the former period and near the inflection point at an appropriate time interval that can be approximately regarded as a straight line. If the statistics of the fixed number of years roughly correspond to any of the aforementioned scenarios, then the three periods can be obtained through a linear growth model. Numerous science disciplines at present have transitioned into normal science from original science. The literature in many knowledge domains tends to increase linearly in this scenario.

Menard's research shows that a discipline in the general literature growth rate changes according to three different periods as follows:

- ① Stable stage—increasing in line, science was at its birth;
- ② Increasing stage—increasing rapidly at an exponential rate, science was developing;
- ③ Cycle stage—stability and growth occur alternately, science had matured.

Rescher believed that the scientific development process was not a process of addition but a process of subtraction. The main findings at present are the opposite of past conclusions because information absorption is accorded with the “effect of decreasing return of law” and literature growth from the perspective of economic analysis. The development of disciplines and shifting of scientists interested in paper distribution do not only exhibit an exponential growth state, but are also conditioned by many factors and have many distribution states. Therefore, the number of literature is increasing following different growth models.

(2) Social environment influences the literature information growth rule

Scientific environment conditions, including politics, economy, culture, education, and other social conditions, clearly affect the number of science literature laws. These conditions determine the quantity and quality of scientific research investment and scientific research personnel. Thus, they also determine the number of scientific research achievements and approximately reflect the results of the number of scientific literature. The community requirements and provisions for financial, human, and material resources frequently determine the emergence and development of a science field. Scientific literature is not only required by the law of scientific development, which is a theoretical growth model in itself, but also by different social and environmental conditions of restriction. For example, the First and Second World Wars, and the damages inflicted by these wars on the development of science and technology, resulted in two “rough” periods. Scientific literature also underwent two major declines, which caused the pattern of certain disciplines to change from an exponential growth curve to a logical one. Given the current shortage in resources, the society has been unable to develop a scientific system that provides an index input. The exponential growth in the 1970s resulted in mostly linear publications, coupled with an amount that corresponded to the general scientific literature growing characteristics. This scenario caused most of

the scientific literature to transition from exponential to linear growth from the past to the future. Socio-environmental conditions frequently cause an increase in scientific literature, which exhibits random process characteristics and changing growth rate. Carrier technology, increase in scientific literature publishing technology, effects of computers and modern communication technology, and many other factors can also be considered random processes.

2.7 Applications of Literature Information Growth Laws

Research on scientific literature growth laws plays an important role in science history, science, and theoretical and practical information science. Their applications mainly include the following aspects.

2.7.1 *Applications to Science of Science*

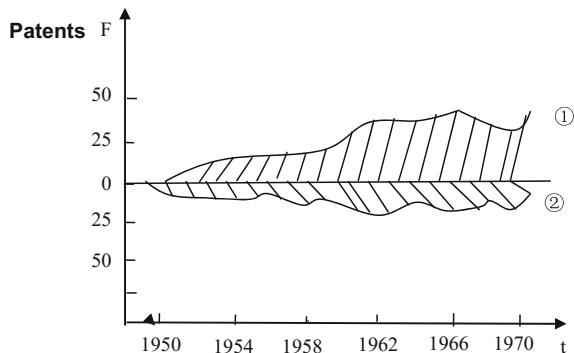
The increase in the number of documents changes the rules for judging and predicting the growth of scientific knowledge. It then explores the law of development of science as a whole, which is a commonly applied method in the history of science and scientific research. In this case, G. Gilbert proposed that establishing the following hypothesis would be necessary before working on studies: all knowledge should be contained in published literature, and each document should contain the same amount of knowledge. A. Doyle provided a more vivid metaphor: “Like saving money to provide the interest, knowledge generates new knowledge that in the increases in knowledge is just as banks offering compound interests—Growth in any time will produce a fixed percentage of the number of recent.”

In a large-scale study of science and technology, the increasing number of scientific literature laws generally simulates the development process of science and technology, as well as explores its laws. The price of conclusions on the exponential growth law of science and technology indicates scientific literature quantities according to the index law, which is from the growth based on a conclusion. Therefore, the laws of scientific literature research are efficient paths in the study of science history and theory applications.

2.7.2 *Applications to Information Research*

From the perspective of information science, studies on the increasing law of science literature is a reliable method for scholars in information science research,

Fig. 2.5 Patent distribution of semiconductor diffusion technology in the world ① and ② curves indicate the number of changes in relevant patents in Japan and other countries in the world, respectively (1949–1970)



allowing them to understand the trend of science development, and to make scientific predictions.

The increase in the number of scientific literature can denote a country with a technological development process and achievement levels. In all literature types, patent documents are the most sensitive indicators of science and technology development. Hence, scholars frequently utilize the change in the number of patent documents in intelligence analysis. Figure 2.5 shows the changing curve of the relationship between the number of patents for Japanese semiconductor diffusion techniques and the number of patents from other countries from 1949 to 1970. From the perspective of science, studies on the growth laws of science literature are reliable methods for scientific researchers to analyze information science, as well as to understand the trend of science development and scientific predictions. Japan started this research trend two years later than other countries. It then caught up with other countries in 1964, and eventually became the leader in 1968. The need for technology import provides a reliable decision basis for the current study. Therefore, changing the number of literature can fully comprehend a branch or technical area of the entire formation process, development, and future trend; select a scientific research topic; and identify a technical solution to provide a quantitative basis.

2.7.3 Application to Literature Information Management

Research on growth laws began with studies on library management. Libraries or intelligence agencies determine reasonable allocation of funds, principle of data collection, collection increase strategy, storage space expansion measures, information processing, and transmission of exchanges in new technology applications. These processes are contained in the number of scientific literature, which is an important decision basis of future growth trends. For example, a literature center

can be increased by studying its collection of laws and trends, as well as by determining the budget and size of collection development. The detailed decision must add the number of shelves and area of library expansion.

However, several factors and problems, such as research tools, mathematical models, research methods, and reliability of results, still affect these studies.

Chapter 3

Literature Information Obsolescence Law

It is objective law of scientific development that scientific knowledge updated, new knowledge produced, the old knowledge replaced along with the progress of science and technology constantly. As the record of scientific knowledge, metabolism of scientific literature should be bound to ensue. Literature obsolescence is an inevitable and general social phenomenon. Obsolescence and its laws is an important topic of Scientometrics and bibliometrics, also an important topic of information metrology too. People launched a number of studies around literature obsolescence, such as literature obsolete mechanism, measure, the mathematical model and its influencing factors since the Gosnell's research (C.F. Gosnell 1943). Those studies mainly around three aspects: the first is the theory of literature obsolescence and then explore the dynamic law of literature information dissemination; the second is study on research methods and quantitative description method, in order to accurately grasp the principle and its dynamic mechanism of obsolescence law; the third is the application research of Literature obsolescence, in order to guide books of choose and buy, collection shelf circulation and other activities better, to raise the utilization ratio of literature and faster service benefit.

3.1 The Concept and Measure of Literature Information Obsolescence

People have long-term exploration on obsolescence law of scientific literature and its characteristics. Gosnell (C. F. Gosnell) was the first to research literature obsolete which from the university of New York. In 1943, he used this topic in his doctoral thesis, and published an article titled “literature obsolete problems in university library” on “the library of universities and research institutions” the following year in March. After that the research literature on the subject is not much. People began to study quantitative indicators and methods to describe

literature obsolete since the late 1950s. Introduce the concept of “half-life” to scientific literature field is a milestone in literature obsolete study. In the ‘60s, the American R.E. Barton and R.W. Kepler did a series of research on “half-life” of scientific literature. They pointed out the meaning of “half-life” in the field of scientific literature, described the obsolete curves, calculated the standard formula, and carried out the preliminary calculation of the “half-life” of literature in nine subjects, the study on literature obsolete promoted one step further. Since the 70s, funding cuts and capacity crisis, made the people in the library pay more attention to the research in this field, to find the law of literature use in past and present and show the development trend in the future. Reinhard (M. B. Line) and Sandison (A. Sandison) put forward theoretical issues such as the relationship between the rate of literature use and the value of information, and research literature obsolete under the premise of considering the literature growth. Allen Kent’s papers on the literature of University of Pittsburgh established a example for the library in improve the ordering principle, provide practical standards literature collection and removal. In the following study, the people do a lot of work around literature obsolete measure index, mathematical model and its influencing factors, and other issues, and has made a certain progress. During the process of research, people formed the different understanding of the literature obsolete and metrics gradually.

3.1.1 The Concept of Literature Obsolete and Intelligence Obsolete

(1) A basic understanding about literature obsolete

Since the Goss Neil’s research, the history of the research on literature obsolete phenomenon has been more than 60 years. As the research on deepening, different understanding concepts of literature obsolete formed, can be summarized as: process viewpoint, state viewpoint, viewpoint of process state dialectical.

1) Process viewpoint

This view is the mainstream of research, think that literature obsolete is a process. Gossnell (C.F. Gossnell) points out that: should notice the other side of the problem in the process of the accumulation of knowledge (characterized by the accumulation of literature), which is all knowledge or its carrier will lose the original value with the passage of time gradually in his doctoral thesis “literature obsolete problems in university library “in 1943, and put forward use” document obsolescence” refined the literature become no longer useful or no longer valid gradually. The book “Basis of Literature Metrology “pointed out that explicitly,” obsolete is a process, is a dynamic concept. The persistent of process even lead to negative the measurement

methods and results in literature obsolete of Burton–Kebler. In fact, process viewpoint is the foundation of the rationality in observation through time.

2) State viewpoint

This view think literature obsolete is a state. But it still seems being in the subconscious, had significance when it involving a particular literature obsolete or not. But in fact, Burton–Kebler was measure the literature obsolete by describing literature obsolete state, but the later researchers were considered it's a description of the literature obsolete process still. Actually, state viewpoint is also very meaningful understanding perspective, is the foundation of the so-called synchronic observation of rationality.

3) Viewpoint of process state dialectical

This view think literature obsolete is a process and a kind of state also. Process is the process of continuous variation, state is the state at some point in the process. The demonstrate Burton–Kebler's measure of literature obsolete. His book “literature metrology” pointed that “the scientific literature obsolete is an objective social phenomenon, is a complex dynamic process” (Junping Qiu 1987) have show this view clearly. The view is helpful to deepen the basic understanding of literature obsolete. It made the diachronic observation and synchronic observation have the rational basis.

(2) Obsolete of literature and information

Generally speaking, the “obsolete” issue including obsolete of literature and information. So-called scientific literature obsolescence refers to the content of scientific literature become obsolete with its “time of life” increasingly, as the value of intelligence source decrease constantly, even completely lose. Scientific literature obsolete is not only an objective social phenomenon, but also a complex dynamic process. To this, the famous former Soviet intelligence expert B.A. Полушкин defined it as, “as a source of scientific information, the scientific literature lost the value as its “time of life” increased, and therefore the scientists and experts less who used it less and less. Obsolete is not scientific literature itself, but the article contains the information”.¹ On the concept of information obsolescence, experts Line M.B. and Sandison A. defined as the phenomenon like “the effective value of intelligence attenuation over time”.² While B.A. Полушкин think that should adopt a “document information obsolescence”, and points out that information obsolete is actually the obsolete of the document information, document information obsolescence is a relative concept, is relative to the object or intelligence users. Intelligence object refers to the phenomenon of objective existence objects, facts,

¹Mihayiluofu, Xu Xinmin, et al. Science communication and Information Science. Beijing:Science and technology literature press, 1980.

²Line M. B.,Sandison A. A obsolescence and Changes in the use of literature with time. J. of Doc, 1974.30(3), 283 ~ 350.

and its nature and characteristics, etc. Obviously, these information objects are usually in the constantly changing, and document information reflects some kind of record state of an object, has the invariance, so the document information of obsolete will inevitably appear accordingly. For information users, document information if the lack of pertinence, or loss of the novelty, practicality and gets no use value, you can think of bibliographic information relative to the information users obsolete. Visible, due to compare different frame of reference of these two kinds of obsolete has a larger difference. Literature obsolete and intelligence are two different concepts. Can think that the information obsolescence is relative to the object of intelligence, and the literature obsolescence is relative to the information users. For librarians and information personnel, concerned mainly is the literature obsolescence issues.

The knowledge of the obsolete problem is that there are many confused: always trying to confirm whether a (class) literature or intelligence obsolete, because it is useful for books information management and user. And attempt to grasp its obsolete process with time, hope that through time metrics on whether or not the obsolete, it is useful for the researchers. But for literature or intelligence unit, and whether it is difficult to explain the obsolete, the obsolete affected by many factors. In fact, the obsolete of literature or availability change tendency and rule of intelligence group, is a kind of value judgment.

(3) Precise definition of Literature obsolescence

Definition 3.1 Assuming Ω is the literature space, χ is the σ domain on Ω , P is the availability measure of literature with χ , said (Ω, χ, P) is the literature space measured by P . Ω is the collection made up of all literature, the element in χ is document set.

Definition 3.2 Assuming $X \in \chi$, and $P(X)$ is the availability measure with χ .

Definition 3.3 For $X \in \chi$, $P_t(X)$ is the availability measure of χ aged t .

Definition 3.4 For $X \in \chi$, $X(t) \subset X$ is document set of X aged t , and $P(X(t))$ is the availability measure of $X(t)$ in observation time.

Definition 3.5 Observing the change of $P_t(X)$ along with t , called diachronic observation of X .

Definition 3.6 Observing the change of $P(X(t))$ along with t , called synchronic observation of X .

Definition 3.7 For $X \in \chi$, when $\forall \varepsilon > 0$, if $\exists T$, and $t > T$, $P_t(X) < P_T(X) \leq \varepsilon$ or $P(X(t)) < P(X(T)) \leq \varepsilon$, so there are again in X

If $T = \max\{t | P_t(X) \leq \varepsilon\}$, X is obsolescence of ε - at aged T ;

If $T = \max\{t | P(X(t)) \leq \varepsilon\}$, X is obsolescence of literature ε - at aged T .

Axiom 3.1 For $\forall X \in \chi$, $\forall \varepsilon > 0$, all $\exists T$, when $t < T$, $P_t(X) < P_T(X) \leq \varepsilon$ or $P(X(t)) < P(X(T)) \leq \varepsilon$. The rule is called literature obsolescence axiom. The former is diachronic obsolete, the latter is synchronic obsolete.

3.1.2 Measure for Literature Obsolescence

In order to measure the speed and degree of scientific literature obsolescence, reveal obsolescence law quantitatively, people put forward some main measure such as half-life, price index, remaining useful indicators from different angles.

(1) The half-life

① The concept of half-life

In 1958, an intelligence science conference held in Washington, scientist J.D. Bernal first use “half-life” as the characterization of the obsolete speed with literature information, refers to the time of half of the published literature information no longer used. This concept is suitable for the diachronic observation of obsolescence.

The concept of “half-life” (synchronic half-life) with literature information given by R.E. Burton and R.W. Kebler in 1960, refers to the time of publishing the newer half of all the using literature in a subject (professional). This time the same as time experienced by half of the subject literature lost useful (diachronic half-life). For example, the half-life of the chemical literature is 8.1 years has determined, means 50% of the total chemical literature which still in use when statistical and study happened was published in the last 8.1 years. Also can say, after 8.1 years, half of chemical literature has gradually lost its use value. Allegedly, under the contemporary condition, the value of science and technology literature will be lost more than 30% if the publication delay 1.5 to 2 years.

The concept of “half-life” with literature information given by R.E. Burton and R.W. Kebler also known as “the median citation age”. The journal citation report (JCR) which is the byproduct of the science citation index (SCI), put forward Citing half-life and Cited half-life through integrated the two definition.

According to precise definition method for literature obsolescence, half-life can also be precisely defined:

Definition 3.8 For $\forall X \in \chi$, if $P_T(X) = \frac{1}{2} P_0(X)$, then T called diachronic half-life of X .

Definition 3.9 An observation time selected, for $\forall X \in \chi$, $X = \{X(t)|t = 0, 1, 2, \dots, K\}$.

If $\sum_{t=0}^{T_1} P(X(t)) = \frac{1}{2} \sum_{t=0}^K P(X(t))$, then T_1 called the average synchronic half-life of X . If $P(X(T_2)) = \frac{1}{2} P(X(0))$, then T_2 called the distributed synchronic half-life of X .

② The applicability of the half-life

As early as in 1963, the famous American literature metrologists D. Price is expanded the applicable scope of “half-life” and pointed out that a paper “half-life” is about 1.5 years in his research. It means that half of all the other papers (Citing literature) is published in 1.5 years after the paper cited (Cited literature) published. Visible, the meaning of “half-life” are different between a certain subject document information and a paper (or one year). The former is relative to the number of cited literature, the latter is relative to the cited literature.

But many studies have demonstrated that journal articles “half-life” is much shorter than 1.5 years. In 1970, M.B. Line, the British scholar at the university of bath technology pointed out that study “half-life” should consider the growth rate of literature information, so he added exponential growth to calculate in his study. In 1980, Brown studied the chemical journal literature “half-life” under the premise of considering the literature growth. Broadly speaking, the concept of “half-life” can be applied, to the different main body, can be divided into such as the half-life of subject literature, the half-life of journal literature, etc.; from the perspective of citing, can be separated different concepts as “citation half-life” and “cited half-life” also. Different subjects suitable for different concepts. Synchronic half-life is usually evaluate obsolete trend of a subject or professional literature rather than a single documents. Diachronic half-life can be obsolete half-life of a subject literature, a kind of journals, and even a piece of literature.

③ The calculation of half-life

The calculation of half-life can be obtained according to the definition of half-life.

- 1) mapping method. Made citation frequency distribution table from statistics, drawing a map with citation cumulants or citation cumulant as the ordinate, published age of citation as abscissa, find out abscissa T corresponding the ordinate where half of citation cumulants or citation cumulant in the picture, then T is the request results.
- 2) the quantitative model calculation method. Established a literature obsolete model used statistics, find out the half-life formula according to the definition, and then obtained the results added corresponding data. This will be discussed further when discuss the mathematical model and obsolete index of literature obsolescence.

(2) Price index

① The concept of price index

In 1971, D. Price found that half of all the citation of the investigated literature published in a year published in 5 years after the statistical analysis of the science citation index (SCI). Inspired by this, Price proposed a quantitative index for a literature obsolescence measure in knowledge domains, namely the appellation of “price index”. He thought five years can be used as the standard to distinguish the utilization degree of literature information, the literature with less than 5 years’

publishing life called “real useful literature, the life more than five years called “archives” literature. The price index is the ratio between citation number and the total citation of literature with publishing life no more than 5 years in a certain knowledge areas, the index as a measure of the obsolete speed and degree about literature obsolescence. Its computation formula for:

$$P_r(\text{price index}) = \frac{\text{number of cited literature p less than 5 years publishing duration}}{\text{the total citation}} \times 100\%$$

② The comparison of price index and half-life

In general, for literature in a subject or field, the “price index” is larger and the half-life is shorter, means the literature obsolete speed is faster. As two quantitative indicators of measuring scientific literature obsolescence. They are both from the perspective of literature being used, but reflect the literature obsolete situation in a different way. Price thought, the ratio between the number of “there are now” citation and the number of “archives” citation is an more important characteristic than the “half life” of citations.

Literature “half-life” can only measure the obsolete condition of literature in a subject or all in general, and “price index” can be used in literature of a particular field, can also be used to evaluate the obsolete characteristics of a journals, an institution, or even a particular author or an article. The concept of half-life can be applied to general information obsolescence, the price index is only applicable to the literature information.

To promote the concept of price index.

Although the concept of price index only from citation statistics analysis, but we also can promote the concept of the price according to the precise definition of literature obsolescence:

Definition 3.10 When $\forall X \in \chi$, $P_r(X) = \frac{\sum_{t=0}^4 P_t(X)}{\sum_{t=0}^K P_t(X)}$ as diachronic price index for X.

Definition 3.11 An observation time selected, When $\forall X \in \chi$, $X = \{X(t) | t = 0, 1, 2, \dots, K\}$, $P_r(X) = \frac{\sum_{t=0}^4 P(X(t))}{\sum_{t=0}^K P(X(t))}$ as synchronic price index for X.

Obviously, the original definition of price index is in line with the original Definition 3.11.

(3) The surplus benefit index

① Concept of the surplus benefit index

British B.C. Brookes introduced the concept of journal benefit as evaluation index of obsolescence with surplus benefit of journal. The number of used literature from A certain year of a journal by user is called the journal benefit. The surplus benefit

means the journal's benefit also reserved after several years, which is a measure of obsolete degree about the journal.

② The applicability of the surplus benefit index

When measure obsolete degree by journal benefit, is useful only for information demand of a certain type and content about several kinds of journals.

③ The surplus benefit index calculation

With surplus benefit index as evaluation index of obsolete, means that the surplus benefit index for literature availability measure, and observe the literature obsolescence law. A corresponding quantitative model of literature obsolescence could be established. Brookes assumed journal benefit satisfied negative exponential model, and assume that annual citation amount as the benefit measurement for selected journals in the year. Then the total benefit of selected journals is:

$$U = \sum_{t=0}^{\infty} Ca^t = \frac{C}{1-a}$$

In this formula, C for the citation quantity of selected journals in the first year within its "lifetime", α is obsolete coefficient. So, the surplus benefit i years later is:

$$U(i) = \sum_{t=i}^{\infty} Ca^t = \frac{Ca^i}{1-a} = Ua^i$$

3.2 Research Methods of Literature Obsolescence

3.2.1 Statistical Data Analysis of Literature Management

Statistical data analysis of literature management applicable to research in a given unit of literature and information obsolete conditions and "local" obsolete. It is very useful for guiding eliminate and report the literature. For the corresponding statistical index, the specific statistical methods mainly include books borrowing number statistical analysis and literature circulation statistical analysis. The data can be also used such as about literature circulation, literature belong to the plane and document copy. Make a comprehensive judgment on literature obsolete based on frequency of these data with time distribution. For the library and Information Department, the important is the state of collection obsolete and grasp the dynamic collection of obsolete process in a long-term monitoring process gradually. To some extent, literature management data can reflect the actual situation of used literature in the literature information center objectively. Combinating these statistical analysis with regular business work such as literature choose and buy and collection

elimination is helpful to organization for the transition from the traditional experience management to scientific quantitative management.

Burrell considered the influence of negative exponent law on obsolescence when research the correction of mixture model in book circulation. He still research on a variety of eliminating methods on the basis of the frequency data of circulation. Goughlin verified Burrell's model, Tague's research was disappointed, Gelman showed the β Two model fitting method, Burrell admitted the success of the β two model later, and pointed out its shortcomings. Someone in China inspected the unreliability of using literature using frequency distribution fitting the literature obsolescence. But the existing application is using some more optional standard, remains to be perfect. Now should be seen, statistical data analysis of literature management is a useful method for the quantitative study of networked information obsolescence. Information resources networking system also have this management statistics function.

3.2.2 *Citation Analysis*

Citation analysis is to study the literature obsolescence through collect all the documents about this field in a certain period of time, and the statistical data about the citation of an article published and cited time, etc. Citation analysis is the most widely used and the most effective method of research currently. Applied to research on the literature obsolescence of a subject or a professional. Applied to the so-called "universal" obsolescence research. This method is also divided into diachronic method and synchronic method.

Diachronic method is in consistent with the process viewpoint, synchronic method is consistent with the state viewpoint. View that think two methods to describe literature obsolescence from different angles is in consistent with viewpoint of process state dialectical.

A lot of research literature obsolescence from the perspective of two, but the observation from the two perspective whether the agreement always can't reach a consensus, whether has the rationality is no unified understanding also.

Line and Sandison (1974) think: there is no reason to assume that the same obsolescence measure by diachronic method and synchronic method. Someone in China pointed out the unreliability of synchronic method, completely denied the method.

Stinson and Lancaster (1987) confirmed the rationality of synchronic method. The synchronic and diachronic obsolescence measure can be tested through data comparison, and that "there are differences between the two methods, from different sides reflects the literature obsolete phenomenon, cannot be ignored" was proved.

The two methods have to be fully comparative study in all directions by using quantitative method and analysis techniques yet.

(1) Diachronic method

Diachronic method of citation analysis is a method to statistical analysis the data literature to be cited in the following years after a specific literature (collection) produced.

① Statistical method of diachronic data

A method to reflect the dynamic process of the literature (collection) obsolescence by data statistical about literature to be cited in the following years after a specific literature (collection) produced objectively. Obviously, the time span will be larger and large amount of statistical data in this statistic.

② Analysis method of diachronic data

This method made a specific literature (collection) for statistical observation object, made statistics of cited data after the publication at a certain time interval, and ordered in time sequence, research the change of citation data when observation time t increased. Statistical analysis index mainly including the cited frequency and cited rate, etc. There are mainly analysis method such as regression analysis, correlation analysis and model fitting method, etc.

(2) Synchronic method

Synchronic method of citation analysis is a method to statistical analysis data of the literature published over the past year cited by literature published within a certain time.

① Statistical method of synchronic data

To study the subject areas, collected all the documents in a certain time, statistics the data literature published published be cited in this time, so as to reflect the time distribution state of literature obsolescence during this time objectively. Obviously, the time span will be smaller for this kind of statistical. JCR (Journal Citation Reports) made the synchronic observation of the literature of science every year is the most effective and simple method to help us to get the synchronic data.

② Analysis method of synchronic data

In a given period of time, using synchronic observation to observe the distribution of the citation in different fixed number of year. Research the change of citation time distribution data when observing time t increased. The main statistical analysis index include citation frequency and citation rate, etc. Analysis method such as regression analysis, correlation analysis and model fitting, etc.

3.2.3 Mathematical Methods

It is a kind of effective method that using mathematical methods to study and measure literature obsolescence, and also the main method of literature metrology. Specific application of mathematical methods mainly focused on mathematical statistics and mathematical model.

(1) Mathematical statistics

The mathematical statistical method is widely used, also need further development and utilization. In 1985, Burrell corrected the hybrid model of circulation and considered the influence of the negative exponential law of obsolescence. In 1987, he researched the result of a variety of eliminating methods on the basis of the current frequency data also. In 1994, the British W. Glanzel and U. Schoepflin described statistical properties of the literature cited process with a non similarity production process, given the stochastic model of the literature information obsolescence, discussed the influencing factors and application problems about the model.

(2) Mathematical model

The mathematical model is a method of studying the law by using mathematical theory and method, to describe the relationship between the various factors of the literature obsolescence by mathematical expressions and symbols. The mathematical model method has been widely used in the research of literature obsolescence, and has achieved fruitful results and become a classic study. For example, negative exponential model, Brookes accumulation index model, Burton-Kebler obsolescence model and formal and so on.

3.2.4 Comprehensive Analysis Method

Comprehensive statistical analysis method is a method of survey the actual situation of literature obsolescence. This method based on citation data, circulation data, the library used data, data between literature return to rack data and the total literature used data. This method strive to overcome the limitations of citation analysis and circulation statistics by absorb the merits of the above methods, analyze the situation of the use and obsolescence of literature from the angle of practical use of literature. This method is more objective and comprehensive, but it requires the large-scale, many aspects of literature. So it used to study literature obsolescence generally in some professional or subject field.

It should be pointed out that there are some limitations in these methods. Because the demand of information users and the literature using is a random process, the obsolete of the literature is influenced by many subjective factors inevitably. The needs of society, the language barriers, the knowledge of user, the

quality of tool, the service attitude of the librarian and intelligence, the habits and psychology of user, and even the color of cover of book, and so on all sorts of accidental factors will affect the use of literature. There are some problems need to study by the views and methods of system theory, to obtain a certain effect. It is very difficult to find a unified fixed mode and method for the study of literature obsolescence because the literature obsolescence is a very complicated social phenomenon. This also illustrates the difficulty and the long-term nature of the measurement of the literature obsolescence.

3.3 The Mathematical Model and Index of Literature Information Obsolescence

3.3.1 Classical Mathematical Model and Obsolescence Index

Here has two meanings: one is refers to the model which built up by the classical author, the other refers to the model which established with classical mathematics method.

(1) Several classical mathematical model

Research has shown that obsolescence law of literature can use some mathematical models to describe, such as negative exponential model, Burton-Kebler obsolescence model, Brookes obsolescence model, A. Avramescu obsolescence model, etc.

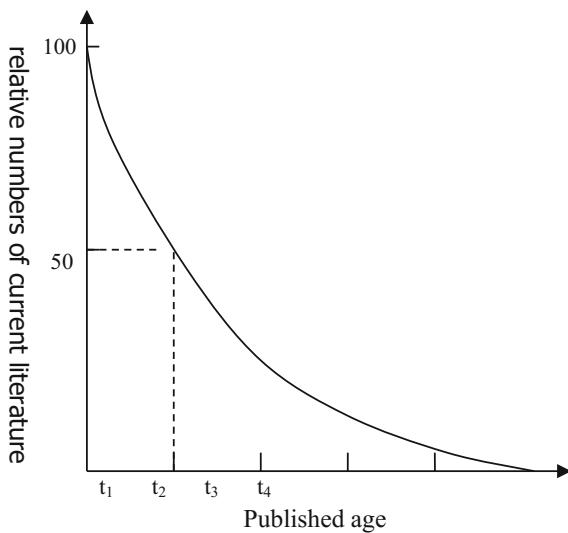
(2) Negative exponential model

As early as in 1958, Bernard proposed the negative exponential obsolescence model which was acquired by using the synchronic data. Commonly used the following function at present:

$$C(t) = Ke^{-\alpha t} \quad (3.1)$$

In this function, t as the literature published age (10 years for the unit); $C(t)$ express the cited frequency of the literature published in t years; K is a constant and varied with different subjects; e refers to the base of natural logarithms, equals 2.71818...; α is the obsolete rate of literature. If the publication age of literature as the horizontal axis, relative amounts of literature used currently as the vertical axis, could be portray a negative exponential curve, which known as the literature obsolete curves. The curve can be express literature obsolete process compared directly. From macro perspective, the negative exponent function described the literature obsolete laws, reflected the attenuation phenomenon of literature utilization ratio, accorded with the actual observations basically. However, throughout the period of literature exchanging, not all the literature utilization ratios are accord

Graph 3.1 The literature obsolete curve



with the negative exponent function law in every phase, and this formula could not directly reflect the relationship between the factors that affect obsolete and literature obsolescence. As a result, the curve also has some deficiencies, needs to be further modified and perfected (Graph 3.1).

According to the definition of half-life, T for half-life, by function (3.1)

$$\frac{1}{2} = \frac{C(T)}{C(0)} = e^{-\alpha T} \text{ take logarithms on both sides, then:}$$

$$T = \frac{1}{\alpha} \ln 2 \quad (3.2)$$

Function (3.2) given half-life, the obsolete rate, two indexes and their relationship. This model cannot be used to describe the time distribution of citation, because the citation distribution in a biggest citation fixed number of year not suitable for the model obviously.

② Burton–Kebler obsolescence model and the revised formula

In 1960, the American librarian R.E. Burton and physicist R.W. Kebler worked together, made a series research on the problems of science and technology literature obsolescence. They chose periodical literature in the field of nine subjects such as physics, chemistry, mechanical engineering, etc. And carried on the statistical analysis and calculation about the citation data, found that the curves depicted according to nine different subjects citation data is very similar in shape. It is a negative exponential curve as the decay curve of the radioactive element uranium 235. So they calculated a standard formula for these curves, later known as Burton–Kebler obsolescence model. The analytical formula is following:

$$y = 1 - \left(\frac{a}{e^x} + \frac{b}{e^{2x}} \right) \quad (3.3)$$

In the formula, $a + b=1$; y is the relative quantity of literature in one subject areas which are still in use after a certain period of time; x as the time and 10 years for unit. According to the above formula, when $y = 0.5$, the half-life of literature can be calculated. Burton worked out the half-life of nine subjects literature which he had researched on this basis. In field of physics and chemistry, the half-life of radioactive materials is related to its composition. According to this truth, Burton and Kebler pointed out that the periodical literature of one subject is composed of various components, many literature of related subjects have the possibility of using. In this way, half decay curve is bound to be dominated by literature with different content. Obsolete speed of published literature influences not only by its subject property, but also by other factors, especially the type and nature about itself. Therefore, Literature obsolescence law is restricted by many factors, Burton–Kebler obsolescence equation has not reflect the relationship between the various factors completely. In addition, this formula is more complicated, and difficult to calculate accurately.

In 1980, the Soviet scholars B.M. Мотылев studied the obsolescence Eq. (3.3) in detail, examined the differences between actual statistical sequence and the theoretical calculation, ξ^2 Test showed that the differences are very significant. And then, he revised the formula, and put forward the following correction:

$$y = 1 - \left(\frac{a}{e^{x-0.1}} + \frac{b}{e^{2x-0.2}} \right) \quad (3.4)$$

In the formula, $a + b = 1$

$$a = \frac{e^{1.8}(1 - y_x) - 1}{e^{0.9} - 1} = 3.4596 - 4.1447y_x \quad (3.5)$$

In the formula, y_x for the relative ratio (decimal) of measured total citation in 10 years.

$$x_{hl} = 10 \times [\ln(a + \sqrt{a^2 + 2b}) + 0.1] \quad (3.6)$$

In the formula, X_{hl} for the half-life(year).

Using the revised formula, real half-life of literature can be calculated, and automatic statistics of the literature cited data required to be done, operation is more simple and easy.

The two equations is consistent, just for citation age the former rules “0” while the latter is “1” only. Or, take the revised formula as the correction of delay in literature information publication. They also do not fit with the description of the citation time distribution.

③ Brookes accumulation exponential model

In 1970, B.C. Brookes put forward that the attenuated process of number of sci-tech periodical literature cited over time obey simple negative exponential model approximately from the perspective. The form also the same as formula 3.1, can be expressed as:

$$C(t) = Ke^{-\alpha t} \quad (3.7)$$

In the formula, $C(t)$ express the cited frequency of the literature published in t years; K is a constant and varied with different subjects; e refers to the base of natural logarithms, equals 2. 71818...; α is the obsolete rate of literature; t means the time.

In 1971, on the basis of negative exponential model of citation frequency, B.C. Brookes proposed the accumulation exponential model of literature obsolescence. From mathematical perspective, this has narrowed the modeling error caused by the random error when citation frequency statistics be made.

Assumed in formula 3.7: $b = e^{-\alpha}$

$$\begin{aligned} Y(t) &= C(t) + C(t+1) + C(t+2) + \dots = Kb^t + Kb^{t+1} + Kb^{t+2} + \dots \\ &= Kb^t(1 + b + b^2 + \dots) \end{aligned}$$

Set up: $M = K(1 + b + b^2 + \dots)$ and the accumulation exponential model is:

$$Y(t) = Mb^t \quad (3.8)$$

In the formula, $Y(t)$ is the number of papers published t years (including t years) ago from citations (literature cited age is equal and more than t); M is a constant, equal to the total amount of the citations; b is the obsolete coefficient and $0 < b < 1$; Half-life could be calculate easily:

$$T = -\frac{\ln 2}{\ln b} \quad (3.9)$$

B.C. Brookes has derived a very good method for calculating the b value:

In (3.7) when $t = 0, 1, 2, 3\dots$ the geometric decline sequence of literature obsolescence:

$$K, Kb, Kb^2, \dots, Kb^t$$

Then set k for the number of citations which published more than i years ago in a certain literature collection, and L for the number of citations which published less than i years ago in the collection. i can be obtained from the 3.7:

$$k = Kb^i + Kb^{i+1} + Kb^{i+2} + \dots = Kb^i(1 + b + b^2 + \dots) = (k + L)b^i$$

So as to be able to find out:

$$b = \left(\frac{k}{k+L}\right)^{\frac{1}{i}} = \sqrt[i]{\frac{k}{k+L}} = \sqrt[i]{\frac{k}{M}} \quad (3.10)$$

Therefore, as long as get the quantity of citation M through statistics and selected an i suitable, b can be calculated more accurately, according to the experience, i can choose 6, 7 or 8 generally.

④ A. Avramescu equation

For different quality and different kinds of literature obsolete trend, Romania literature metrologists A. Avramescu made some research, and putted forward the following mathematical model:

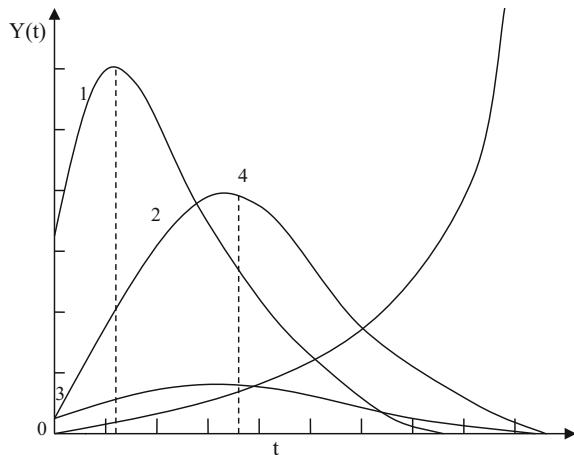
$$Y(t) = C_0(e^{-\alpha t} - e^{-mt}), m > \alpha \quad (3.11)$$

In the equation, $Y(t)$ for the citation frequency; C_0 for the transmission amplitude α for time decay rate of citation frequency; m for the initial increment.

Different C_0 , α , m corresponding to different literature obsolete trend.

- 1) When C_0 , α is bigger, m is small, describing the process that the literature has attached great importance when published, but rapidly obsolete cause people lost their enthusiasm soon. The total number of uses not much (Graph 3.2, Curve 1).
- 2) When C_0 is bigger, α , m is small, describing literature is widely accepted and obsolete slower. The total number of uses is much bigger (Graph 3.2, Curve 2).

Graph 3.2 Typical citation frequency distribution



- 3) When C_0, α, m is smaller, describing although literature obsolete speed is slow but most people has not paid importance. The total number of uses not much (Graph 3.2, Curve 3).
- 4) When m is almost zero, $\alpha < 0$, describing the paper has been recognized later by its high quality. At the beginning, it could not be used widely but has been known later, would not show obsolete for quite a long time as the use time increases gradually (Graph 3.2, Curve 4).

Can be seen, A. Avramescu equation describes the citation frequency distribution, when $m \gg \alpha$,

$$Y(t) \approx C_0 e^{-\alpha t} = C_0 a^t$$

In the formula, $a = e^{-\alpha t}$ for the obsolete coefficient.

The importance in results of A. Avramescu's study is investigated detailed in the using process of a single document, revealed the complexity and diversity of using literature and obsolete process. But, the mathematical model is only theoretical in the information dissemination process, needs to verify through a lot of statistical data.

(2) Commentary of domestic research on mathematical model

Domestic research on obsolescence mathematical model is effective, and a lot of great enlightening significance model are putted forward.

For example,

① Dynamic differential equation

$$\frac{dy}{dt} = ay(t) - by^\alpha(t) \quad (3.12)$$

In the formula, y is the relative amounts of literature in one subject field which is still in use after a certain time t . t is the time, or the publishing time. $\alpha \cdot \beta$ is constant and $\alpha > 0, \beta > 0$, the concrete relates subject characteristics, the types and nature of the literature; $\alpha > 1$; $\alpha \cdot y(t)$ be multiplication, the greater $y(t)$ showed the higher citation rate published in time t and more influence, the more chance to be used by others in the future. Influence on $\frac{dy}{dt}$ is positive, $-by^\alpha(t)$ is called a competitive item. Literature growth will reduce the use of literature. If the citation rate y of literature published in time t higher, the faster modify and update the content of literature and information which published in this period time, the influence of $\frac{dy}{dt}$ is negative, and $\alpha > 1$. The model is theoretical and not based on citation statistics, cannot be used in citation time distribution in principle. It is Verhulst model when $\alpha = 2$ particularly, and (3.12) into a parabolic equation, symmetry, do not conform to the asymmetry of the distribution of the citations.

② Series correction

Series correction is:

$$y = 1 - \sum_{i=1}^n a_i e^{-ix} \quad (i = 1, 2, \dots, n) \quad (3.13)$$

$$\sum_{i=1}^n a_i = 1 \text{ (Boundary conditions)}$$

Obviously, when $i = 2$, the above formulas is Burton–Kebler obsolescence model. Series obsolescence formula is closer to the actual statistic results shown by actual regression data, it has more ideal simulation result. Of course, this type of statistical processing need to use polynomial regression method, are more complex than the linear regression.

③ Literature obsolescence formula

Science and technology literature obsolescence law can describe by the following mathematical model:

$$R(t) = k_0(1 - e^{-\alpha t}) \quad (3.14)$$

In the formula, t for age, α for obsolete constant, $R(t)$ for the frequency of literature accumulated cited in t years (note: citation age \leq the total quality of citations t).

④ Mathematical model of Citation time distribution

If introduce of the obstacle factors in literature communication, namely to consider it needs a transfer and selection process when literature from be published to be referenced, the mathematical model of citation time distribution established:

$$R(t) = R_0 \left(1 - \frac{\beta}{\beta - \lambda} e^{\lambda t} + \frac{\lambda}{\beta - \lambda} e^{\beta t} \right) \quad (3.15)$$

In the formula, λ for literature obsolete coefficient and β citation block coefficient, $R(t)$ for cumulative citation quantity (literature age equal t years or less), R_0 for cumulative total citations.

After these models proposed, extensive research has carried on their equivalence and relations, and made a correction aim at the literature cite delay effect.

However, they are all experience or a priori model, has bad applicability and difficult to be used. Models abused for did not distinguish between citation time distribution data and literature obsolete data in the study.

(3) Research of the mathematical model Literature obsolescence and citation time distribution

Quantitative research to literature obsolescence and citation time distribution occupies an important position in the research of literature metrology. Many famous information experts have made researches home and abroad. The two studies are related closely and different, difficult to separate. A basic comb has given above mentioned, but it main focus on aspects of obsolescence, here the characteristics of citation time distribution model would be further discussed.

Citation time distribution mathematical model affected by the obsolete factors, but also influenced by hindering obsolete factors, can be described by the following models:

$$R(t) = R - Ce^{-\int f(t)dt} + De^{-\int g(t)dt} \quad (3.16)$$

In the formula, R for accumulative total citation; $R(t)$ as the accumulative total citation which the age of literature less than or equal to t ; $f(t)$ for time functions of obsolete coefficient; $g(t)$ as time functions of citation block coefficient; C, D is constant. The first exponential term is obsolete influence function, the next exponential term is blocking factors influence function.

When $f(t)$ and $g(t)$ is constant, and the obsolete coefficient and citation block coefficient are changing with time, the type can be further simplified as following:

Assuming $f(t) = \lambda$, $g(t) = \beta$, (3.16) became

$$R(t) = R - Ce^{-\lambda t} + De^{-\beta t}$$

Assuming $\beta/\lambda = d$,

Through the boundary conditions

$$\begin{cases} R(t)|_{t=0} = 0 \\ R(t)|_{t=\infty} = 0 \end{cases} \quad \begin{cases} R'(t)|_{t=0} = 0 \\ R'(t)|_{t=\infty} = 0 \end{cases}$$

Can be obtained $C = \frac{dR}{d-1}$, $D = \frac{R}{d-1}$, then citation time distribution model is changed to:

$$R(t) = R - \frac{dR}{d-1} e^{-\lambda t} + \frac{R}{d-1} e^{-d\lambda t} \quad (3.17)$$

Citation quantity in some time should be the derivative for $R(t)$:

$$R'(t) = \frac{\lambda R}{d-1} (e^{-\lambda t} - e^{-d\lambda t}) \quad (3.18)$$

(3.16) is dynamic mathematical model of citation age distribution; (3.17) is static mathematical model of citation age distribution; (3.18) is mathematical model of citation time distribution. (3.16) and (3.17) is accumulated distribution model, (3.18) are non-accumulated distribution model.

Derivation (3.18) and order

$$R''(t) = 0 \text{ then}$$

$$e^{-\lambda t} - de^{-d\lambda t} = 0$$

$$e^{-\lambda t} = de^{-d\lambda t}$$

Take logarithm on both sides, then:

$$\lambda = \frac{\ln d}{(d-1)t} \quad (3.19)$$

In (3.19), t has special meaning, it should be a maximum of age limit of the citation, then the citation distribution reach maximum.

To determine the value of λ and d need a equation, can get the relation between λ and d select points uniformly into the model, get λ and d simultaneous (3.19) respectively, then into model obtained model-fitting distribution, made grey correlation analysis with the actual statistical data as the reference sequence, set up the actual data model-fitting with the larger correlation finally.

3.3.2 Grey Dynamic Model (GM) and Obsolescence Index

This model is a posteriori model with a strong applicability, set up based on the complexity of literature obsolete factors and reality of incomplete information.

According to the definition of literature information and obsolescence law above mentioned, here GM and obsolescence index of literature information could be given also.

(1) Diachronic observation on GM and obsolescence index

Do diachronic observation on $X \in \chi$, X is the literature collection at the age of K actually.

Theorem 3.1 The literature collection $X \in \chi$ at the age of K for the object, its current measure is $P_K(X) = \chi^{(0)}(1)$, its measure in last year is $P_{K-1}(X) = \chi^{(0)}(2)$ By analogy, obtain the discrete function:

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(K+1))$$

The 1-AGO (An accumulation):

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(K+1))$$

Among them: $x^{(1)}(k) = \sum_{m=1}^k x^{(0)}(m)$

Then the literature diachronic obsolescence in accordance with the GM model:

$$GM(1, 1) \left\{ \begin{array}{l} \hat{x}^{(1)}(k+1) = (x^{(0)}(1) - \frac{u}{a})e^{-ak} + \frac{u}{a} \\ \hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) \\ \alpha = |a| \quad \text{aging rate} \\ T = \frac{1}{\alpha} \ln 2 \quad \text{half-life} \\ u \quad \text{grey action} \end{array} \right. . \quad (3.20)$$

Proving: according to the conditions set by theorem and literature obsolescence axiom, according to the grey system modeling method can obtain:

$$\hat{x}^{(1)}(k+1) = (x^{(1)}(1) - \frac{u}{a})e^{-ak} + \frac{u}{a}$$

$$\hat{x}^{(0)}(k+1) = (x^{(1)}(1) - \frac{u}{a})(1 - e^a)e^{-ak}$$

α reflected $\hat{x}^{(0)}$, the rate of change for $\hat{x}^{(1)}$, we define $|a| = \alpha$ for obsolete rate, reflect the rate of change about availability measure of X just right, is reasonable.

Assume the half-life is T, according Definition 3.8, so:

$$\begin{aligned} \frac{1}{2}\hat{x}^{(0)}(K+1) &= (x^{(1)}(1) - \frac{u}{a})(1 - e^a)e^{-a(K-T)} \\ &= (x^{(1)}(1) - \frac{u}{a})(1 - e^a)e^{-aK} \cdot e^{aT} \\ &= \hat{x}^{(0)}(K+1) \cdot e^{aT} \end{aligned}$$

Then $e^{aT} = \frac{1}{2}$, $T = \frac{1}{\alpha} \ln 2$

u is the grey action in GM, reflect the influence strength on the system by environment, we introduced directly, used to reflect the influence strengthen on the observe object X by environment (Ω , χ , P), very appropriate.

(2) Synchronic observation on GM and obsolescence index

Theorem 3.2 Selected an observation time and $X \in \chi$, besides $X = \{X(t) | t = 1, 2, \dots, K\}$, $X(t)$ is the literature collection of X types at the age of t. Assume $P(X(K)) = \chi^{(0)}(1)$, $P(X(K-1)) = \chi^{(0)}(2), \dots$, $P(X(0)) = \chi^{(0)}(K+1)$ the discrete function is:

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(K+1))$$

The 1-AGO (An accumulation):

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(K+1))$$

Among them: $x^{(1)}(k) = \sum_{m=1}^k x^{(0)}(m)$

Then the literature synchronic obsolescence in accordance with the GM model:

$$GM(1,1) \left\{ \begin{array}{l} \hat{x}^{(1)}(k+1) = (x^{(0)}(1) - \frac{u}{a})e^{-ak} + \frac{u}{a} \\ \hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) \\ \alpha = |a| \quad \text{agingrate} \\ T_1 = \frac{1}{\alpha} \ln 2 - \frac{1}{\alpha} \ln (1 - \frac{u}{a} (x^{(0)}(1) - \frac{u}{a})^{-1} e^{aK}) \quad . \\ \quad \text{averagesynchronichalf-life} \\ T_2 = \frac{1}{\alpha} \ln 2 \quad \text{distrubitionsynchronichalf-life} \\ \quad u \quad \text{greyaction} \end{array} \right. \quad (3.21)$$

Proving: According to the conditions set by theorem and literature obsolescence axiom, according to the grey system modeling method can obtain:

$$\begin{aligned} \hat{x}^{(1)}(k+1) &= (x^{(1)}(1) - \frac{u}{a})e^{-ak} + \frac{u}{a} \\ \hat{x}^{(0)}(k+1) &= (x^{(1)}(1) - \frac{u}{a})(1 - e^a)e^{-ak} \end{aligned}$$

α reflected $\hat{x}^{(0)}$, the rate of change for $\hat{x}^{(1)}$, we define $|a| = \alpha$ for obsolete rate, reflect the rate of change about availability measure of X just right, is reasonable.

According Definition 3.9, then:

$$\frac{1}{2} \hat{x}^{(1)}(K+1) = \hat{x}^{(1)}(K - T_1 + 1)$$

$$\text{Promptly } \frac{1}{2} ((x^{(1)}(1) - \frac{u}{a})e^{-aK} + \frac{u}{a}) = (x^{(1)}(1) - \frac{u}{a})e^{-a(K-T_1)} + \frac{u}{a}$$

$$\text{Then } e^{aT_1} = \frac{1}{2} (1 - \frac{u}{a} (x^{(0)}(1) - \frac{u}{a})^{-1} e^{aK})$$

By the sequence structure of theorem and literature obsolescence axiom: $a < 0$

$$\text{Obtained } T_1 = \frac{1}{\alpha} \ln 2 - \frac{1}{\alpha} \ln (1 - \frac{u}{a} (x^{(0)}(1) - \frac{u}{a})^{-1} e^{aK})$$

Resemble the method beg half-life in Theorem 3.1 can get:

$$T_2 = \frac{1}{\alpha} \ln 2$$

u is the grey action in GM, reflect the influence strength on the system by environment, we introduced directly, used to reflect the influence strengthen on the observe object X by environment (Ω, χ, P), very appropriate.

Deduction 3.1 Selected an observation time and $X \in \gamma$, and $X = \{X(t)|t = 1, 2, \dots, K\}$, then for average synchronic half-life(T_1) and distribution synchronic half-life(T_2) of X,

$$\text{satisfied: } T_2 = \lim_{K \rightarrow \infty} T_1$$

Proving: learned from Theorem 3.2:

$$\begin{aligned}\lim_{K \rightarrow \infty} T_1 &= \lim_{K \rightarrow \infty} \left(\frac{1}{\alpha} \ln 2 - \frac{1}{\alpha} \ln \left(1 - \frac{u}{a} (x^{(0)}(1) - \frac{u}{a})^{-1} e^{aK} \right) \right) \\ &= \frac{1}{\alpha} \ln 2 = T_2\end{aligned}$$

3.4 The Mechanism of Research and Analysis of Literature Information Obsolescence

Literature obsolescence is a complex social phenomenon and also a very complicated dynamic process. A lot of research on literature obsolescence for many years, the influencing factors is various and mechanism complicated, can be examined from the following several aspects generally.

3.4.1 Several Types of Literature Obsolescence

(1) Static and dynamic

Schreiber-Herbert made obsolescence is divided into static and dynamic, think literature growth is to distinguish the two different types of obsolescence.

- ① static obsolescence: the static obsolescence is the obsolete process not considered the effect of literature growth factors.
- ② dynamic obsolescence: the dynamic obsolescence is the obsolete process considered the effect of literature growth factors.

The different understanding for the two kinds of obsolescence is the lead of the apparent and true obsolete metrical dispute. M.B Line and A. Sandison (1974) believe that should revise the growth of obsolescence, for the obsolete curves based on the actual data shows “apparent” drop only. B.C Brookes (1970) found measure with no correction reflects the authenticity of obsolescence instead, E. R. Stinson and F. W. Lancaster (1987) confirmed this, L. Egghe proved indexes growth leads to more obsolescence in a synchronic condition, the diachronic condition take the opposite.

(2) Local and universality obsolescence

- ① local obsolescence: the local obsolescence refers to the utilization ratio of literature in one library or document information unit reduced. The main research method is literature management statistics analysis method.
- ② universality obsolescence: the universality obsolescence refers to the obsolescence in one subject or world. The main research method is citation analysis method.

3.4.2 Several Circumstances of Literature Information Obsolescence

The metabolism of literature is a very actual problem. Due to the rapid development of modern science and technology, the original immature theory was replaced by the more mature one with the passage of time; the imperfect method complemented by the more perfect one; backward technology is updated by the more advanced one; wrong and one-sided data corrected by more objective truth, and thus the old literature become void. The literature has much valuable several years ago would become obsolete increasingly with the development of science and technology, or even lose their vitality. Therefore, this kind of literature obsolescence is a common phenomenon. From the perspective of literature utilization, scientific literature obsolescence has the following situations:

(1) Obsolete information in literature

Literature content was proved to be unreliable by later, even wrong, this kind of literature cannot be used again, of course.

(2) Literature contains the information is obsolete

The content of Literature information is correct, but for has entered the broader social communication field and has been widely accepted, the original documents don't need to be used any more.

(3) Replaced by updated Literature

Literature content is correct, but replaced by the new literature of more new and comprehensive content, and also gradually less being used by the user as the extension of time.

(4) Use reduction caused by research interests falling

Literature content is correct, but no longer being used by the user for some reason (such as social needs) cause people research interests fall or attention transfer.

3.4.3 Factors Affecting the Literature Information Obsolescence

Many studies on the obsolete mechanism have shown that: The root causes of the scientific literature obsolescence is the growing and update of scientific knowledge. As we know, the development of science is not simple accumulation of facts one by one fact, but through improve, develop and update the countless creative theory continuously, to form today's knowledge base. As the continuation of the human society, the process of scientific knowledge growth is made up of "stack" and "update" two aspects about knowledge. The metabolism of the scientific literature is bound to cause due to the knowledge update. So, research to the scientific literature obsolescence is the exploration of scientific knowledge fixed speed essentially; Literature obsolete coefficient is a reflection of the fixed rate of scientific knowledge. Literature is a complex random aggregate, the literature metrology has asymmetry, and the irreversible characteristics of times, so literature obsolescence is a very complex phenomenon. Scientific literature obsolescence influenced by many factors, its mechanism can analyze from the following five aspects:

(1) Growth of literature

In the dynamic law of literature, growing and obsolete are two aspects of things, and they are illustrate the correction rate of scientific knowledge from different sides, that is the progress of science. So the literature obsolescence is linked with the growth firstly. We all know, the science made a breakthrough when the scientific literature increased mess, refers to the rate in superposition, improve and update the scientific knowledge is accelerated. At this time, the old literature with the content not perfect and comprehensive been forgotten gradually, reference frequency reduced rapidly. The knowledge old literature contained will not die, but its application value tends to zero gradually, becoming the "file". On the other hand, If literature growth slowly for some reason, obsolete curve gradient would flatten out at this time. According to Price statistics, basic theory research literature had a sharp drop when the first and the second world war broke out, obsolete curves flattens out clearly also. obsolete rate quick recovery almost at the same speed again until in 1926 and 1950 respectively after the war, return to pre-war values. Besides, due to the growing literature information, there is more literature for reference objectively, and people tend to reference the new literature in practice, so the literature obsolete rate would change also. In general, the faster the literature growing, obsolete speed up accordingly. The faster the literature growing, the half-life shorter. Therefore, the emergence and growth of new literature is one important factors contribute to the literature obsolescence.

(2) Literature subject characteristic

Literature contents belong to subject with different nature and characteristics, the difference of obsolete rate is very big. Generally speaking, the literature half-life of basic theory subject is long, and in the application technology subject the half-life is

shorter and obsolete faster relatively; the half-life of literature with a long history subject is longer than emerging subjects; half-life of subjects literature more stable is longer than the subjects literature going through a major change in content or technology. Some subjects, such as electron, metallurgy, chemical industry and other fields, the literature half-life is shorter because of the active research, abundance input resources and knowledge update quickly. And other subjects, such as animal taxonomy, geography and so on, its development is the accumulation of knowledge mainly but not correction, therefore these areas are more stable relatively, has a long half-life in generally and obsolete slowly for historical records can work for a long time. In some subjects, such as sociology and machinery manufacturing, etc., the number of literature in a rapidly obsolete and archives literature is same, they are between the former two.

(3) Different stages of subject development

In the whole period of subject development, every subject undergo different historical stages such as the birth, development and relatively mature. Even the same subject, the literature half-life is not the same in different development stage, the obsolete curve is not all accords with negative exponential curve also. When the subject at the birth and the beginning of development, the literature quantity increase exponentially, literature obsolescence accords with negative exponential function relation because the original literature is less, its corresponding obsolete curves show as the negative exponential curve. With the deeper subject research, subject development entering relatively mature period, the growth of literature can not keep the original exponential growth any more, the growth rate of literature become smaller, its corresponding obsolete curve flattens out, half life become longer. It reflects the science fixed rate slowed objectively, but does not mean the stagnation of scientific research. Instead, it marks the subject has entered a relatively mature stage on the one hand, the scientific value of literature reached a certain depth, make the extension of literature use life. On the other hand, suggests that the result of the scientific activities mainly lies in the accumulation of knowledge rather than the correction. When the number of knowledge accumulation once reached a certain amount, there will be a leap from the quantitative to qualitative changes, and make the subject into a new height and level, to derive new branch at the same time likely, and make the literature growth presents exponential rate; then literature obsolete curve is restored to negative exponential curve.

(4) Type and nature of literature

The obsolete speed of literature depends not only on literature subject content, but also related to its type and nature. Usually, science book has a longer “half-life” than journal articles, science and technology reports, meeting documents, etc. And classic works has a longer half-life than general works, theoretical publication has a longer half-life than Communications informative publications, articles to discuss longer than introductory articles, peer-reviewed literature obsolete slower than research papers, and so on.

(5) User requirements and information environment

User's characteristics and information environment quality could not be ignored which are the factors influencing the obsolete documents. For example, different quality of user has different documents requirement. The backbone of the scientific research workers interested in the latest literature, while the new workers need to background information and historical documents. Even the same reader, he demand for literature also has different characteristics in different periods for different research purposes. The literature useless for someone is still useful for else, such as professional workers in history. So from view of knowledge users, the fixed number of year on literature use is vary from person to person. Different countries or regions on the fixed number of year of literature use is not same entirely. Science developed countries interested in new literature published recently, while scientific and relatively backward countries need to consult literature of a period of time to get the experience of other countries.

3.5 The Application of Literature Information Obsolescence Law

Literature information obsolescence law is one of the basic law of literature information flow. It reveal the law of the document and information work and the characteristics of scientific development from the view that the literature utilization decreased with the passage of time.

A. Imihajlovic thought the study of literature obsolete issues is meaningful because of the literature utilization in the future could be forecast in a reliable way which on the basis of analysis of the literature obsolete. Then guiding the whole literature organization in some degree further. Therefore, the study of literature obsolescence law has an important significance both in theory and practice.

3.5.1 *The Application in Document Information Management*

(1) To guide the eliminate and optimize collection

In the document and information work, eliminate obsolete documents timely is an important link to optimize library collection and improve the efficiency of document information service. A good literature eliminated can be solve the crisis of library space effectively, on the one hand, also improve the efficiency of information retrieval. The detection rate of useful literature will be increase because of separated the obsolete documents from the useful literature. Due to the rapid growth of the

scientific literature, library and information organizations are almost encounter space crowded problems which bring serious difficulties to the circulation without exception. Their immediate task is eliminate obsolete documents. Therefore, “please go” obsolete documents, free up valuable space, make sure useful literature is not submerged in the obsolete literature. That is crucial to determine the specific applicable period of the literature, and the literature obsolete data has important practical action in eliminate documents.

(2) Provide evidence for formulate reasonable literature work principles

Research on literature obsolescence law provides a scientific basis for make working principle of various literature has different half-life. For subject or professional has short “half-life”, documents work should be take the time and more efficiency; strengthen the literature reports, carry out SDI services; implement open access journal of these subjects; The literature of obsolete, can be saved as microcopy or put it in the closed stack room. Fixed number of the open literature year be determined according to obsolete data. To develop a reasonable management system, etc.

(3) Evaluation of the literature

Obsolete index of the scientific literature is very necessary and beneficial for mastering the characteristics of literature, identify the limitation of literature, to determine the literature value.

3.5.2 The Application in the Study of Science and Technology

Literature is an existence form that science can touch, also the main information source which scientists are most interested in and depended on, from which they learn the results of predecessors or others. Therefore, research on the obsolescence law which is the significant characteristic of literature can show the speed of scientific development, revealing the rules of the development of science, reflects how the human inherit and development of scientific knowledge. Literature Obsolete is associated with its subject nature. Subject nature and the development stage could be judged according to the literature obsolete index data. If research on literature obsolete properties of a particular technology field, the development speed, suitable time, and term may be eliminated of this technology can be roughly determined. Therefore, it is indeed a new important way to study on science of science and technology to reveal the development of science and technology process and regularity by study on literature obsolescence law.

Chapter 4

Concentration and Scattering Distribution of Literature Information: Bradford's Law

4.1 Background of Bradford's Law

4.1.1 Founder: Bradford

Samuel Clement Bradford (1878–1948) was a world-famous philologist and chemist. He was born on October 1, 1878 and graduated from the University of London. He obtained a doctorate in science in 1922. He founded Bradford's scattering law, which is the main foundation of bibliometrics.

Bradford did not major in library science and literature, but he loved and was engaged in library work for a long time. He served as the leader of the science library in South Kensington until his retirement in December 1937. This important lending library later developed into a national science and technology library, which is one of the largest science and technology libraries in the world.

Bradford actively participated in the activities of the international library community. He participated in the editing of the “Universal Decimal Classification” (UDC) and published numerous articles. He passionately advocated for and promoted the use of UDC. He also later served as the director of the International Committee of Taxonomy.

Bradford and Pollard co-created the British Society for International Bibliography (BSIB) in 1927 as the English branch of SIB. Bradford first served as the vice president of the society, and then as president in 1945. SIB was later renamed as FID, and BSIB and the Society for Special Libraries and Intelligence Agencies (Aslib) were combined in 1948. Bradford became the honorary editor of the journal of the said society since 1939. He was a member of the British Library Society and served on its board for many years. He remained in the leadership position of the National Central Library and Finance Committee in the United Kingdom.

Bradford was a prolific writer. He did not only edit and publish many catalogues but also produced numerous works on classification theory, practice, and cataloging

theory. His primary works included “Theory of Science and Applied Scientific Work of Classification,” “Original Classification,” “Science and Technology Organization Directory,” and “Cataloging.” He also published many papers, including “Universal Decimal Classification Origin, Purpose, Structure, and Use,” “Fifty Years of Literature,” “All Literature Work of Science and Technology,” and “The Information Source of Special Discipline.” Bradford was also a physical chemist, who performed many related studies and wrote numerous papers on kinetic theory of liquids and solutions, colloidal theory, and raising topics such as roses.

Bradford focused primarily on the study of philology in his later years. He completed his monograph “Literature Work” in 1948, which became one of his most influential writings. Bradford conducted a systematic theoretical summary of literature work in his book. He explored the nature, origin, and purpose of literature work. He also elaborated on the alphabetical subject index, UDC, indexing work, abstract work, library science, and a wide range of other issues. He focused on a systematic discussion on the literature decentralized law, which largely contributed to the birth of bibliometrics.

Bradford died on November 14, 1948. Several magazines, such as the “Library World,” published articles in his memory. “Documentation,” a magazine sponsored by Aslib, published commemorative articles and research papers in a special series in 1977 to commemorate the 100th birth anniversary of Bradford, which also provided excellent materials for the study of his academic thought and scientific contributions.

4.1.2 *Background of Bradford Law's*

The development of Bradford's literature scattering law is not accidental, but has a certain objective background.

- (1) Document scattering is a common objective phenomenon. In scientific research and documentation work, Bradford deeply acknowledged the scattering of scientific literature. He determined that papers in a certain discipline would commonly disperse in magazines or journals in other disciplines. For example, papers on cybernetics can be published in neuroscience journals, papers on mechanical hearts can appear in physics journals, and genetics theses can be dispersed in agronomy journals. The scattering of scientific literature is a widespread phenomenon, but determining how to identify their scattering law is the most critical issue. Bradford believed that the literature dispersion law could be deduced qualitatively from the principles of scientific unity and could be derived from the number of papers contained in relevant journals.
- (2) Principle of scientific unity. Although different scientific disciplines exist, they are still part of a larger body of knowledge. This principle lays the ideological foundation for Bradford's law. Bradford believed that given the principle of

scientific unity, every science and technology discipline would be relatively associated with any other discipline. Therefore, a phenomenon occurs wherein a type of literature in a discipline appears in the journals of other disciplines. This condition is an important foundation to understand the literature scattering law of Bradford. However, the following questions must be addressed: how does the relationship among disciplines affect the relationship among literature types? what is its ratio? what are the scattering characteristics? Bradford summarized these characteristics in a study, which stated that a specialized journal could contain useful papers in other disciplines. A paper that is useful to an expert may not only appear in a journal of that expert's discipline, but may also occur in the journals of other disciplines at certain times. The number of journals of other disciplines decreases if these journals exhibit close relationships with the journals to which the expert belongs to, which evidently presents an inverse relationship. For example, papers related to library automation must be published in library science journals. Simultaneously, they may appear in magazines on electronic technology, modern technology, and data processing. The sizes of the journals of other subjects depend on the closeness degree between these journals and library automation. For example, 100 papers are about library automation and 60 of which are attributed to library science; thus, the remaining 40 papers will be scattered in journals from other disciplines. The number of these journals depends on how close their relationships are with library science. If the relationship is sufficiently close, then each journal issues 4 papers. These 40 papers will be scattered in 10 journals. Otherwise, each journal only issues an average of 2 papers, and these 40 papers will be distributed in 20 journals.

Bradford also believed that several journals would always have contents that were closer to certain disciplines, whereas other journals would always have contents unrelated to the subject. In particular, for the core journals of a discipline or the few journals that contain many contents of this subject, the papers that involve the subject content must be more than those that deal with the content of other subjects. Hence, the idea of dividing a journal into several regions is established. Bradford's approach is described as follows. The journals are divided into several regions according to the number of papers of the journals in a given subject. The journal number increases along with the decreasing number of papers published in each region, which is similar to the previously described inverse relationship.

- (3) Literature statistical study is the foundation of Bradford's law. Since the 20th century, several scholars have performed studies on literature statistics, which has positively affected the formation of Bradford's law. Bradford's comprehension of the literature law started from the literature statistics because of this practical requirement. Abstract magazines will have a storage function by the 30th century, and the reporting and retrieval of literature will increase because of the rapid increase in the number of scientific literature. However, a common phenomenon is the existence of redundant and missing documents in these

abstract journals, which has resulted in Bradford's speculation. Is there an underlying relationship between desperation and incompleteness? Bradford began to explore the literature system to identify the reasons and the inherent law of literature. Moreover, he used the quantitative method to conduct his research work; in particular, data were systematically summarized and analyzed from literature statistics, and then the quantitative law of the literature information flow was derived. Bradford conducted many statistical investigations, mastered the scattering characteristics of literature, and determined several inherent laws in his long-term studies on scientific literature. He also deduced these literature statistics in terms of mathematics and obtained the same results with the theoretical supposition, which laid the foundation for the formal establishment of Bradford's scattering law.

4.2 Formation of Bradford's Law

4.2.1 *Proposal of Bradford's Law*

We refer to papers on a particular subject, discipline, or field as "related papers." Related papers in a journal are not subject to uniform distribution, but have clear concentration and are affected by the scattering law. Experts have been aware of this phenomenon, but extensive studies on this topic have only begun in the mid-20th century. The famous British literature scientist Bradford first established the literature scattering law and proposed the well-known Bradford's law of scattering, which is abbreviated as Bradford's Law.

Bradford selected applied geophysics and lubrication as objects, and then organized his colleagues at the Library of Science Museum to prepare statistics on the documents within these two subjects (i.e., a total of 490 journals and 1727 papers). The distribution is shown in Table 4.1, where A is the number of journals, B is the number of papers in each journal, C is the sum of A , D stands for the accumulation of $A \times B$, and E is $\log_{10}C$.

Bradford aligned the journals according to their numbers of papers in decreasing order and then applied three different methods in his analysis.

- (1) Regional analysis. Bradford divided the journals of the two aforementioned disciplines into three zones according to the annual average papers: (a) journals with more than four papers; (b) journals with more than one paper, but less than four papers; and (c) journals with only one paper. The partition results are listed in Table 4.2. After the analysis, he concluded that the number of papers in each zone was nearly congruent, and the numbers of journals in successive zones were distributed in a geometric progression, where the common ratio was approximately 5.

Table 4.1 Distribution of literature in applied geophysics and lubrication journals

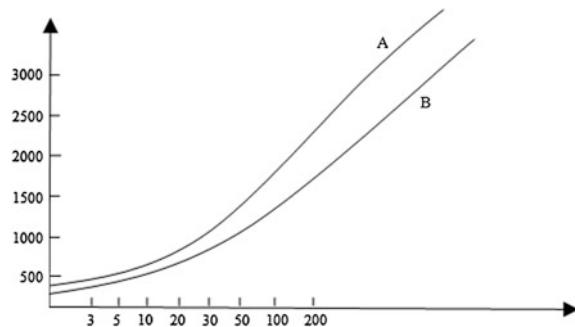
Applied geophysics					Lubrication				
A	B	C	D	E	A	B	C	D	E
1	93	1	93	0	1	22	1	22	0
1	86	2	179	0.301	1	18	2	40	0.301
1	56	3	235	0.477	1	15	3	55	0.477
1	48	4	283	0.602	2	13	5	81	9.699
1	46	5	329	0.699	2	10	7	101	0.845
1	35	6	364	0.788	1	9	8	110	0.903
1	28	7	392	0.845	3	8	11	134	1.041
1	20	8	412	0.903	3	7	14	155	1.146
1	17	9	429	0.954	1	6	15	16	1.176
4	16	13	493	1.114	7	5	22	196	1.342
1	15	14	508	1.146	2	4	24	204	1.380
5	14	19	578	1.279	13	3	37	243	1.568
1	12	20	590	1.301	25	2	62	293	1.792
2	11	22	612	1.342	102	1	164	395	2.125
5	10	27	662	1.431					
3	9	30	689	1.477					
8	8	38	753	1.580					
7	7	45	802	1.653					
11	6	56	868	1.748					
12	5	68	928	1.833					
17	4	85	996	1.929					
23	3	108	1065	2.033					
49	2	157	1163	2.196					
169	1	326	1332	2.513					

Table 4.2 Journal distribution in different zones

Zone	Number of papers	Applied geophysics		Lubrication	
		Number of journals	Number of papers	Number of journals	Number of papers
1	>4	9	429	8	110
2	1–4	59	499	29	130
3	1	258	404	127	152

- (2) Image observation. Bradford depicted the curve distribution of the papers in these two journals (Fig. 4.1). The x -axis shows the logarithm of the cumulative number of journals, whereas the y -axis shows the cumulative number of papers. He observed that if the flexion in the lower part was omitted, then the image

Fig. 4.1 Distribution curves of related papers in the two subjects: A—Applied geophysics (1929–1932) and B—Lubrication (1931–1932)



would become a straight line. Bradford concluded that the number of journals in Zones 1, 2, and 3 also proportionally and simultaneously increased under this context.

- (3) Mathematical derivation. Bradford also performed a mathematical derivation of his empirical data and formula.

Let m_1 , m_2 , and m_3 be the number of papers in three different zones; n_1 , n_2 , and n_3 be the number of journals in the corresponding zone; r_1 , r_2 , and r_3 be the average number of papers of each journal in the corresponding zone; and evidently, $r_1 = \frac{m_1}{n_1}$, $r_2 = \frac{m_2}{n_2}$, and $r_3 = \frac{m_3}{n_3}$. The rank order principle is $r_1 > r_2 > r_3$, whereas $n_1 < n_2 < n_3$.

Bradford assigned the same number of papers in each zone:

$$n_1 r_1 = n_2 r_2 = n_3 r_3 = m_1; \text{ thus, } n_2/n_1 = r_1/r_2 = a_1 \text{ and } n_3/n_2 = r_2/r_3 = a_2.$$

In the preceding formula, a_1 and a_2 are the constants with values larger than 1. However, Bradford assumed that $a_1 = a_2 = a$. Thus, we can obtain the following:

$$n_2 = a_1 n_1 = a n_1 n_3 = a_2 n_2 = a_1 a_2 n_2 = a_1^2 n_1 = a^2 n_1. \text{ Therefore, } n_1 : n_2 : n_3 = 1 : a : a^2. \quad (4.1)$$

Bradford believed that the derivation of Formula 4.1 was consistent with the result of the regional analysis.

In the statistical studies, Bradford determined that although disciplines are different, their papers have the same distribution in corresponding journals. The regular fact shows that Bradford has drawn the conclusion of the scattering law of literature. He published a paper entitled “Sources of Information on Specific Subjects” in the column of “Books and Documents” in the “Journal of Engineering” in January 1934. This paper was the first to quantitatively describe the empirical law of literature scattering, and eventually became a popular material on the historic significance of bibliometrics.

4.2.2 Establishment of Bradford's Law

Although Bradford evidently presented the scattering law of scientific literature early in 1934, his research results did not attract the attention of the public for a long time. His monograph entitled "Documentation" was published approximately 14 years later, immediately after he died in 1948. This monograph included his popular paper published in 1934. It was then expanded as Chapter IX of a book entitled "Documentary Chaos," which attracted the attention of many scholars, including Vickery.

Vickery was a British documentalist who worked at the research laboratory of the British Imperial Chemical Company. He was the first expert to publish papers on Bradford's law, in which he did not only fully affirm Bradford's work, but also referred to this distribution as "Branford's scattering distribution" and to its result as "Bradford's scattering law." He also creatively afforded several amendments and supplements to the law. The research results of Vickery have not only contributed significantly to uniting the distributions of figure and law in structure and making the law more integrated in form, but they have also enriched the content of Bradford's distribution theory, which has significantly contributed to the establishment and development of Bradford's law. Afterwards, Bradford's law was widely recognized by the Academia of International Library and Information Science, in which the work of Vickery undoubtedly played a decisive role. Thus, although Bradford presented the law, its discovery, establishment, and circulation could be attributed to Vickery.

In addition to Vickery, many other documentalists have conducted studies on Bradford's law. Among them, F.F. Leimkuhler and B.C. Brookes were the most popular. The former was a scholar who contributed to the development of zone description; the latter was the one who formulated and developed image analysis. The theory, mathematical description, and application of Bradford's law gradually became perfect because of the joint efforts and contributions of many scholars, which also made experts realize the sense of the law and led to its establishment and rapid development.

4.3 Basic Content of Bradford's Law

4.3.1 Elaboration of Bradford's Law

The basic principle of Bradford's law is composed of two parts: zone description and image description.

- (1) Zone description. Bradford wrote in his book "Documentation Work" that if scientific journals were arranged according to the number of papers that belonged to a subject in decreasing order, then journals could be divided into a

core zone and several areas with the same number of papers as the core zone. Thus, the numbers of journals in the core and subsequent zones are distributed as $1 : a : a^2 \dots$

The literal representation of the law is based on the zone analysis method for ranked journals. If the arrayed journals are divided according to the number of papers within a certain period (i.e., a year) into three zones and each zone is provided with an equal number of papers (i.e., each zone equally obtains $1/3$ of the total number of papers), then the articles in the first zone (the core zone) is from n_1 journals that are less in quantity but are the most efficient. The second zone (the relevant area) contains n_2 journals with a large number and medium efficiency. The third zone (the peripheral zone) consists of n_3 journals that are the largest in number but are the least efficient. Hence, the number of journals in the three zones exhibits the following relationship:

$$n_1 : n_2 : n_3 = 1 : a : a^2 \quad (a > 1), \quad (4.2)$$

where a is Bradford's constant or the “proportional coefficient.” This constant is approximately 5 in the case of the data used by Bradford. The empirical formula, Formula (4.2), is a zone representation of Bradford's law.

Suppose that 248 journals on a certain subject exist, with 660 published papers. Bradford's scattering law states that the “core” journals in the first zone has only 8 types, the “relevant” journals in the second zone are $8 \times 5 = 40$, and the number of “peripheral” journals in the third zone is $8 \times 52 = 200$. Each zone will also publish 220 articles as shown in Table 4.3.

Table 4.3 shows that the number of papers of the journals in each zone gradually decreases, whereas the number of core zones gradually increases. The core zone contains an average of 27.5 papers in each journal, with the highest information density, followed by the relevant zone with an average of 5.5 papers in each journal, and then by the peripheral zone with an average of only 1 paper in each journal. From the investigation of Vickery, Formula (4.2) can be generalized as follows:

$$n_1 : n_2 : n_3 : \dots = 1 : a : a^2 : \dots \quad (4.3)$$

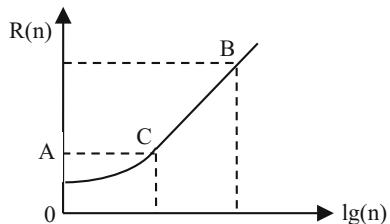
That is, the equation adapts the law that is suitable for the three zones to that applicable to multiple zones.

- (2) Image description. Bradford's law can also be characterized by using images. If we take the logarithm of the number of journals ($\lg n$) as the horizontal

Table 4.3 Zone distribution of Bradford's law

Zone	Number of journals	Number of papers
1	8	220
2	40	220
3	200	220

Fig. 4.2 Curve distribution of Bradford's scattering law



coordinate and the corresponding cumulative number of papers as the vertical coordinate to describe an image, then we can obtain a curve. We call this curve “Bradford’s scattering curve.”

The early graph of Bradford’s law is frequently expressed as the form shown in Fig. 4.2. The scattering curve AB consists of two parts: a rising curve AC that corresponds to the core zone and a line CB that corresponds to the subsequent zones. Later studies show that the inflection point C is the demarcation point for the core area.

Bradford drew another conclusion from the perspective of image as follows:

$$n_1 : (n_1 + n_2) : (n_1 + n_2 + n_3) = 1 : b : b^2.$$

Let $n_{1-2} = n_1 + n_2$, $n_{1-3} = n_1 + n_2 + n_3$. We can then obtain the following equation:

$$n_1 : n_{1-2} : n_{1-3} = 1 : b : b^2, \quad (4.4)$$

where n_1 stands for the number of journals in the core area; n_{1-2} is the cumulative sum of journals in the core and second zones; n_{1-3} indicates the sum of journals in all the three zones, i.e., the total number of journals N ; and b is the scattering coefficient. The value of b differs from that of a in Formula (4.2).

The application of Formula (4.4) is more extensive than that of Formula (4.2). Brooke’s well-known equation is the generalization of Formula (4.4). Another generalization to Bradford’s law by Vickery also demonstrates the validity of the conclusion obtained by image description.

- (3) Comparisons of zone and image descriptions. Vickery and Wilkinson pointed out that several contradictions could not properly unify the two forms of Bradford’s law, namely, zone description and image description. In particular, Wilkinson assumed that image description would be more accurate than a literal one after conducting an in-depth study in 1972. Several Chinese studies have also pointed out that the results of the zone expression, i.e., Formula (4.2), and the image description, i.e., Formula (4.4), are not only impossible to be consistent, but are also impossible to set up. This method is proven as follows.

First, if Formula (4.4) is established, then $n_1 : (n_1 + n_2) : (n_1 + n_2 + n_3) = 1 : b : b^2$. The following deductions are also obtained:

$$\begin{aligned} n_2/n_1 &= b - 1 \\ n_3/n_1 &= b(b - 1) \\ n_3/n_2 &= b \\ n_2/n_1 &\neq n_3/n_2 \end{aligned}$$

Thus, Formula (4.2) cannot be established, and $n_2/n_1 = n_3/n_2 = a$. Furthermore, the following can be deduced:

$$\frac{n_1 + n_2}{n_1} = \frac{1+a}{1} = \frac{1+2a+a^2}{1+a}, \text{ while } \frac{n_1 + n_2 + n_3}{n_1 + n_2} = \frac{1+a+a^2}{1+a}.$$

Hence, $\frac{n_1 + n_2}{n_1} \neq \frac{n_1 + n_2 + n_3}{n_1 + n_2}$.

Formula (4.4) cannot be established, which also proves that these two types of description in mathematics are not equivalent.

The preceding discussions show that Formula (4.2) is roughly summarized by making approximations for statistical data, which is an approximate empirical method. We also can search in an array of statistical data, where nearly every group of data is approximately subject to Formula (4.2).

The image description method is based on the same statistical data used in zone description, but it takes the logarithm of the number of journals and obtains Formula (4.4) under the approximate requirement that the papers in the three zones are equivalent. Therefore, image description is feasible from a mathematical perspective. Moreover, image description is also near the actual distribution of literature in practical applications, and even its approximation degree is more accurate than that of Formula (4.2), which has also been proven by Wilkinson using four sets of empirical data. However, zone description has two approximate requirements: (1) the number of papers in the three areas should be approximately equal and (2) the ratio of the number of regional journals in successive zones should be approximately equal to the constant value. These requirements are relatively rough. Nonetheless, the law of information measurement shows that any mathematical model can only provide an approximation of the actual scenario. Therefore, neither Formula (4.2) nor (4.4) can accurately fit into the statistical data. They only approximately present the law of literature distribution.

Bradford expounded on the apparently contradictory formulas in related papers and monographs regardless of the conclusion of the image description. He insisted that Formula (4.2) should be the expression of the literature scattering law. However, he also retained the two contradictory methods and conclusions. These two formulation types are consistent with the actual condition to a certain extent. Therefore, experts frequently take both zone and image descriptions as the basic contents of Bradford's law in terms of validity and utility, as well as in the applications of Formulas (4.2) and (4.4).

Notably, Brazilian scholars Maia et al. pointed out in 1984 that the literal expression result was consistent with the graphic result that stemmed from the

former one. They deduced a generalized empirical formula, which was akin to the modified version of Brooks presented by Bradford that started from the literal expression of Bradford's law. These researchers also tested the formula on the two groups of data originally used in Bradford's research. Their results show that the theoretical curve based on the zone description of Bradford is nearly in line with the empirical curve based on the graph description. This scenario shows no discrepancy between the two forms of expressions, but consistency is observed.

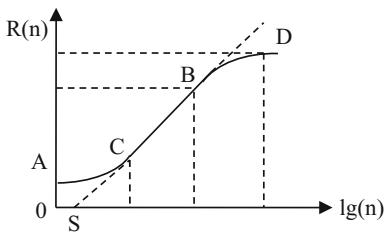
4.3.2 Consistency of Theory with the Practice of Bradford's Law

- (1) Theoretical explanation to Bradford's law. Why does the distribution of scientific literature follow Bradford's law? This answer to this question can be explained in terms of scientific development and literature activity.

The development of science always follows a certain rule. However, modern science and technology intersect with each other, and small subjects are inclined to develop into larger ones. Bradford summarized the literature scattering law from the law of science unity. After a new subject emerges, the first batch of literature will be published in several journals and will attract an increasing number of documents. Subsequently, several new journals will be released. After periods of development, consolidation, competition, and elimination, a certain number of journals will be dedicated to the subject with the largest literature and the highest quality. Authors will be willing to publish their papers in these journals, thereby leading to the appearance of "core journals," which is presented as the rising power curve in Bradford's graph. This phenomenon is the "accumulative effect" in the distribution of literature. However, the papers in this subject can be published in a number of other journals, thereby producing the phenomena of centralization and scattering of literature, given that the development of objects is frequently not determined based on a single factor. As mutual penetration occurs between science and technology, many new disciplines will emerge, along with other edge disciplines that will publish their papers in these journals. A "defined factor" is bound to take effect to inhibit the infinitely increasing number of core journals and given the limited capacities of journals.

To balance these documents, editors have to formulate plans and policies to control the number of documents. Many new journals emerge to fulfill the growing needs of publishing papers. Thus, with the passage of time, the number of journals in this subject will be in direct proportion to the number of papers, thereby making the graph of Bradford's law linearly ascending. The demarcation point between the curve and the line objectively reflects the effect of the defined factor. In addition, the graph of Bradford's law will continue to rise along the curve.

Fig. 4.3 Modern Bradford's scattering curve



- (2) Groos droop. In 1967, Groos used Keenan–Atherton's data and determined that Bradford's curve would not stretch endlessly after entering a straight part, but instead, would bend and sag, thereby splitting the figure for Bradford's law into three parts: the rising curve, the line, and the bent droop, as shown in Fig. 4.3. We call the bent droop (BD) "Groos droop."

With the help of "A Bibliography on Information Science and Technology" codified by Santa Monica (an American company), Canadian scholar Pope gathered statistics on 7368 articles published in a collection of 1011 journals from September 1964 to January 1970. He drew the corresponding graphics between $R(n)$ and $\lg n$, which were proven to be composed of three parts. In particular, when $1 \leq n \leq 10$, we obtain the curve AC of $R(n) = an^\beta$. When $10 \leq n \leq 200$, we obtain the line CB , which is represented as $R(n) = K \lg n/s$. When $n > 200$, the line starts to bend downward and we obtain the curve BD . Theoretically, the practical meaning of the droop phenomenon is currently being discussed in academic circles.

We can obtain relevant data from Fig. 4.3. If these data are substituted into Brookes' formula $R(n) = N \lg n/s$, then the collection of all the journals in "A Bibliography on Information Science and Technology" should be 1418 pieces rather than 1011, and all the relevant papers should be 8900 pieces instead of 7386. From the data, the numbers of journals and literature from Pope's actual statistics are less than the theoretical values, accounting for only 72 and 83% of these values, respectively. Thus, Bradford's curve will sag because of the insufficient volume of journals and documentations, which fails to represent all literature. The Groos droop merely reflects the differences between the expected and actual values.

- (3) Reasons for the differences between theory and actual practice. Bradford's law and actual literature distribution exhibit good consistency; however, differences still exist. The underlying causes of these differences fall within two aspects: the defects of a theory itself and the distortion in reflecting objective reality. The literature scattering curve of Bradford's law presents an S-shaped trend because of the occurrence of Groos droop. However, Brookes' mathematical formula consists only of two parts, which correspond to the rise of the curve and the straight portion, without a mathematical description for the droop; thus, a large error occurs. Therefore, the formula is an approximation in itself, and

deviation is fundamental, but a flaw is found with the further study of Bradford's law. The theory and mathematical expressions of Bradford's law will be continuously supplemented and amended to reflect the objective laws of literature distribution in a more accurate manner. In addition, practical activities, as well as the statistical insufficiency of a number of related journals and documents, cause the last part of Bradford's graphic to bend downward, which can be concluded as follows. First, strict prerequisites, which refer to rigid disciplinary boundaries, must be fulfilled when using Bradford's scattering law. However, the cross-penetration of modern science and technology, as well as the increasing emergence of interdisciplinarity, has led to a statistical error, as evidenced by scholars who tend to classify literature in computer science from another perspective, although it should belong to information science. Second, bibliographies, indexes, abstracts, and books are generally referred to during the current application of Bradford's law. The codification of such referencing tools demonstrates selection and processing courses toward original documents, which is likely to change the original state of objects and leads to the generation of variation. Third, interference, such as journals being renamed or ceasing publication, will be encountered, along with the duplication of publication during the gathering of statistics of various data, thereby causing inaccurate reports. Given these factors, when n reaches a certain value, it will force the graphics to deviate from being straight and drop, which results in the differences between theory and practice.

4.4 Development of Bradford's Law

In this section, we discuss the development problem of Bradford's law and its distribution theory starting from the research efforts of several important representatives based on a review of its development process.

4.4.1 *Development Process of Bradford's Law*

Over half a century, Bradford's law has developed from simple to complex. It can be divided into three stages according to the historical era and research conditions, as follows.

- (1) Creation stage. The creation and dissemination phases before the 1960s were based on Bradford's significant essay published in 1934. However, during the early period between 1934 and 1948, Bradford's law did not arouse the attention of the academic community. In 1948, the publication of Bradford's monograph literature work changed the situation, i.e., his theory began spreading and attracted the attention of some authors. For example, in the

publication year of the monograph, the British scholar Vickery first studied his literature distribution theory. Although Bradford's related papers and monographs are undoubtedly necessary foundation and important symbols of the formation of this law, Vickery's timely recognition, correction, and creative promotion played an extremely important role in the formal establishment and widespread of Bradford's law. From 1948 to 1960, the research reached a relatively silence stage.

- (2) Theoretical stage. The 1960s was an important historical period for Bradford's law and distribution theory. With unprecedented active research and an increasing number of papers, a remarkable "Bradford's favor" gradually formed. During this period, the development of Bradford's law focused more on theoretical research, particularly on law test, revision, and improvement. Attempts to find a more precise empirical distribution formula to enrich and improve its scientific nature and accuracy were conducted. Famous and important research works and achievements during this period included that of British statistician Kendall, who first proposed a version of Bradford's law that was highly similar to Zipf's law in structure based on his research on the application of Bradford's law to operational research bibliography, which was soon endorsed by the academic community. Since the mid-1960s, more in-depth and systematic theoretical discussions of Bradford's law have occurred. In 1967, Lyme Cooley analyzed Bradford-type data, derived new distribution formulas using applied statistical theory and methods, and developed a regional description method. In the same year, Groos performed studies and improved the structure of Bradford's curve, thereby determining that research on droop was highly significant. A year later, Brookes discussed the practical application technology of scattering law, proved that the actual distribution function should be $R(n) = k \lg n$, and developed the image description method, which paved the way for the practical application of Bradford's law. With further research on laws, two initial schools of thought emerged: the regional school represented by Leimkuhler and Goffman and the graphic school represented by Brookes and Kendall.
- (3) Comprehensive development stage. Since the 1960s, the development of Bradford's law experienced a transition from purely theoretical research to widespread applications, thereby forming a complete development situation that focused on both theoretical and applied studies. At present, research activities exhibit all-round progress and applied research papers are published in huge numbers. Moreover, in-depth theoretical studies have gradually developed into a rich and complete theoretical system.

Overall, the entire process and prospects of Bradford's law development have been adapted to the development and evolution law of the entire scientific literature system. The objective requirements of science and society are the driving force of the theory.

4.4.2 Vickery's Inference on Bradford's Law

In early 1948, Vickery carefully studied the literature distribution law and then promoted and amended Bradford's law, thereby making it more precise and general. Vickery's work contributed significantly to the improvement and growth of Bradford's law.

- (1) Two inferences of Vickery. In his study, Vickery did not only identify inconsistencies of the reasoning process, but also presented new and unique insights. He was the first one to extend the law to more general cases. He rededuced the results of image descriptions in accordance with linear distribution, namely, Formula 4.4, which was inconsistent with Formula 4.2. Then, Vickery discontinued Formula 4.4. After conducting logical deduction using Formula 4.2 (the zone expression), he noted that Bradford's distribution remained a curve rather than a straight line, excluding the lower bending portion. He also believed that Bradford's law was composed of language and image descriptions, with the former describing its theory and the latter demonstrating observational data and tracks.

After his comprehensive study of Bradford's law, Vickery determined that this law should not be limited to three zones because it could also be applied to cases with more zones. For example, nothing can be done with the number of selected areas. However, the scaling factor should be changed correspondingly with different partitions. Vickery proved that ranking journals could be divided into any number of zones. He derived an inference similar to the image expression, i.e., Formula 4.4:

$$n_1 : n_{1-2} : n_{1-3} \dots : n_{1-m} = 1 : V : V^2 : \dots V^{m-1}, \quad (4.5)$$

where $n_1, \dots, n_k (k = 2, 3, \dots, m)$ denotes the cumulative number of journals from the 1st to the k th zone, m indicates the divided areas, and V represents the scattering coefficient (or Vickery's coefficient).

In general, Formula 4.5 is called Vickery's formula.

Meanwhile, Vickery identified that the following formula should be workable according to Bradford's law:

$$n_1 : n_2 : n_3 : \dots : n_m = 1 : a : a^2 : \dots : a^{m-1}. \quad (4.6)$$

Formula 4.6, an extended form of Formula 4.2, is called Bradford's formula.

The scattering coefficient a can be determined in this manner. Let the total number of ranking periodicals be N ,

$$\begin{aligned} N &= n_1 + n_2 + \dots + n_m \\ &= n_1(a^{m-1} + a^{m-2} + \dots + a + 1) \\ &= n_1(a - 1)/a - 1. \end{aligned}$$

It can be derived from the preceding equation as follows:

$$N(a - 1) = n_1(a^m - 1). \quad (4.7)$$

From Formula 4.7, a conclusion can be drawn that for a given set of journals, if N is a certain value and m is determined using n_1 , then we can easily obtain a . Moreover, the formula does not only enable the determination of a , but also shows that the divided areas can be increased in an appropriate manner.

Formulas 4.5 and 4.6, developed by Vickery, are important inferences of Bradford's law, which are commonly referred to as Vickery's inference (or Vickery's modification). Formula 4.6 is an approximate form of Vickery's formula (or Vickery's approximate formula).

- (2) Proof of Vickery's formula. The formula is proven by starting from the graphics with similar triangles and other geometric data. We can also provide further proof of Vickery's formula, i.e., Formula (4.5) using Смольков's unified equation, which is from the perspective of physical process of literature scattering.

Смольков's unified equation is given as follows: $R(n) = K \lg(q_1 n + q_2 e^{-\beta n})$. If n is sufficiently large, then we can obtain an approximate result: $r(n) = K \lg q_1 n$.

The curve coincides with the straight line after Point c with increasing n , as follows:

$$\lim_{n \rightarrow N} [R(n) - r(n)] = 0.$$

Even at Point c, the difference between $R(n)$ and $r(n)$ is insignificant. In general, the relative error of this difference is less than 10% according to Soviet scholars. When the curvature is smaller, the difference is also smaller. Therefore, the approximate formula can be used to replace the unified equation. Assuming four divided areas, we can obtain the following equation according to the principles of the partitions of the ranking journal:

$$\begin{aligned} \frac{1}{4}R(n) &= R(n_{1-4}) - R(n_{1-3}) \\ &\approx r(n_{1-4}) - r(n_{1-3}) = K \lg n_{1-4}/n_{1-3}; \end{aligned}$$

$$\begin{aligned} \frac{1}{4}R(n) &= R(n_{1-3}) - R(n_{1-2}) \\ &\approx r(n_{1-3}) - r(n_{1-2}) = K \lg n_{1-3}/n_{1-2}; \end{aligned}$$

$$\begin{aligned} \frac{1}{4}R(n) &= R(n_{1-2}) - R(n_1) \\ &\approx r(n_{1-2}) - r(n_1) = K \lg n_{1-2}/n_1; \end{aligned}$$

From the three preceding formulas, a conclusion can be drawn that $n_{1-4}/n_{1-3} = n_{1-3}/n_{1-2} = n_{1-2}/n_1$.

Let the ratio be v . Then, $n_1 : n_{1-2} : n_{1-3} : n_{1-4} = 1 : v : v^2 : v^3$,

which is Vickery's formula. We can also provide the same proof in the case of m regions.

Thus, we further theoretically validate the correctness of Vickery's inference, which further illustrates the correctness of the graph description method.

(1) Comparison between Vickery's formula and Bradford's formula. As already noted, although Vickery's formula (4.5) and Bradford's formula (4.6) cannot coexist mathematically, they can be used as approximated formulas for research on the scattering law of papers. The actual distribution of journals and literature is more subject to Formula 4.5, and Formula 4.6 should be the approximation of Formula 4.5. The following discussion is about the theoretical error between the two formulas, as well as the conditions for their convergence.

The following can be derived according to Bradford's formula:

$$\begin{aligned}\frac{n_2}{n_1} &= v - 1; \\ \frac{n_3}{n_1} &= v(v - 1); \\ &\dots \\ \frac{n_m}{n_1} &= v^{m-2}(v - 1)\end{aligned}$$

Then, we obtain

$$\begin{aligned}n_1 : n_2 &= 1 : (v - 1), \\ n_2 : n_3 : \dots : n_3 &= (v - 1) : v(v - 1) : \dots : v^{m-2}(v - 1) \\ &= 1 : v : v^2 : \dots : v^{m-2}.\end{aligned}$$

Evidently, $n_2 : n_3 = n_3 : n_4 = \dots = n_{m-1} : n_m \neq n_1 : n_2$.

In general, Formulas 4.5 and 4.6 are equivalent with the scattering coefficient $a = v$ when located in the district outside the core area. However, the situation will differ when the core zone is included. In fact, the core area, such as an exceptional case, makes the regional description method and the image description method incompatible.

If $v \gg 1$, i.e., when Vickery's coefficient is sufficiently large, then we consider $v - 1 \approx v$. Evidently,

$$n_2 : n_3 = n_3 : n_4 = \dots = n_{m-1} : n_m = n_1 : n_2; \text{ therefore, } n_1 : n_2 : n_3 : \dots : n_m = a : v : v^2 : \dots : v^{m-1} \approx 1 : a : a^2 : \dots : a^{m-1}.$$

When v is sufficiently large, Vickery's formula and Bradford's formula are consistent, and they share the same scattering coefficient. A sufficiently large v indicates that relevant papers are highly concentrated in the core area of only a few

journals. The average rates of published articles also significantly differ, and the number of journals in one district is considerably more than that in another area. In this case, the discipline factor plays a decisive role whereas divided areas only have a certain effect.

The use of the Bradford's formula (Formula 4.6) instead of Formula 4.5 will cause several errors during calculation because the scattering coefficients are small for various professional disciplines, and thus, the convergence conditions of the two is impossible to fulfill with several distribution values v that are less than 10. The absolute and relative errors of the two formulas are determined using the following formulas.

The absolute error for the periodical number in the i th area is as follows:

$$\Delta n_i = n_1 v^{i-2}. \quad (4.8)$$

Its relative error is as follows:

$$\delta = \Delta n_i / n_i = 1 / (v - 1). \quad (4.9)$$

For example, given a set of ranking journals, $v = 6$ and $n_1 = 10$. Then, we can resolve the absolute error caused by the regional method with respect to the image method according to Formula 4.8.

$\Delta n_2 = 10$, $\Delta n_3 = 60$, ..., etc. The relative error is $\delta = 20\%$ based on Formula 4.9.

However, the actual error is smaller than the preceding calculation error.

4.4.3 Leimkuhler's Contribution to Bradford's Law

In 1967, F.F. Leimkuhler conducted exploratory research on Bradford's law. He started from the text description of Bradford's law and summarized the empirical data using statistical theoretical methods. He deduced the standardized formula for paper distribution in hierarchically arranged periodicals, thereby developing a regional description approach and making significant contributions to the theoretical development of Bradford's law.

For example, a group of journals is given. The journals are ranked in descending order according to their published papers in a certain subject and divided into m areas, with each area containing the same number of relevant papers. Let the ratio of the number of journals in each district be d_1, d_2, \dots, d_m .

Among which, d_i represents the percentage of the number of journals in the i area, thereby accounting for the total m areas. Hence, we have

$$d_i = a_m d_{i-1} - a_m^{i-1} d_1, \quad i = 1, 2, \dots; \quad (4.10)$$

$$m = 2, 3, \dots; \quad 0 < d_i < 1, a_m > 1;$$

where a_m refers to Bradford's constant when dividing into m areas. Evidently, the sum of d_i is 1, and we obtain the following equation according to Formula 4.10:

$$1 = d_1 + d_2 + \dots + d_m = d_1(1 + a_m + a_m^2 + \dots + a_m^{m-1}).$$

The sum of the geometric progressions enclosed in parentheses in the preceding equation is $\frac{a_m^m - 1}{a_m - 1}$.

$$\text{Therefore, } 1 = d_1 \frac{a_m^m - 1}{a_m - 1}, \quad d_1 = \frac{a_m - 1}{a_m^m - 1}.$$

When these values are substituted into Formula 4.10:

$$d_i = \alpha_m^{i-1} d_1 = \alpha_m^{i-1} (\alpha_m - 1) / (\alpha_m^m - 1). \quad (4.11)$$

From Formula 4.11, the cumulative proportion of the former j areas from the m areas, $D_{j,m}$, is as follows:

$$\begin{aligned} D_{i,m} &= \sum_{i=1}^j d_i = \sum_{i=1}^j \frac{a_m^{i-1}(\alpha_m - 1)}{\alpha_m^m - 1} = \frac{\alpha_m - 1}{\alpha_m^m - 1} \sum_{i=1}^j a_m^{i-1} \\ &= \frac{\alpha_m - 1}{\alpha_m^m - 1} \cdot \frac{a_m^j - 1}{a_m - 1} = \frac{a_m^j - 1}{\alpha_m^m - 1}, \quad j = 1, 2, \dots, m \end{aligned} \quad (4.12)$$

The m areas share the same number of papers, and thus, the contained papers in the first j areas is equal to all the papers multiplied by j/m .

Leimkuhler proved the relationship that $\alpha_m = a_2^{2/m}$ under the situation that m belonged to even and odd cases, respectively. With the introduction of coefficient b , let $b = a_2^2$ and $\alpha_m = a_2^{2/m} = b^{1/m}$. Formula 4.12 can be rewritten in a more general form. Therefore,

$$D_{j,m} = \frac{b^{j/m} - 1}{b - 1}, \quad j = 1, 2, \dots, m. \quad (4.13)$$

In Formula 4.13, a simple exponential function of j/m and a parameter b should be involved in the required minimum ratio of the number of journals. Formula 4.13 is frequently used to define the minimum number of journals that contained a specified number of relevant papers among all the periodicals. Leimkuhler obtained the distribution rate of papers, i.e., Bradford's distribution by promoting the formula to the actual value of all the variables and obtaining the inverse function.

Let x be the ratio of journals that owns the most number of papers to all the journals, where $0 \leq X \leq 1$. $F(x)$ is the ratio of the papers contained in the x part to all the journal papers. Then, from Formula 4.13, we derive the following:

$XF = \frac{b^F - 1}{b - 1}$, $0 \leq x \leq 1$, $b > 1$, where $F = j/m$. As the inverse function of X_F , the preceding formula can also be written as $(b - 1)X = b^F - 1$. After transposition of the logarithm, we obtain $F \ln b = \ln(1 + bx - x)$; thus, $F(x) = \frac{\ln(1 + bx - x)}{\ln b}$. Given that $\beta = b - 1$, the equation can be simplified into:

$$F(x) = \frac{\ln(1 + \beta x)}{\ln(1 + \beta)}. \quad (4.14)$$

$F(x)$ is Bradford's cumulative distribution function. The probability density of the continuous function can be calculated as

$$f(x) = F'(x) = \frac{\beta}{(1 + \beta x) \ln(1 + \beta)}. \quad (4.15)$$

From Formulas 4.14 and 4.15, Bradford's law has extended to cases with any number of areas. This extension is only theoretical. The numbers of papers and journals in each district tend to be extremely small when partitioning. Consequently, the error increases. Therefore, partitions cannot be excessive. However, we can make the application of Bradford's law more convenient by cancelling the restriction of dividing a paper into equal areas. A proportion of the number of papers can be arbitrarily fixed and the number of journals required in Formula 4.14 can be obtained. For example, to obtain 80% of the paper, the determined $F(x)$ is the required minimum number of journals as long as $X = 0.8$ and the corresponding β values are substituted into Formula 4.14.

4.4.4 Brookes' Description of Bradford's Law

The famous British information scientist B.C. Brookes asserted that the guiding ideology was highly evident when Bradford found the scattering law of scientific papers. However, he failed to describe it using mathematical formulas because of a major omission. Consequently, scholars only understood the importance of this ideology 20 years later. Brookes seized this crucial issue, and for the first time, described Bradford's empirical law using a mathematical formula. He then developed an image description method to complete the aftermath of this important law, which had been generally appreciated by the intelligence community. Many scholars agreed that Brookes established the mathematical formula that was exactly in line with Bradford's law and described the distribution of the law systematically and completely.

In 1968, Brookes first deduced the formula for Bradford's law as $R(n) = K \lg n$. Then, he introduced a parameter S to correct the formula by considering the uneven changes of journal order number n , as well as the number of papers. He creatively presented the following two-part mathematical expressions to describe Bradford's law:

$$R(n) = \alpha n^\beta, \quad 1 \leq n \leq c \quad (4.16a)$$

and

$$R(n) = K \lg n/S, \quad c \leq n \leq N. \quad (4.16b)$$

These two equations represent the curved and straight portions of the image, respectively. The drooping curve section that follows the straight part (Fig. 4.3) is only approximately satisfied using Formula 4.16b.

In the formula,

- $R(n)$ cumulative number of related papers of n ;
- n journal rating serial number (level);
- α number of related papers in the first level $R(1)$, which is related to the number of articles in journals with the highest paper rate;
- C number of journals in the core area, namely, the n value of the intersection, where the curve changes into smooth linear;
- N total number of ranking journals;
- β parameter related to the journal number of the core area, the value is equal to the curvature of the curved portion with a total of less than 1;
- K parameter that is equal to the slope of the linear portion in the scattering curve, it can be calculated using an empirical method. When N is sufficiently large, $K = N$;
- S parameter whose value is equal to the n point value, where the straight portion of the line intersects with the horizontal axis (Fig. 4.3)

Brookes conducted further study on Formula 4.16b. He assumed that when counting from the lowest rate of the journal papers, the final increment should be less than 1 because of the limited number of journals in each discipline. For the calculated journal, at least one related paper should be present. If N is considered the ordinal number of the last ranking journal with only one related paper published in a certain year, then N is sufficiently large. The incremental expectations of $r(n)$ from $R(N - 1)$ to $R(N)$ is 1, i.e.,

$$\begin{aligned} r(n) &= R(N) - R(N - 1) \\ &= K \lg N/S - K \lg (N - 1)/S \\ &= K \lg N - K \lg (N - 1) \\ &= -k \lg (1 - \frac{1}{N}) \approx \frac{K}{N} = 1, \end{aligned}$$

where $K = N$, which indicates that the value of the line slope is equal to the total number of periodicals. Thus, Formula 4.16b can be rewritten as: $R(n) = N \lg n S$, $c \leq n \leq N$.

Therefore, the slope of the line can be marked from the expected total papers. However, the horizontal axis, on which people draw Bradford's distribution pattern, is a logarithmic coordinate, and the density of the vertical axis is considerably higher. Thus, we cannot consider $N = 1$ simply because the slope of a graph is 1; instead, we need to observe the scale of $R(i)$.

Brookes' mathematical formula for the description of Bradford's law is also called the graphic presentation of Bradford's law. The proposition of this formula has not only improved the law theoretically, but has also paved the way for its practical use, thereby significantly promoting its application to library and information and documentation work. Thus, Brookes contributed significantly to the improvement and theoretical development of Bradford's law.

Brookes et al. found that the S value increased with the expansion of the ranges of disciplines, which was also relevant to the stage of discipline development. Thus, S can be used as a reference during the comparison of discipline ranges and their maturity levels. Meanwhile, the value of C is associated with that of S .

4.4.5 Unified Equation of Смольков

In 1977, И.А. Смольков, an information scientist from the former Soviet Union, proposed a unified equation instead of the two formulas of Brookes to determine the scattering law of journal articles.

Among ranked journals, those that are far from the core journals publish less relevant papers. Hence, the dispersion process in ranked journals can be regarded as a physical attenuation process. During this process, the increment ($\Delta R(n)$) of related papers gradually decreases with the increase in journal rank (n). Therefore, the following is assumed:

$$\Delta R(n) = \varphi(n) \cdot \Delta n, \quad (4.17)$$

where $\varphi(n)$ is a function with the journal number (i.e., rank) n as the variable. If n is sufficiently large, then $\varphi(n)$ is close to a constant. Δn is the increment of a journal (i.e., the increment of rank).

When n is small, the intensive effect of the literature is similar to the "Matthew effect," which begins working. The physical process of the intensive effect is that relevant papers exclude papers from other subjects. As n increases, this function is weakened. Therefore, we establish a function that reflects this intensive effect and label it $\varphi(n)$: $\varphi(n) \propto c_1 - c_2 e^{-\beta n}$, where C_1, C_2 are constants, and β is the attenuation coefficient of the intensive effect.

Evidently, this function has the following features. When $n = 1$, the function has the minimum value, i.e., its attenuation effect is not apparent. When n increases, the result of $\varphi(n)$ gradually increases to a constant value C_1 , i.e., the attenuation effect gradually increases to a certain value. Meanwhile, considering that $\Delta R(n)$ decreases with the increase in $R(n)$ and the decrease is a decay process, $\Delta R(n)$ is proportionate to the exponential function $e^{-rR(n)}$ and r is a dispersed attenuation coefficient of literature.

From the preceding discussion, if we ignore the small variable that is higher in order than Δn , then the differential equation can be obtained as follows:

$$dR(n)/dn = Ae^{-rR(n)}(c_1 - c_2e^{-\beta n}), \quad (4.18)$$

where A is a ratio.

When this equation is integrated and logged, we derived

$$R(n) = K_1 \ln(q_1 n + q_2 e^{-\beta n} + q_3), \quad (4.19)$$

where $K_1 = 1/r$, $q_1 = rAc_1$, $q_2 = rAc_2/\beta$, and $q_3 = rAc_3$, in which c_3 is an integration constant.

Formula (4.19) can be transformed into a common logarithm as follows:

$$R(n) = K \lg(q_1 n + q_2 e^{-\beta n} + q_3), \quad (4.20)$$

where $K = K_1/M$ ($M = \lg e = 0.43429$).

The calculation result shows that the value of q_3 is extremely small, i.e., accounting for only 0.1% of the sum of the three items in the parentheses in Formula 4.20. Therefore, the constant can be ignored entirely. Accordingly, we obtain the following equation:

$$R(n) = K \lg(q_1 n + q_2 e^{-\beta n}), \quad (4.21)$$

which is the unified equation proposed by Смольков to describe the scattering of papers.

Studies have shown that the distribution curve of papers in some disciplines obtained from the preceding equation can fit well with the actual data, thereby further proving the correctness Vickery's formula.

4.4.6 Theory and Development Trend of Bradford's Law

Journals are among the main carriers of information. The community of information science attaches considerable importance to exploring the distribution laws of scientific papers in journals. After Bradford established his famous Bradford's dispersion law, scholars had presented a dozen empirical distribution formulas. Several scholars also used the theory of mathematical statistics to perform statistical interpretation from different aspects, thereby forming a rich theoretical system known as Bradford's distribution theory.

(1) Basic content of Bradford's dispersion law

The basic content of Bradford's distribution theory is to probe into the distribution of scientific papers in journals and its application. It mainly encompasses a law, nearly a dozen empirical formulas, and four mathematical models.

Bradford's scattering law is the core content of Bradford's distribution law. It has two key points. The first point is rank arrangement, which forms the ordered directory of journals. The second point is to determine the distribution of relevant papers in the main source. The specific research methods include regional analysis and image analysis. Although the specific values of these two methods are unequal, they exhibit the distribution law of papers in journal clusters. Bradford's law is the most basic law of Bradford's distribution theory and is the necessary foundation for the existence of this theory.

Thereafter, scholars further sought to determine relations between relevant papers and the number of journals, thereby leading to the emergence of the empirical distribution formula and a dozen empirical formulas to describe Bradford's distribution. Among which, the more well-known are the linear distribution formula and the mathematical expression proposed by Brooks that consists of two parts, the general formula for Bradford-Zipf's law proposed by Kendall, the standardized formula proposed by Lyme Cooley, and Krakow's unified equation. In addition, several scholars believe that certain statistical laws can be explained by virtue of the methods from probability theory and mathematical statistics, although the generation and distribution of literature are random. At present, numerous models are available for Bradford's distribution. The four main representatives of the law are Simon's stochastic model, Brookes' mixed Poisson distribution model, Lannan's order flow model, and Alamaisiku's scientific potential diffusion model.

Bradford's distribution theory has evident characteristics, i.e., the vast majority of concrete objects under observation focus on a few main sources, their behavior regularity is affected by man-made controlling factors, and the scores of empirical distribution formulas are based on the frequency rank of occurrence of specific objects in the main source.

(2) Development trend of Bradford's dispersion law

From the current situation, research mainly focuses on two aspects. First, scholars are preparing statistics to verify Bradford's law and search for applications. In this regard, many scholars believe that laws conform to statistical results, and therefore, laws are widely recognized and developed. Second, researchers are searching for a universal and precise theoretical explanation of the empirical distribution formula, which has achieved considerable progress. However, problems, such as redundant formulas, varying arguments, lack of a unanimous conclusion, and the absence of a coherent combination of theory and practice, also exist. These issues are the manifestations of an immature theory, which requires further addressing.

The distribution of scientific papers is subject to many factors and objective conditions, including subjectivity and ambiguity. Such mathematical expressions of man-made quantitative laws are daunting and complicated. If Bradford's distribution theory makes a breakthrough in mathematical performance, then it will be perfect and universal. Therefore, the future research trend will focus on using several tools of probability theory, stochastic process theory, and fuzzy mathematics, in addition to considering the comprehensive effects of various factors and combining the actual work of literature, to search for more precise and generalized distribution formula and mathematical model. The current main research directions and problems that should be solved are as follows.

- ① Statistical data and rigorous mathematical methods should be used to strictly test Bradford's law. The pros and cons of each formula should be compared to establish and seek for a more precise and standardized mathematical model at the earliest possible time.
- ② A comprehensive research on the mechanism of Bradford's distribution should be conducted to search for a scientific and uniform theory interpretation.
- ③ Applicable conditions and limitations should be analyzed and studied, as well as combined with real-life practices, to enhance applied research. The theory should be used to guide literature information work, thereby saving money and time, as well as effectively improving the efficiency of library and information services.

4.5 Applications of Bradford's Law

The development of modern science, particularly with scientific knowledge, is highly differentiated and integrated, thereby making the distribution of scientific literature complicated. This trend significantly influences science and literature information work. Therefore, quantitatively investigating Bradford's law and its distribution theories has important theoretical value and practical significance. On the one hand, the study of Bradford's law and distribution theory can further determine the inherent law of literature and information flow as well as provide a reference for the theory establishment of information science and new mathematical

models to promote the development of informetric theory. However, similar to other laws in informetrics, Bradford's law can function as a basic theory for the scientific management of works in library and information science. From the second half of the 1960s, the information science academia has reinforced application studies of Bradford's law, and the scope of this law has expanded to a wide range of disciplines. Many social phenomena and objects are in line with Bradford's distribution law, which is regarded as one of the universal laws of human society. Therefore, Bradford's law and its distribution theory do not only play a pivotal role in informetrics, but also have a considerable influence on other related fields. This distribution theory, which originates from the field of scientific literature and reflects the common phenomenon of man-made factors that play decisive roles, has extensive application prospects.

4.5.1 Basic Method for the Application of Bradford's Law

At present, the application of Bradford's law focuses on the use of ranking technology and analysis methods. Thus, although this law was created based on the scattering of scientific papers in journals, it can also derive different applications. In addition, these methods are models for a variety of applications. These applications have three uniform steps: ① selecting the statistical tools and obtaining the original data, ② ranking the statistical data, ③ analyzing the statistical materials and obtaining the result of the statistical analysis. The selection of statistical tools is contingent on the object and purpose of different studies, and ranks are based on the number of published papers of journals, which is a key step to the entire study. The generation of Bradford's data and the determination of Bradford's constant value are dependent on rank, which is the same for different applications with varying data. This main characteristic is common to the applications of Bradford's law. However, listing all five columns just as Bradford has done is unnecessary when we use rank arrangement to generate Bradford-type data, although determining which method to use should be considered.

Two analytical methods are currently available: zone analysis and graphical analysis (analytical analysis is also included), both of which are gradually stereotyped by stimulating the basic method of Bradford. At present, these methods appear to have become standardized.

- (1) Zone analysis divides journals into three zones based on Bradford's law, thereby making the numbers of papers in each zone approximately equal. Let the ratio of the number of papers in each zone be $1 : a : a^2$. However, it is determined using Vickery's method, with not only more than three zones, but also the value of a difference according to specific conditions. Simultaneously, the zones can be divided finely until only a related paper has been published in each journal in the last zone. Therefore, the analysis results will be more precise than the results obtained by simply dividing journals into three zones. For

example, American scholar Worthen used zone analysis to examine the condition of monographs on several subjects of medical science. He mastered the basic distributions of the monographs of these disciplines and finally determined their core presses.

- (2) Graphical analysis plots the ranked statistical data according to the method proposed by Bradford, and then analyzes the curve. When conducting regional analysis, Bradford-type data only require two columns (A and B), and then we can obtain Bradford's constant. When conducting graphical analysis, Bradford-type data must have two sets of accumulated data, namely, n , whose logarithm is the abscissa, and $R(n)$, which is the ordinate. The image is merely the corresponding curve of n and $R(n)$. This method has been documented by Brookes and is widely accepted for graphical analysis. Canadian scholar Popper identified the applications of core information science journals through image analysis, which was a typical representative of this approach.

4.5.2 Main Region of the Application of Bradford's Law

The application scope of Bradford's law is extensive and has played a guiding role in determining core journals, formulating procurement strategies and store policies, optimizing collections, testing works, learning about the tendency of readers, and retrieving and utilizing literature. Meanwhile, in literature information work, we should not only focus on how to gather the most valuable documents, but also fully consider actual economic benefits. The significance of Bradford's law in this respect is that we can provide the quantitative bases for an information department with limited funds but highly valuable information to help formulate scientific policies.

- (1) Determination of core journals. One of the most basic and common applications of Bradford's law is selecting core journals. This application can be modeled directly from Bradford's method. In recent years, this application has been widely applied to various disciplines, such as chemistry, medicine, agriculture, oceanography, and information science. We can determine the core journals for a certain subject using either regional analysis or graphical analysis.
- (2) Literature search. Through the use of the mathematical formulas of Bradford's law, we cannot only estimate the total number of papers in n journals, but also evaluate the efficiency of document retrieval through calculation.
- (3) Investigation of monograph distribution. We can grasp the basic distribution of monographs in a certain field and determine the core presses in this field by analyzing the publication conditions of the monographs.

The research objectives are presses and monographs, and thus, we should select an apropos catalog that can show the publication condition of a subject. Worthen found that after analyzing the monographs and publishers of 5 titles in "The latest

Table 4.4 Bradford's distribution of monographs per publisher

Number of publishers	Number of monographs per publisher		
1	41		
1	20		
1	19		
1	17		
1	16		
1	15		
1	14		
3	11		
3	9		
2	8		
7	7		
9	6		
8	5		
11	4		
15	3		
44	2		
216	1		
Title	Number of publishers	Number of monographs	
Diseases	179	255	
Cardiac disease	154	220	
Cerebrovascular disease	86	127	
Vasculature disease	87	105	
Arrhythmia	51	63	
	557	770	
Region	Number of monographs	Publisher	Bradford's constant
		1	
1	113	5	
2	113	10	2
3	111	17	1.7
4	111	28	1.6
5	106	50	1.7
6	96	96	1.9
7	120	120	1.3
			Average 1.7

catalog of American NLM," the distribution of books against publishers substantially conforms to Bradford's distribution (Table 4.4). In the first zone, the 5 publishers published 113 books, which is more than half of the total number of publishers that only publishes a monograph (216). In the second zone, 113 monographs are published by 10 publishers. The scaling coefficient of the number of publishers in each zone tends to be a constant.

This distribution can evidently be used to guide library procurement; thus, buyers can determine where a large number of monographs are published to secure a detailed list of publishers. From actual statistics, a deviation is found for the empirical estimate. For example, Saunders, a publisher of medicine with a good reputation in the United States, is located in the lower section of the second zone. In addition, the public service department is considered a peripheral one although it is located in the core zone. Furthermore, this distribution may also provide readers with a guide to retrieving documents.

- (4) Maintenance of dynamic collection. Previously, the collections and services of a library are based on experience. Such chaotic status is due to our failure to determine the minimum value of the useful collection of information sources. Taking full advantage of Bradford's law will be reasonable.

Statistical results have shown that the distributions of the circulation of journals and their readers are subject to Bradford's law. Table 4.5 shows the distributions of journal circulation and readers in the Allen Memorial Medical Library in March 1968. The circulated journals for the aforementioned month are distributed as follows: 11 journals are borrowed 113 times; 16 journals, 108 times; and 21 journals, 107 times. Thus, the number of journals with nearly the same circulations in the successive zones forms the geometric ratio: 1:1.4:1.42:...:1.47. In addition, 86 readers and 76 circulated journals are found in the first, second, third, and fourth zones. From the former, we can continually gather statistics on the minimum and successive zones monthly or quarterly, which is contingent on utility density and draw the curves of the circulation gradient to predict the circulation demand for the

Table 4.5 Bradford's distribution of journals circulation and readers

Region	Number of circulations	Number of readers	Bradford's constant	Region	Number of circulation	Number of readers	Bradford's constant
1	113	11	—	1	118	13	—
2	108	16	1.5	2	108	18	1.4
3	107	21	1.3	3	107	24	1.3
4	110	28	1.3	4	109	31	1.3
5	110	38	1.4	5	109	40	1.3
6	110	55	1.4	6	108	54	1.4
7	109	93	1.7	7	108	105	1.9
8	109	109	1.2	8	109	109	1.1
	876	371	Average 1.4		876	394	Average 1.4

following year, guide purchase, and built apropos collections. From the latter, we can first establish a minimum core of readers, and then use appropriate theme titles in a corresponding retrieval tool to identify areas in the core, in which readers are interested in, thereby finding the latest publications under these titles. Subsequently, the main core of the theme title that is exhibited by the core authors is set up. This method still applies Bradford's law to summarize the distribution of journal articles under each title to find the core and widely used journals to serve accurate collections. The collection remains an ordered status with the minimum core of circulating journals and another minimum core of the subject, in which core authors are interested in, and thus it can offer useful materials to readers.

- (5) Measurement of the integrity of search tools. In library management, testing the integrity of abstract indexing and catalog tools is a critical task. This process is also significant for readers to determine the integrity of the retrieval and evaluation of collections. Through the use of the ranking methods and mathematical expression of Bradford's law, we can determine how many information sources should be abstracted in the interests of a certain cover degree on the one hand, and evaluate the integrity of retrieval tools in a certain subject on the other hand, while providing scientific foundations for the selection and utility of these tools by comparing actual statistical data with the theoretical values obtained from Bradford's law.
- (6) Comparison of subject amplitude. On the basis of Bradford's law, we can obtain different core zones and Ss when analyzing papers and journals with different subjects. Meanwhile, the differences among subjects are determined by comparing different cores and Ss. In general, parameter S may represent the maturity degree of development and the domain extent of a subject, thereby providing an instructive reference for amplitude determination.

In addition, the number of journals appearing in two core areas can also be a criterion for the overlap degree of two disciplines. If one journal is important to two subjects, then these subjects are overlapping. Furthermore, if the important journals for the two subjects are more, then the overlapping degree of these two subjects is higher. At present, we can address these problems using Bradford's law and quantify judgement.

- (7) Guiding readers to use journals. Bradford's law is of practical significance to instructing readers to read literature, thereby enhance the usage rate of journals.
- (8) Guiding the work of journal subscription. Through Bradford's law, we can determine the core journals of a certain subject and provide a basis for journal ordering. Simultaneously, the law can be used to determine the types of journals that should be ordered and addressed by photocopying papers to guide ordering work. Moreover, Bradford's law is of guiding significance to formulate reasonable collection policies and fund allocations.

Bradford's law is a successful attempt to use a mathematical model to illustrate a certain phenomenon in the social field, which is anomalous to several laws in social

Table 4.6 Number of journals and corresponding papers in the science citation index

	Number of journals	Number of papers (%)
20	20	
100	43	
500	70	
2200	100	

sciences, such as the 20–80 principle and the Pareto effect. Garfield searched all the journals in the science citation index database and obtained the data shown in Table 4.6. Among the total of 2200 journals, 100 occupy 43% of all the relevant papers, whereas 23% of all the journals (500) occupies 70% of all the papers. This result is highly similar to the property distribution regularity discovered by Italian economist Pareto in the 19th century, i.e., the rule of “80:20.”

Therefore, several scholars have proposed extending Bradford's law and the rank arrangement skill to other social sciences in addition to literature work. For example, Brookes declared that Bradford's rank arrangement skill could be applicable to some social phenomena; he even suggested building a subfield, namely, the statistic of individuality based on Bradford's law. Further research will broaden the prospects for the application of Bradford's law and its distribution theory.

4.5.3 *Conditions and Limitations of the Application of Bradford's Law*

From the large number of application searches using Bradford's law, the successes alternate with failures because the law is initially obtained from empirical observations. Although it has been developed and improved, Bradford's law still has several disadvantages. Only if it sufficiently satisfying some of the conditions below, Bradford's law can be established, and the application of Bradford's law is frequently limited by the factors.

- (1) The subjects, domains, or projects of papers should be clearly presented.
- (2) The journal list of related subjects, domains, or projects should be sufficient.
- (3) The time span of the journals analyzed should be clearly defined to ensure consistency in literature statistics.

Chapter 5

Word Frequency Distribution of Literature Information: Zipf's Law

Dr. George Kingsye Zipf is a professor at Harvard University. He is also a famous linguist and psychologist. Zipf is knowledgeable and has performed numerous works in various fields. In 1935, Zipf verified the research achievements of his predecessors on the word frequency distribution rule using considerable statistical data and performed an in-depth and systematic theoretical research to formally establish this law. To honor his contributions, the law was named after him. Zipf's law quantitatively determines the literature vocabulary frequency distribution rule, which is one of the basic laws of literature informetrics. This law is highly significant to book information, information resource management, and science and technology management field.

5.1 Theoretical Basis of Zipf's Law: Principle of Least Effort

In April 1948, Zipf finished his monograph "Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology," which was first published a year later. The book has over 500000 words, a refined language, a clear hierarchy, and a closely bound theme. It cites a huge amount of data and facts to make the principle of least effort incisive. This monograph has a considerable influence, and several scholars call it "a monumental work, a masterpiece." The principle of least effort effectively explains the inner factors and mechanism of Zipf's law; it is also the theoretical basis of Zipf's laws.

5.1.1 Principle of Least Effort

In a related research, Zipf found that every person he encountered in his daily life was bound to have a certain degree of movement in his environment. He saw the movement as walking on some kind of road. However, the road that people walked on in their own environment was not the same as Zipf's in all of the activities. A person who is in a relatively static state will need the sustaining movement of material and energy to output, cycle, and input a series of processes that will complete metabolism. The movement of material and energy is also working in a certain manner. We can compare the body of a person to the polymerization of materials at varying speeds and different ways through the system of the body. In a macroscopic universe, the system of people and their external environment as a whole take different roads and move at various paces.

In addition, Zipf emphasized the concepts of "movement" and "road." He aimed to illustrate that regardless of which type the movement of each person belonged to, it would be restricted by simple basic laws in a certain manner, thereby making every effort to choose the least effort. In various types of sports, people act consciously or unconsciously according to basic rules. Zipf considered this as he established a common law known as the principle of least effort. "Least" has a subjective meaning. Objectively, each person has his/her concept of "least," and these concepts are not exactly the same.

To understand Zipf's principle of least effort, we use a simple example. A person going from point A to point B can take a variety of different roads; however, he/she has to choose a road. Therefore, we have to consider from the perspective of economic, security, time, and the combination of subjective conditions, such as physical, and objective conditions, such as regional environment and other factors, and then choose a path that most conforms to their own conditions and requirements, making their own force smallest. In this manner, the power consumed by the person is economical. Such a choice is based on the principle of least effort. It is a reflection of a person's desire, which is also the result of his/her efforts. In all types of choices, people consciously or unconsciously follow the basic behavior selection rules, i.e., the principle of least effort.

5.1.2 Principle of Least Effort and the Word Frequency Distribution Law

Zipf asserted that when we use language to express and communicate ideas, the principle of least effort makes us feel like we are in the opposite directions from two forces, namely, simplification and diversity of force. When talking or writing, these two kinds of forces present as that on the one hand, we want to be understood by each other; on the other hand we want to be short as much as possible. From this point of view, the speaker only uses one word to express all the concepts for the

least effort (i.e., simplification). By contrast, his/her listeners achieve the least effort (i.e., diversity of force) when each concept is expressed in a single word. Simplification and diversity of force achieve balance, and the frequency distribution of the natural language vocabulary presents a picture of the hyperbola. This effort is different from that in physics.

After the least effort principle is presented, several scholars abroad conducted research about this principle and applied it to many areas, including book intelligence work. For example, during the late 1960s, one scholar applied the law to the reasonable position of a library or information center in a city, attempting to make all the people who could use it exert minimum effort. Several scholars have applied this principle to study the optimal scheme of books arranged in a library, completely break the traditional method in a certain order, and solve how librarians can exert the least effort when retrieving the books needed by readers.

Evidently, Zipf proposed the principle of least effort" according to the famous benefit maximization principle in economics (i.e., the economic man hypothesis). Zipf used word frequency distribution based on profound thoughts, and thus, people regarded him as the main contributor to word frequency distribution, which was called Zipf's law.

5.2 Formation and Establishment of Zipf's Law

In the literature, the use and frequency of different vocabularies are exhibiting a certain regularity. To discover and present the law, several scholars have conducted explorations. These studies and results of word frequency distribution laid the necessary foundation for the formation of Zipf's law.

5.2.1 Appearance of Frequency Dictionary

As early as 1898, German linguist F.W. Kaeding wrote the first frequency dictionary in the world (German frequency dictionary or Haufigkeits Wörterbuch der Deutschen Sprache). Kaeding wrote this dictionary through a clause that contained a sample size of 1.1 million words and statistics for the occurrence frequencies of each word in the total sample. At the beginning of this century, American education expert and psychologist E.L. Thorndike wrote the Teacher's Word Book of 20000 Words and the Teacher's Word Book of 30000 Words to produce a work with numerous frequency statistics about the English vocabulary. At present, a frequency dictionary has many varieties worldwide, including ordinary frequency dictionaries and professional frequency dictionaries.

A frequency dictionary is actually a type of vocabulary. In a vocabulary, every word provides its frequency in certain lengths of the clause. The number of accumulations of frequency information about a word in a different language is

large; hence, scholars are eager to theoretically generalize this information. In a frequency dictionary, word frequency and serial number are two of the most basic quantitative indicators. They depict the statistical properties of a word in the glossary, and thus, scholars have emphatically studied the relationship between the two basic quantitative indexes in a word table to determine the frequency of word distribution.

5.2.2 *Estoup's Found*

In 1916, famous French stenographer J. Estoup discovered the word frequency distribution of a quantitative form in a long article. He was involved in the research on improving the stenograph system and observed the following rules. Suppose you have documents containing the N word (N should be large). These words are ordered according to their diminishing absolute frequency N that appears in the literature or according to natural numbers from 1 (the word with the largest absolute frequency) to L (the word with a minimum absolute frequency). The words are ordered with the serial number, then the vocabulary is prepared (Table 5.1). Estoup found that the absolute frequency product of a word and its corresponding serial number r is generally stable under constant K as follows:

$$n_r \bullet r = k. \quad (5.1)$$

5.2.3 *Condon's Formula*

In 1928, E. Condon, a physicist from the Bell Telephone Company, found regularity in the research on raising the capacity of telephone lines for communication. His findings are as follows. From the information of Dewey and Ayres on word frequency statistics, the logarithmic $\lg r$ of the serial number of a word is expressed by the x -coordinate, whereas the ordinate denotes the logarithmic $\lg n_r$ of the absolute frequency of a word (Fig. 5.1).

Condon found that the distribution relationship between $\lg r$ and $\lg n_r$ is close to a straight line AB, with an angle α between line AB and the x -coordinate.

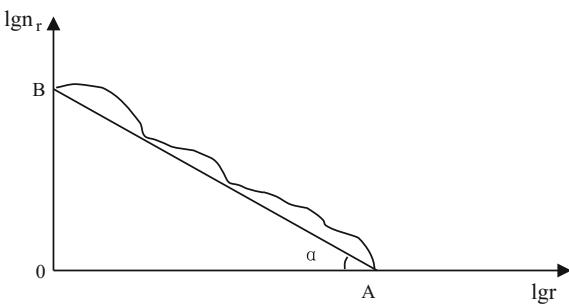
If $\operatorname{tg} \alpha = \gamma$, then

$$\lg(r^\gamma \bullet n_r) = \lg K,$$

Table 5.1 Absolute frequency of a word and its corresponding serial number

Serial number of a word	1 2 ... r ... L
Absolute frequency of a word	$n_1 n_2 \dots n_r \dots n_L$

Fig. 5.1 Word frequency distribution



where K is a constant. Thus,

$$n_r = \frac{K}{r^\gamma}.$$

After numerous trials, Condon found that when $\alpha = 45^\circ$, namely, $\gamma = \tan \alpha = \tan 45^\circ = 1$. Therefore, the preceding mathematical formula can be transformed as follows:

$$n_r = Kr^{-1}.$$

Then, the following result can be obtained from the total length of the sentences N divides the two sides of the equation:

$$\frac{n_r}{N} = \frac{K}{N} r^{-1}.$$

Let $\frac{F_r}{N} = f_r$ (relative frequency of rank r words).

Then, $\frac{K}{N} = C$ (C remains a constant); thus,

$$f_r = Cr^{-1}, \quad (5.2)$$

which is Condon's quantitative formula.

However, Condon asserted that verifying whether C was a constant would require further experiments. If C is a constant, then we can use the following method to determine the value of C .

When the test times $t \rightarrow \infty$, f_r (frequency) will become P_r (probability), and thus, the formula can be given as follows:

$$P_r = Cr^{-1}.$$

The sum of all the word probabilities is 1, i.e.,

$$\sum_{r=1}^L P_r = 1.$$

For a given sentence, vocabulary capacity $L(0 < r \leq L)$ is known When Cr^{-1} is substituted into P_r , then

$$\sum_{r=1}^L Cr^{-1} = C \sum_{r=1}^L r^{-1} = 1.$$

Thus,

$$C = \frac{1}{\sum_{r=1}^L r^{-1}}.$$

The value of C can be obtained. In Dewey's data, for example, $L = 10161$. When the preceding formula is inserted, $C = 0.102$ is obtained.

Condon successfully achieved and reported the aforementioned results. Simultaneously, he identified and hoped that people could use a wider range of experimental materials to test whether C is a constant. Accordingly, Condon presented this quantitative formula, but did not fully confirm it.

5.2.4 Zipf's Research and the Establishment of Zipf's Law

Zipf used the work of his predecessors as bases, focused on the problem that his predecessors failed to solve, and conducted bold explorations, thereby officially founding the word frequency distribution law. In 1935, Zipf conducted a systematic research on frequency distribution using a large number of statistical data. He first examined the reliability and properties of C in Condon's formula. Zipf mainly worked according to the frequency dictionary composed by M. Hanley for J. Joyce's medium-length novel "Ulysses." The sentences in the dictionary have a capacity of 260432 words and the vocabulary has 29899 words. In this manner, Zipf could inspect Condon's results on a considerably larger scale and unequivocally investigate whether C is a constant.

Furthermore, Zipf estimated the value of C according to the formula $Pr = Cr^{-1}$. He determined that in this formula, when $r = 1$,

$$P_r = Cr^{-1} = C.$$

Thus, c is the probability that the serial number of this word is 1 (i.e., the highest frequency word). From the test, Zipf obtained $C = 0.1$; hence, C is a constant.

However, numerous facts illustrate that for most European languages, nearly no word from any language can meet the following condition: the serial number is 1 and the relative frequency is 0.1 (generally, it is less than 0.1). For example, in the English language, the serial number of the word “the” is 1 and its $\text{Pr} = 0.071 < 0.1$.

The preceding statement is extremely easy to prove using mathematics.

When Condon's formula is used to solve the C value, we derive

$$C = \frac{1}{\sum_{r=1}^L r^{-1}},$$

which can be obtained approximately as

$$C \approx \frac{1}{\ln L + \beta},$$

where β is Euler's constant and $\beta \approx 0.5772$. From the preceding formula, when the values of L are 5000, 10000, 50000, and 10000, the values of C will be 0.11, 0.10, 0.09, and 0.11, respectively. These values are similar to (but slightly less than) 0.1.

In this way manner, Zipf made several modifications to his original idea and identified that C is not a constant, but instead, a parameter. Its avg is $0 < C < 0.1$. For $r = 1, 2, \dots, L$, parameter C is given as

$$\sum_{r=1}^L P_r = 1.$$

Afterward, Zipf reached a similar conclusion based on the word frequency statistics of some other brands, thereby proving that the word frequency distribution of a single parameter equation $f_r = Cr^{-1}$ or $P_r = Cr^{-1}$ is correct.

Zipf conducted numerous challenging statistical and calculation works, determined the nature of C, demonstrated the quantitative relationship between description word frequency and level of serial number, and provided a profound theoretical explanation to the principle of least effort, thereby making a significant contribution to present the distribution. Consequently, scholars use his name to refer to this law. In most of the literature of information science and linguistics, the single parameter word frequency distribution law is known as Zipf's law.

5.3 Basic Content of Zipf's Law

5.3.1 *Textual Representation of Zipf's Law*

In 1949, Zipf published his masterpiece “The Human Behavior and the Principle of Least Effort.” Under the guidance of this principle, Zipf conducted numerous

studies on the important tool of human communication and language, aiming to prove that the distribution of natural language vocabulary follows a simple law. On the basis of previous research, he collected a large number of statistical materials and performed system analyses. He found that in any article, word frequency would obey the following rules.

If we make each word's frequency in a long article (approximately 5000 words) in statistics, and arrange them according to the high frequency words in the former, the low-frequency words in descending order, and rank them with natural numbers, namely, the highest frequency word's level is 1, the frequency of the second rank is 2,..., the smallest frequency word's class is D (L). If F_r is used to indicate frequency, then r indicates serial number grade; hence,

$$F_r \bullet r = C, \quad (5.3)$$

where C is a constant. However, this constant is not absolute and will fluctuate up and down around a center value. This formula and the quantitative form that Zipf has previously validated are consistent with each other, and scholars have also called the formula Zipf's law (or first Zipf's law).

Zipf's law cannot only determine the absolute frequency of words and also the relative frequency of words.

If N is the total vocabulary (word capacity) of an article, then f_r is the relative frequency of the word of class r , i.e.,

$$f_r = cr^{-1}, \quad (5.4)$$

where c remains a constant, and $c = \frac{C}{N}$; however, $f_r = \frac{F_r}{N}$.

In general, Formula (5.3) is called the absolute frequency representation or frequency representation of Zipf's law, whereas Formula (5.4) is called the relative frequency representation or frequency representation of Zipf's law. These formulas are manifestations of different forms of the same rule.

Zipf used the principle of least effort to explain the law. He argued that for any language, the function of words used with high frequency would not be too large. The value of the word meaning itself is small in this occasion, and thus, the power to pass these words is small. Consequently, the product of word frequency and serial number level is basically stable as a constant.

5.3.2 *Image Description of Zipf's Law*

We can obtain a hyperbola, namely, Zipf's distribution curve (Fig. 5.2) according to the statistical data on word frequency and grade number available in the literature (Table 5.2), as well as by establishing a rectangular coordinate system of F_r with r and using the x -coordinate to indicate the word's level serial number r and the ordinate to indicate the corresponding frequency F_r .

Fig. 5.2 Zipf's distribution curve

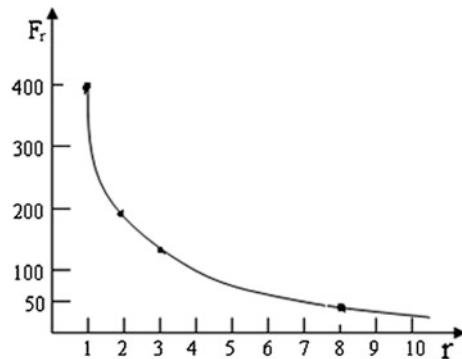
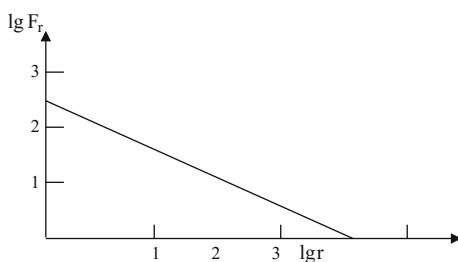


Table 5.2 Word frequency and number of statistical data in the literature

r	F _r	lgr	lgF _r
1	400	0.0000	2.6021
2	200	0.3010	2.3010
3	133	0.4771	2.1239
4	100	0.6021	2.0000
5	80	0.6990	1.9031
6	66	0.7782	1.8195
7	58	0.8451	1.7634
8	50	0.9031	1.6990
9	44	0.9542	1.6435
10	40	1.0000	1.6021
...	...		

Fig. 5.3 Zipf's distribution logarithmic curve



If the grade number r and frequency F_r use logarithmic coordinates, then the image in Fig. 5.2 will become a straight line, i.e., Zipf's distribution logarithmic curve (Fig. 5.3).

This type of distribution is known as Zipf's distribution.

5.3.3 General Mathematical Form of Zipf's Law

If the image in Fig. 5.3 is expressed as the equivalent mathematical formula, then the result is as follows:

$$\lg r + \lg F_r = \lg C.$$

In general, through analytical geometry, we can identify any straight line that slopes as b, which can be expressed as follows:

$$b \lg r + \lg F_r = \lg C.$$

To rewrite the equation in Fig. 5.3 in a similar form, we derive

$$F_r \bullet r^b = C. \quad (5.5)$$

If $b = 1$, then Figs. 5.5 and 5.3 are the same type. This result is consistent with the amendatory Zipf's law presented by Joyce. It is a general form of Zipf's law.

5.3.4 Applicability of Zipf's Law

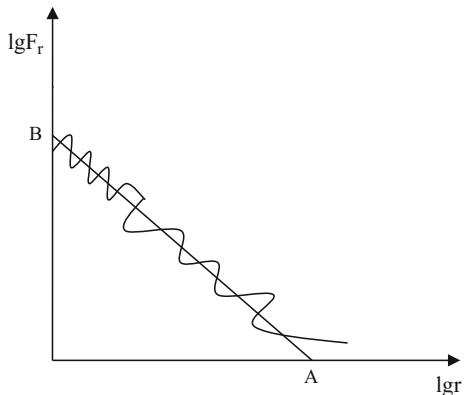
In general, Zipf's law corresponds to the actual situation of word frequency distribution in Western literature; it quantitatively presents the word frequency distribution of literature information. However, the problem of word frequency distribution is highly complex, thereby causing the scope of the aforementioned formula to have certain limitations. In particular, for words with particularly high and low frequencies, this formula cannot successfully reflect the distribution.

For example, the r value in the formula corresponds to only one F_r (or Pr) value, and thus, the word of same frequency cannot exist in a clause. This finding is inconsistent with the objective facts of language. The experiment proved that when $15 < r < 1500$, the word group capacity of the same frequency is not large, but $r > 1500$; that is, when the frequency of the word is smaller, the word group capacity of the same frequency is significantly increased. Evidently, this formula cannot properly describe the distribution of low-frequency words.

In fact, Fig. 5.4 is an incomplete straight line, i.e., a stepped broken line (Fig. 5.4). The graphics show that for low-frequency words, words with the same serial number are numerous; by contrast, for high-frequency words, words with the same serial number decrease as frequency increases. Inversely, the number of words with the same serial number increases as frequency decreases. This phenomenon cannot be described effectively using the aforementioned formula.

Evidently, Zipf's law has certain limitations in correctly reflecting the objective law of word frequency distribution. Zipf only used the general statistical method and did not use mathematical theory to perform further research on the preceding

Fig. 5.4 Word frequency distribution function of the image



results. The law is merely based on experience. Simultaneously, Zipf's law is based on English, and the succeeding research is mostly confined to the Indo-European language. A highly significant difference exists between English and Chinese, and numerous problems need to be further studied and discussed. To address these limitations, later researchers conducted thorough discussions, thereby enabling Zipf's law to achieve development.

5.4 Development of Zipf's Law

After Zipf, many scholars have performed extensive and in-depth research on Zipf's law. In summary, research on this law has focused mainly on two aspects. The first aspect involves adding parameters to correct Zipf's formula and enable it to more accurately describe the word frequency distribution of literature in a more general sense. The second aspect involves studying the experience law and theoretical basis of Zipf's distribution to extensively explore its application prospect from different angles. These studies have strongly promoted an all-round development of Zipf's law and its distribution theory.

5.4.1 Joos's Double-Parameter Formula

As early as 1936, shortly after Zipf published his research results, American linguist M. Joos presented a correction to Zipf's single parameter word frequency distribution law. Joos explained that in Zipf's formula, $P_r = Cr^{-1}$, C is a parameter, and so is the negative exponent (expressed as γ) of r . Therefore, when words in the dictionary continuously increases, γ will also increase, i.e., angle α in the image will become larger; when words in the dictionary are few, γ tends to decrease, i.e., angle α in the

image will become smaller. Evidently, γ is not always equal to 1, and angle α is not always 45° . That is, γ is not a constant, but a parameter. If parameter $\gamma = b$, then

$$P_r = Cr^{-b}. \quad (5.6)$$

Among these, $b > 0$, $C > 0$. For $r = 1, \dots, n$, parameter b , C needs to achieve $\sum_{r=1}^n P_r = 1$.

This situation describes Joos's double-parameter word frequency distribution law.

For Joos's formula, when $b = 1$, the formula will become $P_r = Cr^{-1}$, which is Zipf's single parameter word frequency distribution law. Therefore, Zipf's formula is only a special condition of Joos' formula when $b = 1$. Evidently, Joos' formula is more abstract and universal than Zipf's formula; it is also a real substantial revision of Zipf's formula.

5.4.2 Three-Parameter Formula of Mandelbrot

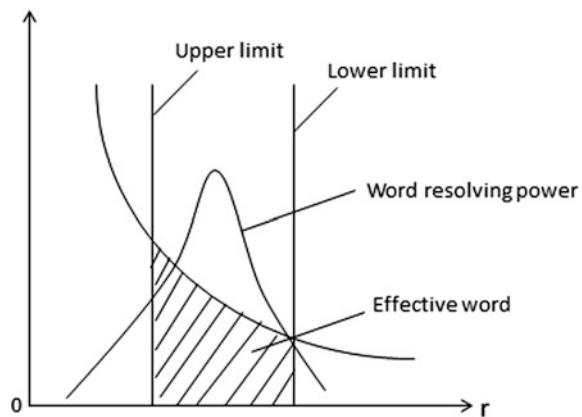
In 1952, the French-American mathematician B. Mandelbrot reinterpreted Zipf's law and fixed its expression using information theory and probability theory methods to study word frequency distribution. Mandelbrot regarded words as letter sequences separated by spaces that exhibited a certain order to compare written language and codes using analog or digital expression format, to view sentence as a word's sequence which encoded by word; and to view the article as sequence of sentences that formed by increase and elimination process of the sentence. Mandelbrot argued that all symbols had a certain value and could provide a priori probability to vocabulary to determine the minimum overall average while the amount of information remained the same. From this point of view, Mandelbrot theoretically presented the three-parameter frequency distribution law through a strict mathematical derivation. Its form is given as follows:

$$P_r = C(r + a)^{-b}. \quad (5.7)$$

Among these, $0 \leq a < 1$, $b > 0$, $C > 0$. For $r = 1, \dots, n$, parameter a , b , and C should achieve $\sum_{r=1}^n P_r = 1$.

The definitions of parameters a , b , and C are as follows.

- Parameter C is associated with the word with the highest probability.
- Parameter b is associated with the quantity of high probability words. For high probability words where $r < 50$, b is the nondecreasing function of r . Parameter b does not decrease with increasing r .

Fig. 5.5 Luhn's assumptions

Parameter a is associated with the number (n) of words. This parameter has greater freedom of choice, and thus, the formula is more flexible and suitable for measuring data under various conditions.

Mandelbrot's modified formula basically maintained the original appearance of Zipf's law, but made its application more extensive. In Fig. 5.5, when $a = 0$, the formula form is $P_r = Cr^{-b}$, which is Joos' double-parameter formula; when $a = 0$, $b = 1$ then evidently, the formula form is $P_r = Cr^{-1}$, which is Zipf's single-parameter formula. Accordingly, Joos' formula and Zipf's formula are only special cases of Mandelbrot's modified formula.

Although the aforementioned study results have improved Zipf's formula, limitations still exist, particularly in low-frequency words that cannot fully satisfy the distribution formula. Therefore, word frequency distribution still requires further research and still have to be perfected, which leads to Zipf's second law of production.

5.4.3 Low-Frequency Word Distribution: Zipf's Second Law

Some scholars argue that Zipf's law is composed of two main laws: high-frequency and low-frequency word laws. The low-frequency word distribution law is known as Zipf's second law.

Zipf's second law was first deduced by B. Booth. Let P_r be the probability that the r -th word will appear, N be the total number of different words that will appear in the overall word collection, and n be the number of times that the r -th word will appear. Hence, we derive

$$P_r = \frac{n}{N}.$$

From Zipf's law, we obtain

$$P_r = Cr^{-1}.$$

When the two preceding formulas are combined for computation, we derive

$$r = \frac{CN}{n}. \quad (5.8)$$

When the frequency is the same, the serial number r is discontinuous. Assume that the serial number is obtained using the largest sorting method, namely, all the words with the same frequency share the same serial number, with the last serial number obtained using the natural ordering method. For example, the word with a frequency of occurrence of 400 is only one, whereas that with a frequency of occurrence of 100 is three. Therefore, the serial number of the word with a frequency of occurrence of 400 is one. However, according to the natural ordering method, the three words with the same frequency of occurrence of 100 can have three order values, namely, 2, 3, 4. We make a unified provision that takes a maximum number 4 to ensure that the word frequency corresponds with each word, and understands it as the function of frequency n as the independent variable. Hence, Formula 5.8 can also be written as follows:

$$r_n = \frac{CN}{n}. \quad (5.9)$$

Evidently, from the definition of maximum sort, some r_n words will appear more than n times and r_{n+1} words will appear $n+1$ times; thus, the number of words that appears only n times is as follows:

$$I_n = r_n - r_{n+1} = \frac{CN}{n} - \frac{CN}{n+1} = \frac{CN}{n(n+1)}. \quad (5.10)$$

Consequently, the number of words that appears only one time is as follows:

$$I_1 = \frac{CN}{1(1+1)} = \frac{CN}{2}. \quad (5.11)$$

Thus,

$$I_n/I_1 = \frac{CN/[n(n+1)]}{CN/2} = \frac{2}{n(n+1)} \quad (n = 2, 3, 4, \dots). \quad (5.12)$$

Table 5.3 Booth's statistics on the word frequency distribution of four articles

Article	W•R•U•1	W•R•U•2	W•R•U•3	EIdridge
Total number (N)	4325	4409	8734	43989
Different number(D)	1001	1211	1698	6002
Frequency of occurrence is 1 (I_1)	541	710	887	2971
Frequency of occurrence is 2 (I_2)	152	227	273	1097
Frequency of occurrence is 3 (I_3)	94	91	151	516
Frequency of occurrence is 4 (I_4)	56	41	90	294
Frequency of occurrence is 5 (I_5)	36	32	62	212

Table 5.4 I1/In comparison of the predicted and actual values

I_n/I_1	Predicted values	Actual statistical values
I_2/I_1	0.33	0.36
I_3/I_1	0.17	0.17
I_4/I_1	0.10	0.10
I_5/I_1	0.071	0.070
I_6/I_1	0.048	0.051
I_7/I_1	0.036	0.035
I_8/I_1	0.028	0.028
I_9/I_1	0.022	0.029
I_{10}/I_1	0.018	0.015

The preceding equation and the length of the body are not related to the constant C and only related to word frequency. Formula 5.12 is the expression of Zipf's second law.

Booth examined the appearance of low-frequency words in four English articles. The numbers of words in the four articles are listed in Table 5.3. Table 5.4 presents a comparison of the In/I1 value that is predicted using Formula (5.12) with the numerical value of Booth's actual statistics (Article 4). As shown in the table, the predicted value is extremely close to the actual value. Thus, Zipf's second law, i.e., Formula (5.12), is more reliable and enables the prediction of low-frequency words in the English literature.

5.5 Applications of Zipf's Law

Many studies have shown that Zipf's law has common significance and wide applications. Zipf's distribution law is not only a generalized formula of informetrics, such as Bradford's law and Lotka's law, which can all be converted into Zipf's distribution form. In a wide social field, many phenomena, such as the distributions of the publication number of scientific literature, urban population, geographical features, and biological species, generally present Zipf's distribution

form or characteristics. Therefore, Zipf's distribution theory is undoubtedly a powerful tool to present the inherent law of social science. Its theoretical value is considerably beyond the scope of philology and information science; its application also infiltrates linguistics, science, economics, sociology, and the entirety of social sciences. As early as 1961, Kendall, the chairman of the British's Royal Statistical Society, spoke highly of the importance of Zipf's distribution in his speech entitled "The Law of Nature in Social science." Afterward, Haitun, a bibliometricist from the former Soviet Union, explicitly identified Zipf's law as the best scientific distribution law for solving social science phenomena.

In the field of library science, information science, information management, and science and technology management, the application of Zipf's law does not only focus on explaining the principle of least effort, but is also mainly used to present the linguistics statistical rule and language processing method. Zipf's law has certain theoretical guiding significance for determining bibliographic information characteristics, designing intelligence systems, formulating indexing principles, controlling glossaries, organizing retrieval documents, and conducting scientific evaluations. Its main applications are introduced in the following three aspects.

5.5.1 Application to Literature Indexing and Thesaurus

Zipf's law is a powerful tool for studying language problems from the perspective of statistics. Document indexing and glossary compilation are the problems that have the most dealings with language and words in library and information science works, and are the ones that most likely come in contact with Zipf's law either directly or indirectly.

(1) Vocabulary compilation

With the advent of computer information retrieval, traditional classification and subject methods can no longer adapt to information organization and retrieval. In the 1950s, the powerful function of descriptor indexing did not only change the entire face of information retrieval, but also present many new subjects to researchers. For example, the control of words, the scale of glossaries, the number of chosen words, and the criteria for choosing words are problems that should be solved. To improve the efficiency of computer information retrieval, the qualities of the descriptor list and indexing are becoming increasingly important. They directly affect the recall ratio, precision ratio, and other aspects of information retrieval. Consequently, scholars are forced to study the problems of preparing glossaries and indexing using linguistics theory and mathematical methods. This type of research involves Zipf's law.

Researchers have identified distribution characteristics by using the frequency of the cited literature and descriptor. They have determined the desirable parameter values according to Zipf's frequency distribution through indexing experiments.

The compilation of some glossaries has used the terms in the original documents, counted their frequencies, and researched their distribution characteristics. Finally, words that correspond to the frequency of use are added into the glossary. After a glossary is established, it is repeatedly amended following the practice of indexing to adapt to real specifications and applications.

Glossaries are compiled by following rules and built based on the scientific method. They can control vocabulary within an appropriate range, thereby improving their quality. Famous foreign glossaries have gradually matured since the 1960s. Most of these glossaries have been compiled in the aforementioned manner, and the compilation process has been tested using certain mathematical methods.

(2) Automatic indexing

Since the development of computers, automatic indexing and automatic classification have become popular topics in the study of intelligence experts both local and abroad. Automatic indexing involves the processing of original information using a computer after the original text is inputted into the system, statistically analyzing the frequency of each word through program control, selecting suitable words for indexing or comparing them with a specific classification system, and finally, classifying them. Word frequency plays a decisive role in the entire process. The values of high-frequency and low-frequency words are not too large for retrieval; hence, they cannot be used for indexing or for identifying the category that words should be placed into. We must also choose words with an appropriate frequency and powerful indexing and classification functions.

This study originated from P. Luhn. As early as 1958, Luhn asserted that word frequency in an article would provide a good strategy to identify effective words. The relative position of effective words in a sentence provides an excellent measurement method for the sentence effect. Therefore, evaluating the effectiveness of a sentence will be based on a combination of two methods. Luhn arranged the words in a given position in the text according to decreasing order of frequency and obtained Zipf's distribution, such as a hyperbola. He assumed that at that time, two critical points could be found on the r -axis to determine a critical range and eliminate high-frequency and low-frequency words that fell outside the range. Moderate-frequency words are effective, and thus, are retained. The resolution of effective words is their capability to recognize text content. These words reach the maximum point at the midpoint of two critical points on the r -axis and from the peak point to reduce both sides near the critical point to 0 (Fig. 5.5).

The aforementioned assumptions have laid the foundation for automatic indexing in information retrieval. Moreover, Luhn personally designed a method based on word frequency for automatic indexing. If literature can be represented using a class name, then each name will represent the words occurring in a paper for a certain class; thus, if a valid word occurs as a member in this class, then the literature is indexed by name. Such a system generally consists of three parts: (1) eliminating high-frequency words, (2) eliminating suffixes, and (3) identifying

the corresponding word stems. After such processing, we can obtain a list of class names, which can represent literature in the retrieval process and can exhibit the effects of index terms or keywords.

In literature indexing, weighting search words is a good means to improve retrieval efficiency. S. Jones designed a highly meaningful weighted method based on Zipf's law and Luhn's assumptions. Luhn assumed that a change in search word resolution would correspond to a change in their frequency level function. The word with the highest resolution is a medium-frequency word, thereby suggesting that an effective word from the literature can be found using this model. Evidently, the same database can be used to provide a weighting system for all types of special literature search words. This system will assign a weighting that is directly consistent with its frequency in the literature for each search word. The weighting method can be spread across the entire document. The vocabulary of a document typically follows Zipf's distribution; i.e., if we count the data where each search word has occurred and in how many literature. Then, these words are arranged according to decreasing order of frequency and a hyperbolic pattern is obtained. Through an experiment, Jones determined that if N articles are available, and a certain search word involves n articles, then the word for the weight of $\log(N/n) + 1$ is provided, and good retrieval effect is achieved.

5.5.2 *Application to Information Retrieval*

In computer information retrieval, a database should be built first. At present, most databases are still document types. These databases consist of one-by-one records. A record represents a piece of literature and is divided into different fields according to various characteristic description of the literature. The field can be divided into author, title, and subject fields according to its attribute, regardless of which type of field is made up of words. Therefore, Zipf's law is closely connected with information retrieval.

The organization of a document during information retrieval is connected with languages. When we establish an information retrieval system, building an inverted file (secondary index) is typical. The size of an inverted file (secondary index) depends on the number of different words and the frequency of each word in the same property field; that is, we need to consider the number of times that each word occurs in different records. For example, the size of the author's inverted file (author index) does not only depend on the number of different authors in the author field in all the records, but also on the total number of authors. In fact, regardless of inverted file type, word frequency is not completely consistent. However, we can attempt to identify the rule. On the basis of Zipf's law, we arrange each word of the description attribute values in order of diminishing frequency. Different levels are formed. Assume that a type of field has D different words in the entire database and its total occurrence is N . Hence,

$$P_r = \frac{\text{occurrence frequency of words with rank } r}{N}.$$

Therefore, the occurrence frequency of words with rank r is equal to NP_r . P_r is the probability of the attribute values of describing the word of rank r that is randomly drawn from related fields. It satisfies the relation as follows:

$$\sum_{r=1}^D P_r = 1.$$

We find that in an inverted file, the occurrence frequency of a word that will be added into the file approximately satisfies the following formula:

$$P_r = A/r. \quad (5.13)$$

A is constant in the formula. A large number of studies have shown that the value of A is close to 0.1. Therefore, this formula and Zipf's formula are equivalent, which indicates that the word frequency characteristic in the depository and Zipf's law are consistent. Through calculation, we can resolve the required memory space of the database.

5.5.3 Application to Science Evaluation

The word frequency analysis method used in Zipf's law has been increasingly applied to scientific evaluation, and currently, to the management of science and technology. Such trend is notable. For example, using the bibliometric analysis of keywords to show the research trends of a subject is a valuable empirical study. In particular, a large-scale word frequency statistics analysis based on network environment can increasingly improve the credibility of research conclusions and gain the favor of science and technology management departments.

- (1) Abroad case study: Report on nanotechnology research and development of the University of Montreal in Canada

In 1997, Robert Delphé, a professor at the University of Montreal in Canada, and his team completed a metrological analysis report on the state of nanotechnology research in the world and submitted it to the National Research Council (NRC) of Canada. Its objective is to provide quantitative information on the development of international nanotechnology to make the decision making of the country more scientific with regard to nanotechnology research and development. The report was based on 79 nanotechnology keywords confirmed by NRC. It used the word frequency analysis method to analyze the nanotechnology thesis output worldwide and the distribution of the nanotechnology patents of countries in the world. The basic

conclusion of the report indicates that nanotechnology is developing rapidly around the world during the 1990s, and the United States and Japan are leading in the field, whereas nanotechnology research in Canada is relatively backward. The report also includes Chinese nanotechnology research direction, paper production, patent application, and the main nanotechnology research institutions. It provides a general outline of the academic status of China in this research field.

The INSPEC database obtained from 79 keywords provided by NRC was retrieved. From the database, 25484 papers were determined to have 50 keywords with higher frequency in 8 years. When the 25484 papers were classified by country, the United States ranked first with a total of 7927 nanotechnology theses, which accounted for 31.1% of the nanotechnology papers worldwide. Among the 10 countries with the most papers, Japan ranked second with 3867 papers, accounting for 15.2% of the worldwide total. China was listed at seventh place with 1020 papers (4.0%). The other countries in the top 10 are Germany (2815), Great Britain (1302), Russia (1296), France (1266), Italy (620), Switzerland (522), and Canada (506). China apparently has a place in the field of international nanotechnology research. The Canada report also researched the agency distribution of 25484 papers and listed 90 research institutions with the most papers in the world. Among these, 6 institutions are from China.

The report also listed 40 national distributions with the highest frequency nanotechnology keywords. The frequency distribution of these keywords can approximately draw the outline of the research field of various countries. Table 5.5 lists the major keywords in Chinese nanotechnology papers and compares them with the major keywords in the United States, Japan, and the world. These two countries were selected for comparison because they were the main power in nanotechnology research worldwide. From Table 5.5, we can see that the top 12 Chinese keywords are basically identical with those of the United States, Japan, and the world based on frequency sorting. In particular, the top 3 Chinese keywords are also the top 3 keywords in the world. The mainstream research direction of nanotechnology in China is consistent with that of the world. However, the difference in the specific order of keywords also reflects the advantages and disadvantages of Chinese nanotechnology research.

Notably, the keyword “nanocrystal” ranks 4th in the United States, 6th in Japan, and 1st in China, with a total frequency of 303, which even surpasses that of Japan. Another keyword, “nanocrystal material,” which is related to “nanocrystal,” ranks 10th in China, and 22nd, 31st, and 21st in the United States, Japan, and the world, respectively. Therefore, China has apparent advantages in the aspects of nanocrystal and nanocrystal material research. Simultaneously, we also identify areas of weakness: “scanning tunneling microscopy” and “atomic force microscopy.” The former keyword ranks 1st in the United States, Japan, and the world, but only 3rd in China, with a frequency of only 85, accounting for 28% of the frequency of “nanocrystal,” which ranked 1st in China. The latter keyword ranked 7th in China, but 2nd and 3rd in the United States and Japan, respectively. The research direction represented by “scanning tunneling microscopy” and “atomic force microscopy” is characterization

Table 5.5 Distributions of the main nanotechnology keywords from China, the United States, Japan, and the world (1989–1996)

Keywords	China			United States			Japan			World		
	Frequency	Order	Frequency	Order	Frequency	Order	Frequency	Order	Frequency	Order	Frequency	Order
Nanocrystal	303	1	730	4	298	6	2899	2				
Fullerene	113	2	860	3	332	5	2618	3				
Scanning tunneling	85	3	1402	1	716	1	3964	1				
Nanostructure	70	4	475	5	140	12	1385	9				
Quantum wire	61	5	454	8	432	2	1833	5				
Electroluminescence	58	6	254	12	332	4	1556	8				
Atomic force microscopy	51	7	870	2	366	3	2195	4				
Quantum dot	51	8	457	7	235	7	1670	7				
Laser deposition	48	9	471	6	152	11	1746	6				
Nanocrystal material	46	10	65	22	8	31	254	21				
Conduction polymer	41	11	275	11	184	8	1029	11				
Giant magnetoresistance	37	12	295	10	167	9	751	12				

and monitoring to nanoscale. The low relative frequency and backward relative order reflect that China is relatively weak in these research aspects.

The report also performs screening analysis of patents about nanotechnology based on nanotechnology keywords using the United States patent database (1987–1996) as the basic source of data. The research results show that the United States and Japan have the most patent numbers, considerably exceeding those of other countries and involving wide-scale technology. The frequency of patent keywords is the highest, accounting for 59% (United States) and 23% (Japan) of the total frequency. The other countries with a large number of patents are Germany, France, Great Britain, and Canada. The report indicates that in the field of nanotechnology, China's maximum number of patents filed for technology is consistent with that of the world.

(2) Domestic study case: Analysis of research hot spots in international library information science from 2008 to 2012

In 2013, based on the 2011 Journal Citation Reports/Science Edition edited and published by the Institute of Scientific Information in 2011, we selected 17 types of foreign journals with high influence as data sources. We used the word frequency analysis method to conduct a statistical analysis of research hot spots in international library information science from 2008 to 2012. The abbreviations for the 17 types of journals are as follows: (1) MIS QUART, (2) J COMPUT-MEDIAT, (3) J AM MED INFORM ASSN, (4) INFORM SYST RES, (5) J INFORMATR, (6) INFORM MANAGE-AMSTER, (7) INT J COMPSUPP COLL, (8) J INF TECHNOL, (9) ANNU REV INFORM SCI, (10) J MANAGE INFORM SYST, (11) INFORM SYST J, (12) J ASSOC INF SYST, (13) SCIENTOMETRICS, (14) J HEALTH COMMUN, (15) EUR J INFORM SYST, (16) J AM SOC INF SCI TEC, and (17) J STRATEGIC INF SYST.

The following is the results of our word frequency statistics (Table 5.6).

Table 5.6 High-frequency words in international library information science for nearly five years (frequency >130)

Order	Popular words	Frequency	Order	Popular words	Frequency
1	Science	454	15	Behavior	184
2	Technology	386	16	Design	177
3	Influence	376	17	Indicators	176
4	Model	373	18	Innovation	174
5	Performance	315	19	Quality	171
6	Systems	312	20	Organizations	170
7	Information	290	21	Networks	166
8	Information-technology	266	22	Information-systems	156
9	Information technology	259	23	h-index	151
10	Communication	259	24	Care	146
11	Knowledge	252	25	European journal	143
12	Management	234	26	Trust	136
13	Internet	227	27	Collaboration	136
14	Perspective	188			

From Table 5.6, we determine that for nearly five years, the research hot spots in international library information science can be divided into six latitudes: measurement, management, technology, network, retrieval, and medical health.

The popular words for the measurement latitude include “science,” “influence,” “indicators,” “h-index,” “journals”, “bibliometrics,” “citation analysis,” “citations/citation,” “index,” “influence factor,” and “publications.” These popular words cover basic theory, indicator measurements, analysis methods, and research objects for metrology. Metrology research has become a relatively independent research field in library information science (LIS).

The popular words for the management latitude include “knowledge,” “management,” “perspective,” “innovation,” “organizations,” “collaboration,” and “adoption.” These words reflect LIS research management. Management is one of the major upper subjects of LIS, and thus, LIS and management exhibit natural affinity.

The popular words for the technology latitude include “technology,” “model/models,” “performance/behavior,” “systems/system,” “information technology/information-technology,” “design,” “information systems,” and “framework.” These words represent the dense colors of technology and the system model for LIS research in the field of LIS research. Technical topics have occupied highly important positions. The scope of the entire international LIS field generally pays high attention to technical research.

The popular words for the network latitude include “information,” “communication,” “Internet,” “networks,” “trust/user acceptance,” “e-commerce/electronic commerce,” “online,” “satisfaction,” and “World Wide Web.” These words reflect the research content of LIS under the network environment.

The popular words for the retrieval latitude include “implementation,” “retrieval/search/seeking,” “relevance,” “communication technology,” “information system research,” “content analysis,” “information-retrieval/information retrieval,” “classification,” and “context.” These words represent important contents in LIS research, i.e., information retrieval.

The popular words for the medical health latitude include “quality,” “care,” “system,” “health/health information,” “risk,” “physician order entry,” “electronic health record,” “women/breast cancer and health-care/health care,” “health literacy,” “public health/public-health,” and primary-care.” These words reflect the development momentum of interdiscipline, i.e., medical informatics. Research in this area is also one of the popular topics in the current international LIS field.

Chapter 6

Author Distribution of Literature Information: Lotka's Law

The basic laws of infometrics are Bradford's law, Zipf's law, and Lotka's law. Lotka's law explains scientific productivity and the relationship between authors and the quantities of their papers.

As an empirical law that describes the distribution of authors and papers, Lotka's law can be explained from various aspects. In the field of infometrics, Lotka's law indicates the distribution of authors during a certain period or within certain subject areas. From the perspective of anthropology, Lotka's law can be understood through the following statement, "...in the process of human development, the contributions to human progress by different individuals with different personalities..." Merton considered such differences the Matthew's effect.

Since the 90th anniversary of Lotka's law in 1926, several scientists have conducted comprehensive research on this law. The main research methods used by scholars are experience, theoretical, and simulation model methods. A conclusion can be drawn by comparing these three methods. Among them, the experience method has achieved the greatest progress, but still exhibits several disadvantages. For example, it cannot explain the difference among distribution models for authors in various disciplinary fields. The theoretical method has achieved advances in solving the mathematical relationship among different distributions of authors, whereas the model method is mainly used in scientometrics and bibliometrics.

After the 1960s, Lotka's law became a research hot spot along with the high-speed development of informatics theory. Researchers have completed numerous comprehensive studies on Lotka's distribution data collection, the determination of the range of parameter values, and the fitting and testing of equations according to the time limitation of Lotka's law and the characteristics of empirical laws, which should be constantly revised. Meanwhile, scientists from different disciplines have contributed to the development and revision of Lotka's law by applying their own academic advantages. Literature has shown that even Internet website coupling also satisfies Lotka's distribution.

6.1 Background of Lotka's Law

6.1.1 Founder of Lotka's Law: Lotka

Alfred James Lotka (1880–1949) was an American demographer. He was born in Poland in 1880 and died in Debanke, New Jersey in the United States in 1949. Lotka studied in France, Germany, and Great Britain when he was young. After he obtained his Bachelor of Science degree from Birmingham University, he went to Universität Leipzig and Kone Le University to enroll in their graduate programs. After he graduated, Lotka worked at the head office of an American university, the National Patent Office, and the National Bureau of Standards. From 1924, he worked at the Metropolitan Life Insurance Company.

Lotka engaged in social and academic activities. He once served as the subeditor of *Scientific American* and as the president of *American Statistical Association* and the *Population Association of America*. Lotka is good in statistics. With regard to science, Lotka studied the dynamic condition of general biology, developed the theory of demographic analysis, and presented the competitive growth rate. In 1925, he published *Physics and Chemistry Basis* as a research on organic sphere using mathematical and physical methods.

In 1926, Lotka, who was then working for an insurance company, published the paper entitled *The Frequency Distribution of Scientific Productivity* in *J. Washington. Acad. Sci.* In this paper, Lotka reported the inverse square relationship between scientific literature authors and their papers. Then, he promoted the classic Lotka's law, which became the earliest and most famous informetrics law. Lotka's law made a pioneering contribution to informetrics.

6.1.2 Background of Lotka's Law

In a glorious page of human history in the 20th century, scientific development did not only progress rapidly, but also experience deeper progress. In the early 20th century, Einstein proposed relativity theory while Planck proposed quantum theory, which developed physics research and influenced every aspect of modern natural science. For example, nuclear chemistry and other new branches were established as a result of the transformation of Dalton's classical chemistry theory into modern chemistry theory. These changes are also reflected on the increasing scientific literature. Meanwhile, scholars realized the productivity of science and technology in all fields after the end of World War II. Therefore, scientists have strengthened their effort to increase their achievements.

Since the beginning of the 20th century, the amount of scientific literature worldwide, particularly those from the United States, has doubled. This increase is absolutely the fastest among scientific achievements in the United States, and document information flow has reached an unprecedented peak. We can clearly see from Fig. 6.1 that from 1839 to 1899, American technical journals have gradually

increased from 150 types in 1839 to 400 types in 1899. Periodical doubling time is approximately 36.5 years ($d = 36.5$). Meanwhile, approximately 500 types of periodicals were available in 1905 and 1000 in 1925; hence, periodical doubling time was 20 years, which was shortened by nearly 16.5 years.

The growth of technical journals increases the number of abstract journals. To fit in the increase, abstract journals become normalized. This change has aroused considerable interest among scholars with regard to the study of document rules. In 1922, W. Hulme, who worked for a patent library, presented statistical bibliography, which provided scholars with a new perspective in literature research.

With the considerable increase in science and technology and technical journals, scientific and technological workers, as the main force that conducts technical research, also experience a surge in number under the aforementioned situation. On the one hand, scientists strive for academic priorities through a large number of published monographs or papers to receive the recognition of society on scientific labor and achievements, particularly the incentive of determining academic priorities. On the other hand, the verification of academic qualifications is mainly based on personal academic works in modern society. Thus, a competition occurs among scientists who published papers about their research. However, inevitable problems exist among different amounts of literature because of personal ability and limitation in objective conditions. Consequently, exploring the distribution of authors is becoming increasingly popular.

Lotka considered the quantities of authors and papers, as well as their uneven distribution in the study of data. Thus, he introduced the concept of scientific productivity, i.e., the amounts of papers written by individual researchers during a certain period. He measured the capability of scientific researchers through this concept.

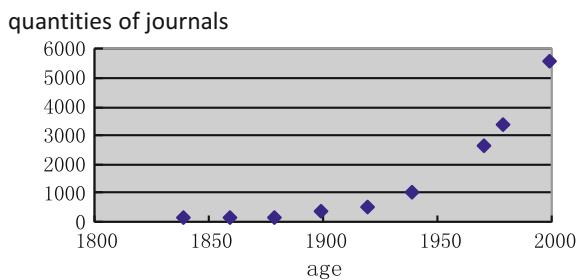
6.2 Formation and Basic Content of Lotka's Law

6.2.1 Formation of Lotka's Law

(1) Collection and conclusion of data

Lotka collected data from chemistry and physics literature as basic data. Chemistry and physics were the fastest developing and most comprehensive subjects in the 20th century, and they could fully reflect the entire process of the development of science. For chemistry, Lotka chose *Chemistry Abstract* (CA), which was published by the chemical abstract service. CA is the universal chemistry journal that includes theoretical chemistry and applied chemistry. The amount of papers under CA considered nearly 98% of all chemistry literature. For physics, Lotka analyzed the *List of Physics History*, which was compiled by Auerbach from Germany and included 1325 scientists and their works in the field of physics in the early 20th century.

Fig. 6.1 Growth of the American journal of science and technology



In Table 6.1, PN denotes the number of papers, AN denotes the number of authors, AP denotes the percentage of authors, and CP denotes the cumulative percentage.

Note: 1 AP can be obtained using the following equation: $f_0(y_x) = \frac{y_x}{\sum y_x}$.

Tip: y_x is the number of authors (AN).

$$2 \text{ CP} = \sum f_0(y_x).$$

Statistics are presented in Table 6.1. In this table, Lotka chose 6891 authors, whose names start with A or B from CA's cumulative index from 1907 to 1916, and 1325 authors from physics.

When Lotka used statistical knowledge to conclude data at that time (i.e., the 20th century), science communication and cooperation were notably minimal. Lotka neglected the role of authors in literature distribution given the small proportion of cooperation in literature. Thus, for cooperation papers, Lotka considered only the older ones without counting the others. However, with the development of technology, cooperation papers have become the main form of scientific activity. The original Lotka's law must be modified to fit in cooperation research papers.

(2) Estimating the value of n

The number of papers for an author is set to write x , and y_x denotes the number of authors who wrote x papers. To observe the early results, Lotka took the logarithm of x and y_x and obtained $\lg x$ and $\lg y_x$. Then, Fig. 6.2 is drawn according to the data. To show a good linear relationship between x and y_x , Lotka deleted the points for prolific authors. To estimate the straight slope in Fig. 6.2, we can use the least squares method. The slope n is used to express:

$$n = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2}, \quad (6.1)$$

where X and Y represents $\lg x$ and $\lg y_x$, respectively; and N denotes the number of investigated data pairs. The result is $n = 1.888 + 0.007$ when the top 30 points from CA, namely, the first 30 data in Table 6.1, are used. In physics, $n = 2.201 + 0.017$ based on the first 17 data. Therefore, Lotka identified the exponent of the equation that was approximately 2.0 based on the two examples.

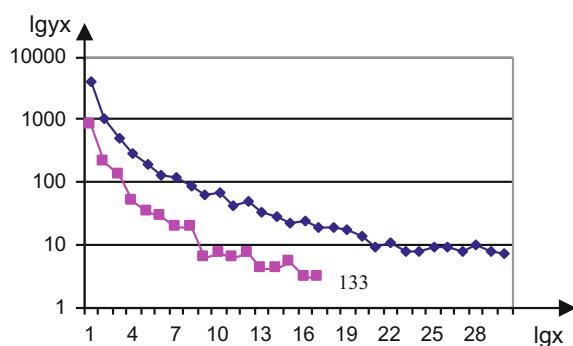
Table 6.1 Lotka's statistics

Chemistry				Physics			
PN	AN	AP	CP	PN	AN	AP	CP
1	3991	0.5791	0.5791	1	784	0.5916	0.5916
2	1059	0.1536	0.7328	2	204	0.1539	0.7456
3	493	0.0715	0.8043	3	127	0.0958	0.8415
4	287	0.0416	0.8460	4	50	0.0377	0.8792
5	184	0.0267	0.8727	5	33	0.0249	0.9041
6	131	0.0190	0.8971	6	28	0.0211	0.9252
7	113	0.0163	0.9081	7	19	0.0143	0.9396
8	85	0.0123	0.9204	8	19	0.0143	0.9539
9	64	0.0092	0.9297	9	6	0.0045	0.9584
10	65	0.0094	0.9391	10	7	0.0052	0.9627
11	41	0.0059	0.9451	11	6	0.0045	0.9684
12	47	0.0068	0.9519	12	7	0.0052	0.9735
13	32	0.0046	0.9566	13	4	0.0030	0.9766
14	28	0.0040	0.9606	14	4	0.0030	0.9796
15	21	0.0030	0.9637	15	5	0.0037	0.9833
16	24	0.0034	0.9672	16	3	0.0022	0.9855
17	18	0.0026	0.9698	17	3	0.0022	0.9878
18	19	0.0027	0.9725	18	1	0.0007	0.9886
19	17	0.0024	0.9750	21	1	0.0007	0.9894
20	14	0.0020	0.9770	22	3	0.0022	0.9916
21	9	0.0013	0.9783	24	3	0.0022	0.9939
22	11	0.0015	0.9799	25	2	0.0015	0.9954
23	8	0.0011	0.9811	26	1	0.0007	0.9962
24	8	0.0011	0.9822	27	1	0.0007	0.9968
25	9	0.0013	0.9836	30	1	0.0007	0.9975
26	9	0.0013	0.9849	31	1	0.0007	0.9983
27	8	0.0011	0.9860	34	1	0.0007	0.9997
28	10	0.0014	0.9875	37	1	0.0007	0.9997
29	8	0.0011	0.9886				
30	7	0.0010	0.9896				
31	3	0.0004	0.9901				
32	3	0.0004	0.9905				
33	6	0.0008	0.9914				
34	4	0.0005	0.9920				
40	2	0.0002	0.9936				
41	1	0.0001	0.9937				
42	2	0.0002	0.9940				
44	3	0.0004	0.9944				
45	4	0.0005	0.9950				

(continued)

Table 6.1 (continued)

Chemistry				Physics			
PN	AN	AP	CP	PN	AN	AP	CP
46	2	0.0002	0.9953				
47	3	0.0004	0.9957				
49	1	0.0001	0.9959				
50	2	0.0002	0.9962				
51	1	0.0001	0.9963				
52	2	0.0002	0.9966				
53	2	0.0002	0.9969				
54	2	0.0002	0.9972				
55	3	0.0002	0.9976				
57	1	0.0001	0.9978				
58	1	0.0001	0.9979				
61	2	0.0002	0.9981				
66	1	0.0001	0.9982				
68	2	0.0002	0.9984				
73	1	0.0001	0.9985				
74	1	0.0001	0.9986				
78	1	0.0001	0.9987				
80	1	0.0001	0.9988				
84	1	0.0001	0.9989				
95	2	0.0002	0.9991				
107	1	0.0001	0.9992				
109	1	0.0001	0.9993				
114	1	0.0001	0.9994				
115	1	0.0001	0.9995				
345	1	0.0001	0.9996				
346	1	0.0001	0.9997				

Fig. 6.2 Lotka's distribution curve note: the diamond denotes chemistry data; the square denotes physics data

The original statement of Lotka's law is as follows: the number of authors who have published x papers occupied the proportion of total authors within a certain period; the proportion is denoted as $f(x)$, which varies inversely as the square of x as follows:

$$f(x) = C/x^2. \quad (6.2)$$

Note: $f(x)$ is the proportion of authors who have published x papers to the total number of authors, x is the number of papers, and C is the characteristic constant in certain subject areas.

(3) Estimating the value of C

When both sides of Formula 6.2 are divided by the total number of authors Σy_x , we obtain

$$\frac{y_x}{\sum y_x} = \frac{c}{\sum y_x} \cdot \frac{1}{x^2}. \quad (6.3)$$

Let $f(y_x) = y_x/\Sigma y_x$, which is a fraction of the authors who published x papers. Let $C = c/\Sigma y_x$, where C is a new constant that represents the number of sampled authors, then Formula—6.3 is derived as follows:

$$f(y_x) = C \times 1/x^2. \quad (6.4)$$

We can recognize Formula 6.3 as another expression of Lotka's law. It indicates that an inverse relation exists between the proportion of authors who published x papers to the total number of authors and number of papers x .

Thus, $x^n y_x = C$,

$$\begin{aligned} y_x &= C/x^n, \\ y_1 &= C(1/1^n), \\ y_2 &= C(1/2^n), \\ y_3 &= C(1/3^n), \\ &\vdots \\ &\vdots \\ y_x &= C(1/x^n), \end{aligned}$$

To obtain the cumulative sum on both sides,

$$\sum y_x = C \cdot \left(\frac{1}{1^n} + \frac{1}{2^n} + \frac{1}{3^n} + \dots + \frac{1}{x^n} \right).$$

When Σy_x is divided on both sides,

$$\frac{\sum y_x}{\sum y_x} = \frac{c}{\sum y_x} \cdot \sum \frac{1}{x^n}$$

After simplifying, let $C = c/\sum y_x$.

$$1 = C \sum \frac{1}{x^n}, \text{ namely } C = \frac{1}{\sum \frac{1}{x^n}} \quad (6.5)$$

Given that $\sum 1/x^n$, it converges when $n > 1$ and diverges when $n \leq 1$. Therefore,

$$C = c/\pi^2 = 0.6079.$$

6.2.2 Content of Lotka's Law

(1) Verbal expression

After analyzing Table 6.1, we determined that in CA, 3991 authors published 1 paper, which accounted for 57.92% of the total authors (6891), and 1059 authors published 2 papers, which accounted for 15.73%. In physics, 284 authors published 1 paper, accounting for 59.71% of the total 1325 authors, and 204 authors published 2 papers, accounting for 15.04% of the total number of authors. Lotka inferred the conclusion that the number of authors who published 2 papers was a quarter of the number of authors who published 1 paper, and the number of authors who published 3 papers was one-ninth of the number of authors who published 1 paper, and so on. This conclusion indicates that the number of authors who published n papers is $1/n^2$ of the number of authors who published 1 paper and the number of authors who published 1 paper is 60% of the total authors. This statement can be expressed as follows:

$$y(x) = y(1)/x^2. \quad (6.6)$$

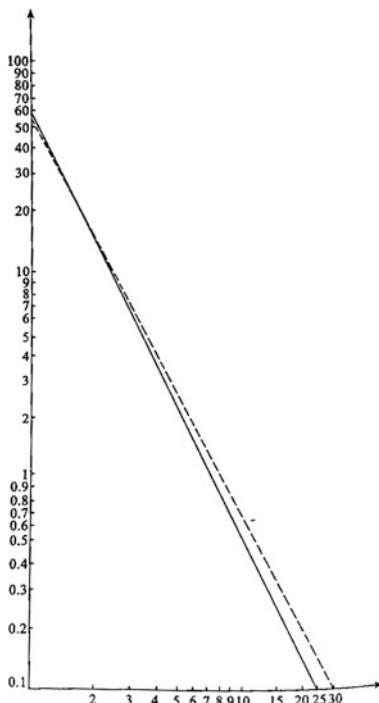
Note: $y(1)$ is the number of authors who published 1 paper; $y(x)$ is the number of authors who published x papers.

(2) Graphical representation

(3) Characteristics of Lotka's law (Fig. 6.3)

Lotka's law actually addresses the distribution phenomenon of the concentration and dispersion of scientific papers based on authors. It has two basic characteristics. First, Lotka adopted the method of frequency sorting, i.e., the ranking of authors

Fig. 6.3 Distribution curve of Lotka's law



according to their frequency of appearance instead of the number of their papers. Second, the concentration and dispersion of papers based on the authors are restricted to a square inverse relationship. This characteristic indicates that Lotka has only provided a single description for the degree of concentration and dispersion. The main objectives of later scholars are to verify and develop the second characteristics of Lotka's law.

6.2.3 Generalized Lotka's Law

In introducing Lotka's law, Formula 6.2 shows the relationship between authors and papers according to the actual statistics in the fields of chemistry and physics. However, as a statistical law of experience, Lotka's law also exhibits an obscure darker empirical feature and relative accuracy. In estimating the values of n and C , Lotka did not choose the precise value, but only a divisor. Thus, we find that the values of the time and interval of the sample data differ in various areas, and the value of n is not simply approximately 2.0. In 1976, Vlachy raised such a constructive problem. In 1985, M. L. Pao drew the conclusion that n belonged to (1.2, 3.5) of 48 groups involved in various fields of research. In accordance with this law, we can deduce the form of the inverted power of Lotka's law, i.e., the general expression is as follows:

$$f(x) = C/x^n, \quad (6.7)$$

where x is the amount of paper, $f(x)$ is the proportion of authors who published x papers to the total number of authors C , and n is a parameter. Formula 6.7 is called the generalized Lotka's law.

In the course of studying of Lotka's law, researchers have attempted to find the common form of Lotka's law to make its application more universal. After Pao's research, Canadian scholar Paul Travis Nicholls proposed the modified form of Formula 6.7 as follows:

$$f(x) = c/x^n \quad c > 0, \quad x = 1, 2, 3, \dots, x_{\max}, \quad (6.8)$$

where x_{\max} represents the maximum capacity within a given writing period of authors.

6.3 Development of Lotka's Law

6.3.1 Verification of Lotka's Law

In 1926, Lotka described the quantitative relationship between authors and the number of papers with a unique perspective. This relationship is called Lotka's inverse square law. However, the results failed to obtain sufficient attention during that time. After 15 years of silence, Lotka's law attracted worldwide attention, and thus, became the prelude to validation works for empirical laws.

We can understand the verification of Lotka's law from various fields of scientific research nearly seven decades after it was developed from Table 6.2.

Since 1926, scholars have hardly stopped validating the relationship between literature distribution and Lotka's law. However, given the different time backgrounds of scholars, their works present varying verifications. From 1926 to 1982, we can identify the number of literature related to the distribution of authors to illustrate this point (Table 6.3).

The data in Table 6.3 were obtained from 19 famous journals within the field of bibliometrics, thereby reflecting the related research history. As shown in the table, the topic gained sufficient attention until 1969. Since Lotka's paper entitled "The Frequency of Distribution of Scientific Productivity" was published in 1926, the number of related papers did not exceed six until 1969. In 1969, seven papers were published, thereby showing that the topic had come to a turning point in history. However, the maximum number of papers published about this subject had been reached. The *Journal of the American Society of Information Science* and the *Journal of Documentation*, which have a common characteristic, have gone through a long and continuous publishing history.

From the 1980s onwards, scholars are no longer limited to basic authentication work, but they also make comprehensive and in-depth research on the distribution of literature authors based on the study of Lotka's and other previous researchers. In the

Table 6.2 Verification of Lotka's law

No.	Verifier	Year	Papers (books)	Research field	Contribution and conclusion
1	A.J. Lotka	1926	Frequency distribution of scientific productivity	Chemistry/Physics	1. Lotka's law: $y = C/x^2$ 2. Inferences of Lotka's law
2	H.T. Davis	1941			1. Presented the distribution formula: $y = C/x$ ($0 < x$) 2. Distribution of scientific authors is close to Pareto's distribution
3	G.K. Zipf	1949	Human behavior and the principle of least effort		Lotka proposed an approximate calculation formula; it is not a strict probability distribution
4	D.H. Leavens	1953		Econometrics	1. Conclusion is consistent with Lotka's law 2. Proposed the K-S test
5	H.A. Simmon	1957	Models for man	Biometrics	1. Lotka's law is suitable for the distribution of scientific authors 2. Statistical tests are unnecessary
6	D.S. Price	1963	Little science, big science		1. Lotka's law is an inverse square law 2. Detected scientific elites using Lotka's law
7	R.A. Fairorne	1969			1. Refer to the relationships among Bradford's, Zipf's, Mandelbrot's, and Lotka's distributions 2. Lotka's law is more suitable for low yield authors
8	L. J. Murphy	1973	Lotka's law in humanities		Lotka's law is applicable to humanities; (note: errors occur in Murphy's fitting)
9	H. Voos	1974	Lotka's law and library science	Information science	1. $n = 3.5$ and 80% of all authors wrote only 1 paper 2. Test results are satisfactory at $\alpha = 0.5$ level
10	A.E. Schorr	1974	Lotka's law and forensic history	Library science	$n = 4$ and 80% of all authors wrote only 1 paper
11	A.E. Schorr	1975		Legal science	Field does not comply with Lotka's law
12	A.E. Rogge	1975		Anthropology	Field conforms to Lotka's law

past 80 years, basic theoretical research mainly involves the distribution of literature authors. From the 1980s until the 2000s, work has focused on applied research.

Research in the 1980s is represented by Pao. In 1986, Pao used 48 sets of data on the distribution of literature authors, including 20 subjects and 3 large library catalogs to verify Lotka's law. The value of n was between 1.8 and 3.8, and authors who wrote only 1 paper accounted for 52–91% of the total number of authors. Only

Table 6.3 Statistics of papers related to the distribution of authors (1926–1982)

Year	Number of papers	Accounted for total (%)	Year	Number of papers	Accounted for total (%)	Year	Number of papers	Accounted for total (%)
1926	1	0.21	1945	0	0	1964	3	0.63
1927	0	—	1946	0	—	1965	3	0.63
1928	0	—	1947	0	—	1966	3	0.63
1929	1	0.21	1948	2	0.42	1967	6	1.27
1930	0	—	1949	1	0.21	1968	2	0.42
1931	0	—	1950	1	0.21	1969	7	1.48
1932	1	0.21	1951	0	—	1970	12	2.54
1933	0	—	1952	1	0.21	1971	9	1.91
1934	1	0.21	1953	4	0.85	1972	26	5.51
1935	1	0.21	1954	2	0.42	1973	27	5.72
1936	0	—	1955	1	0.21	1974	29	6.14
1937	0	—	1956	4	0.85	1975	37	7.84
1938	2	0.42	1957	2	0.42	1976	46	9.75
1939	0	—	1958	2	0.42	1977	36	7.63
1940	0	—	1959	2	0.42	1978	34	7.2
1941	3	0.63	1960	3	0.63	1979	40	8.47
1942	0	—	1961	3	0.63	1980	61	12.92
1943	1	0.21	1962	4	0.85	1981	39	8.26
1944	2	0.42	1963	4	0.85	1982	4	8.50

7 sets of data in this research satisfy the inverted square law proposed by Lotka. In addition, Pao presented a new formula to estimate the value of C. The formula can be calculated using a computer, and thus, it provides a convenient condition to verify and develop Lotka's law through the use of advanced technology. Subsequently, Nichols proposed an extended form of Lotka's law. Furthermore, he and B.C. Griffith presented an improved method to estimate the value of n. He considered the floating range of n values in 1.5–4 rooms.

Y.S. Chen presented his views on Lotka's law based on Simon Yule's research. Xie's research on the parameter estimation problem of Lotka's law is comprehensive. He considered the establishment and testing of Lotka's law model typical examples of an extreme hypothesis; the key step is to test the hypothetical limit of the goodness-of-fit test. However, the proposed test program of Simon was unsatisfactory. In the goodness-of-fit test, the χ^2 test is frequently used; however, the empirical data of Lotka's law do not meet the statistical findings of an independent inspection and distribution with the same requirements. Moreover, the χ^2 test experiences difficulties in combined categories. Although, R.C. Coile proposed a one-sample test through the Kolmogorov–Smirnov (K–S) test, this test still cannot escape the abuse of "suspects." The study on the effects of fitting routine tests conducted by C.J. Cleser and Moore indicates that the confusion of a positive correlation and fitting insufficient is the general phenomena of the applications in the omnibus tests.

Chen believed that estimates of the n value in previous studies frequently neglected problems: ① the hypothesis-independent variable n was from 1 to

Table 6.4 Statistics of the distribution of papers related to authors (1983–2015)

Year	Number of papers	Accounted for total (%)	Year	Number of papers	Accounted for total (%)	Year	Number of papers	Accounted for total (%)
1982	0	—	1993	5	2.646	2004	5	2.646
1983	1	0.529	1994	6	3.175	2005	11	5.82
1984	0	—	1995	4	2.116	2006	7	3.704
1985	2	1.058	1996	2	1.058	2007	8	4.233
1986	2	1.058	1997	3	1.587	2008	6	3.175
1987	1	0.529	1998	6	3.175	2009	10	5.291
1988	1	0.529	1999	7	3.704	2010	11	5.82
1989	4	2.116	2000	2	1.058	2011	12	6.349
1990	0	—	2001	8	4.233	2012	16	8.466
1991	5	2.646	2002	7	3.704	2013	11	5.82
1992	7	3.704	2003	4	2.116	2014	10	5.291

infinity, i.e., continuous; and ② assuming that n was continuous, then no “jump” or discontinuity would occur. Although Nichols noticed the first question during that time, his work did not offer any fundamental breakthrough afterward.

Table 6.4 shows the number of papers with Lotka's theme. As indicated in the table, studies on Lotka's law presented a relatively flat course from 1983 to 2000; from 2001, the number of research papers on Lotka's law began to increase gradually; from 2009 onwards, the number of papers published annually reached 10 or more, thereby showing that new growth points emerged in the study of Lotka's law. Papers were published in dozens of international journals, with *Scientometrics* and the *American Society for Information Science and Technology Magazine* accounting for the largest number. In particular, 48 relevant literature were published in *Scientometrics*, which accounted for approximately 25% of the total number of papers.

Nicholls conducted an ongoing study on Lotka's law. In 1986, he modified Pao's authentication method in two manners and applied it to verify 15 examples from humanities, social sciences, and natural sciences.¹ In 1989, Nicholls provided a systemic summary of the measurement normative model, variables, organization and evaluation data, parameter estimation, and goodness of fit to verify previous research on Lotka's law. He presented a method that could ensure the consistency of the results of Lotka's law.²

Bookstein³ considered that in case of determined assumptions, Lotka's law would be established regardless of which method was adopted to count the authors.

¹Nicholls, P.T. Empirical validation of Lotka's law [J]. *Information Processing and Management*, 1986, 22(5): 417–419.

²Nicholls, P.T. Bibliometric modeling processes and the empirical validity of Lotka's law [J]. *Journal of the American Society for Information Science*, 1989, 40(6): 379–385.

³Bookstein, A. Informetric distributions, part II: Resilience to ambiguity [J]. *Journal of the American Society for Information Science* (1986–1998), 1990, 41(5): 376.

Rousseau and other scholars performed a system analysis on the robustness of Lotka's law and found that this law was disabled when counting score types for cooperation papers, thereby overturning the conclusion of Bookstein.⁴ In addition, Russo found that Lotka's distribution did not hold even when the integer counting method was used to deal with a large number of authors and papers, e.g., more than 100 authors.⁵

6.3.2 Contributions of Fracci

The significant contribution of Fracci to the initial verification of Lotka's law cannot be ignored. His proposition explains the entire validation of Lotka's law. Pao concluded the generalized manifestation of Lotka's law based on the work of Fracci. Moreover, Fracci had conducted a systematic study of Lotka's law since 1972. His conclusions include: ① some validation works did not exceed or even meet the initial level of Lotka's law during that time and ② a specific verification work lacked thorough research in terms of some of the proposed basic concepts.

In the verification of Lotka's law, given that the concept of Fracci was objective and rigorous, later researchers have amended their approach, as well as the study of Lotka's law from a scientific and comprehensive perspective, to promote rapid development in research.

In addition, Fracci discovered the two factors that influenced Lotka's distribution, and he explained his own ideas scientifically. The first factor is that the era and environment of the researcher directly affect the result of a study, i.e., the validation work exhibits certain artificiality. The second factor is that the number of authors in the statistical sample or the statistical sample size exhibits a relationship with the research outcomes. Researchers before Fracci did not notice these factors. The introduction of these two views was constructively significant to the further development of Lotka's law. The original data of Lotka's law had a huge capacity (including a statistically long time span that was more than one statistical sample datum). However, subsequent validations only selected one or two journals, or only a few years as statistical data; thus, the results frequently could not reflect the distribution of Lotka's law.

The verification works of recent researchers are similar to the views of Fracci. W.D. Roland believed that the time selection of sample data was extremely important to study the results of the distribution of literature authors. As academic research on a subject approaches maturity, the effect of the time factor will be increasingly significant given the increase in the number of authors with a small amount of low-yielding papers. Roland used data obtained from "Logical

⁴Rousseau, R. Breakdown of the robustness property of Lotka's law: The case of adjusted counts for multiauthorship attribution [J]. Journal of the American Society for Information Science, 1992, 43(10): 645–647.

⁵Kretschmer, H. and Rousseau, R. Author inflation leads to a breakdown of Lotka's law [J]. Journal of the American Society for Information Science and Technology, 2001, 52(8): 610–614.

Table 6.5 Distribution of authors of logical mathematics from 1974 to 1990

Papers	AF (before 1976)		AF (during 1977–1981)		AF (during 1982–1986)		AF (during 1987–1990)	
	OR	IR	OR	IR	OR	IR	OR	IR
1	39.78	38.78	42.74	42.82	46.15	48.08	53.32	52.33
2	14.42	14.89	14.70	15.38	15.52	15.73	15.01	15.86
3	8.12	8.51	8.34	8.45	8.32	8.18	7.32	7.90
4	5.43	5.72	5.29	5.53	5.33	5.14	4.59	4.82
5	3.84	4.20	3.96	3.97	3.84	3.59	3.17	3.28
6	3.02	3.27	3.05	3.04	2.78	2.68	2.28	2.40
7	2.95	2.64	2.65	2.42	2.52	2.09	2.04	1.84
8	2.21	2.20	2.00	1.99	1.88	1.68	1.50	1.46
9	1.92	1.87	1.97	1.67	1.75	1.39	1.39	1.20
10	1.95	1.61	1.76	1.43	1.43	1.17	1.14	1.00
11	1.65	1.41	1.50	1.24	1.19	1.00	0.94	0.85
12	1.21	1.25	1.08	1.09	0.87	0.88	0.68	0.73
13	1.11	1.12	1.05	0.97	0.83	0.77	0.65	0.64
14	1.25	1.01	1.06	0.87	0.82	0.68	0.64	0.56
15	0.94	0.92	0.74	0.78	0.59	0.61	0.47	0.50
16	0.74	0.84	0.67	0.71	0.56	0.55	0.44	0.44
17	0.79	0.78	0.63	0.65	0.48	0.50	0.38	0.40
18	0.67	0.72	0.60	0.60	0.48	0.46	0.38	0.36
19	0.61	0.67	0.50	0.55	0.38	0.42	0.30	0.33
20	0.54	0.62	0.43	0.51	0.33	0.38	0.26	0.30
Authors	5935	5529	8558	8107	11624	11163	14813	14352
Papers	37161		46496		53965		58281	
	n = 1.381		n = 1.477		n = 1.612		n = 1.719	

Mathematics Bibliography in Europe and America" from 1974 to 1990, including 47000 papers from 15000 authors.

In this table, OR denotes the observation results, IR indicates the inspection results, and AF represents the frequency of authors.

The data in Table 6.5 show the analysis for all the authors (including new ones) in a five-year interval. The data in the second column indicated that 5935 authors were in the field before 1976. New authors of essays comprised 39.78% of the total number of authors, and n was less than 1.4; hence, it did not belong to the validation interval of Pao. In addition, given the statistics of author data between 1981 and 1986 and between 1974 and 1990, the value of n ranges from 1.381 to 1.719. Authors who published more than 20 papers are not included in the table. The statistical results showed that the K-S test was significantly consistent under the level of 0.01.

In addition, each discipline has various stages of development at different times; hence, scholars also vary in writing for the discipline. Moreover, the number of

budding authors is small; thus, their research remains minimal. A breakthrough in academics will be achieved when the subject “pioneer” appears. Therefore, describing the distribution during this period is not a good representation of Lotka's law. During the peak of subject development, a huge increase in the number of authors is observed, academic capacity increases, and core authors emerge. In this stage, Lotka's law can perform effectively in terms of the distribution of authors. In the mature period, the quantity of authors increases, the growth rate slows down, the subjects are varied, and the high-yield data of authors continuously increase. However, the distribution of authors in this stage will be less than that in the second stage. Therefore, the distribution of the total number of authors and the number of authors who continue to write is closely related in different periods given the varying behavior of authors when writing.

Table 6.6 lists the distribution of 435 authors who wrote more than 2 papers and continued to write for 6 to 10 years from 1974 to 1990. “A” represents the percentage of authors who wrote for six years in 1974–1980. “B” shows the case from 1975 to 1980. “C” represents authors who wrote for 10 years from 1974 to 1990. “D” denotes the number of papers in 1971–1990.

From the statistics, authors who continued to write for six years belonged to the discipline during its infancy. The distribution of these authors is skewed, which

Table 6.6 Statistics of authors who continued to write in 1974–1990

Number of papers	A	B	C	D
2	28.05	36.31	14.06	32.22
3	20.92	19.54	17.19	17.82
4	17.47	11.48	7.81	11.72
5	11.03	8.34	10.55	7.49
6	7.13	5.97	8.20	6.27
7	4.83	4.01	8.59	4.27
8	3.45	3.64	5.47	4.67
9	2.99	1.59	8.20	2.64
10	1.61	2.37	5.47	1.73
11	0.69	1.32	2.73	1.42
12	0.69	0.87	0.39	1.39
13	—	1.18	3.13	1.05
14	—	0.64	3.13	1.12
15	—	0.64	0.39	0.81
16	0.69	0.5	0.78	0.71
17	—	0.27	0.39	0.51
18	0.23	0.23	1.17	0.64
19	—	0.09	—	0.54
20	—	0.18	—	0.37
More	0.23	0.82	2.34	2.61

differs from the distribution of Lotka's law. Authors continued to write for 10 years mostly during peak periods; hence, Lotka's law can be used to describe their distribution.

6.3.3 *Development of Lotka's Law in China*

The introduction of informatics in China was relatively late because of historical reasons. Therefore, research in this field is extremely young. China did not study Lotka's law until the early 1980s.

(1) Incipient studies

Wen Shangwu was the first to verify Lotka's law in China. In 1985, Wen published the paper, *Some issues related to Lotka's law*. He chose the catalogues of the Beijing Library as research objects and collected seven data sets, including data from the 3–79 author catalogue boxes of the Chinese Science and Technology Library, the Japanese philosophical discipline bibliography, the STAR cumulative index of author names starting with R in 1979, and part of GRA-I's cumulative authors in 1978.

Wen mainly discussed the extension and applicability of Lotka's law. He adopted Cole's perspective and criticized the works of Murphy and Shaw. His paper was not only the first to verify Lotka's law in China, but also the most comprehensive. His inchoate research laid the necessary foundation for the in-depth development of Lotka's law in China.

(2) Application of Lotka's law

In 1987, Professor Wang Chongde, a famous Chinese scientist, conducted a study on the distribution of authors in the field of information science in China. His study included the following aspects:

① Collection and collation of data

Wang collected five journals that indicated the development status of Chinese information science as the data carrier, including Information Science (1980–1985), Journal of the China Society for Scientific and Technical Information (1982–1985), Journal of Information Science (1980–1985), Library and Information Service (1980–1985), and Technical Information Service (1979–1985). However, work experiences, translations, news, and book presentations were not included (Table 6.7). In the table, PN indicates the proportion of authors that are accounted for, and CN denotes the cumulative proportion of authors that are accounted for.

Table 6.7 Distribution of authors in five journals

Pieces/person x	Number of authors y	Number of papers $x \times y$	PN $y / \sum y$	CN $\sum y / \sum y$
1	1105	1105	0.6936	0.6936
2	235	470	0.1475	0.8411
3	95	285	0.0596	0.9008
4	52	208	0.0326	0.9334
5	31	155	0.0194	0.9529
6	19	114	0.0119	0.9648
7	14	98	0.0087	0.9736
8	10	80	0.0062	0.9799
9	8	72	0.0050	0.9849
10	7	70	0.0043	0.9893
11	5	55	0.0031	0.9924
12	4	48	0.0025	0.9949
13	2	26	0.0012	0.9958
14	1	14	0.0006	0.9968
15	2	30	0.0012	0.9981
16	1	16	0.0006	0.9987
18	1	18	0.0006	0.9993
19	1	19	0.0006	0.9999
Σ	1593	2883		

② Evaluation of the value of n

Formula 1.2 is transformed to obtain the linear relationship between $\log x$ and $\log y$ as follows:

$$\begin{aligned} \log(x^n * y) &= \log C, \\ n \log x + \log y &= \log C. \end{aligned}$$

The linear relationship between $\log x$ and $\log y$ is unsuitable for high-yielding authors. Therefore, on the basis of the approximate calculation formula for outstanding scientists proposed by Price, the data in Table 6.6 show that the total number of authors is 1593, from which 40 are removed, thereby accounting for 2.5% of the total number of samples. Table 6.6 presents a good linear relationship; hence, only 8 high-yielding authors (0.5% of the total), who have published more than 17 papers, are excluded from the table.

Table 6.8 Data for calculating index n

X	Y	X(logx)	Y(logy)	XY	X^2
1	1105	0.0000	3.0433	0.0000	0.0000
2	235	0.3010	2.3710	0.7136	0.0906
3	95	0.4771	1.9777	0.9435	0.2276
4	52	0.6020	1.7160	1.0330	0.3624
5	31	0.7000	1.4913	1.0439	0.4900
6	19	0.7782	1.2787	0.9950	0.6055
7	14	0.8451	1.1461	0.9685	0.7141
8	10	0.9030	1.0000	0.9030	0.8154
9	8	0.9542	0.9030	0.8616	0.9104
10	7	1.0000	0.8450	0.8450	1.0000
11	5	1.0413	0.6989	0.7277	1.0843
12	4	1.0791	0.6020	0.6496	1.1644
Σ		8.6810	17.0730	9.6844	7.4647

Table 6.8 shows the result of using the least squares method to calculate the value of n (Formula 6.1), and retaining the table data of 1/person to 12 person in Table 6.6. Relevant data are assigned to Formula 6.1 as follows:

$$n = \frac{12 * 9.6844 - 8.6810 - 17.0736}{12 * 7.4647 - 10.6810} = 2.2511.$$

The cooperation situation was still not included in the study of Wang. The value of n might have greater volatility when cooperation was considered. The shortcoming of his study was his disregard for the collaborators of scientific papers.

③ Calculation of the value of C

Lotka obtained $n = 2$ when the value of C was calculated. However, throughout the verification of later generations, we realized that $n = 2$ was a special case. The C values of Lotka were calculated against time to obtain $n = 2$. After generations of verification, $n = 2$ was confirmed to be a special case. This case frequently reflects the characteristic of certain subjects. The values of n are not common in different disciplines and even in various stages of development within the same discipline.

$$C = \sum_{x=1}^{\infty} \frac{1}{x^n} = \sum_{x=1}^{p-1} \left[\frac{1}{x^n} + \frac{1}{(n-1)} (p^{n-1}) + \frac{1}{2} (p^n) + \frac{n}{24} (p-1)^{n+1} \right] \quad (6.9)$$

When $n = 2.2511$, substituting it in Formula 6.9 obtains $C = 1/1.4554 = 0.6872$.

When $p = 20$, the error in Formula 6.9 can be neglected, thereby indicating sufficient accuracy. Thus, we can conclude that Lotka's distribution of Chinese authors in information science in recent years is

$$y_x = 0.6872/x^{2.2511}.$$

④ K-S test

To illustrate the consistency of theoretical calculations and actual statistical distribution, Wang conducted the K-S test. When the total statistical number of authors is Σy_x , the value of the statistical tests D can be calculated as $1.63/\sqrt{\sum y_x}$, that is, $1.63/\sqrt{1593} = 0.0408$. When $\alpha = 0.01$, $D_{\max} = 0.0154 < 0.0408$ = the value of D.

The null hypothesis states that the information science authors in recent years has been formed and has begun to develop. Most of the authors (68.72%) wrote only 1 essay. A total of 158 authors, which accounted for 10% of the total authors, wrote more than 4 papers. The total number of works was 36% of the total amount of papers. China currently has approximately 10 part-time workers who conduct intelligence works, i.e., $100,000 = 316$. The number of papers is over 1,600, which account for approximately 55% of the data in Table 2.5. The inference of this article is consistent with Price's law.

To maintain a linear relationship between $\log x$ and $\log y$, Wang deleted high-yielding author data in his research, including 8 authors who wrote more than 13 papers. Therefore, the information science field in China has few high-yielding authors. This result did not increase the value of n and did not affect Lotka's distribution.

(3) The Significance of fractal theory to Lotka's law

N is a dimensional concept in Lotka's law and can be described as D. This concept is the core of the law that represents the frequency distribution of the degree of dispersion of authors, subjects, or topics, and the average degree of the mutual penetration level of the quantitative description. N exhibits a close relationship to various disciplines, theme complexities, development conditions, and rules. The distribution of authors is more decentralized, i.e., a balanced distribution of authors, and the value of D is higher. The estimation method of D can be conducted with reference to the value of n.

$$D_i = \lim_{\delta \rightarrow 0} \frac{\sum_{N=1}^N P_i \ln P_i}{\ln \delta} \quad (6.10)$$

Through the analysis of the D values, the contribution of fractal theory to Lotka's law can be understood in the following aspects.

- ① Different powers of exponent n. The value of n in Lotka's law differs among various disciplines according to verification by scientists, which can be understood by the size of D. Different objects in various disciplines or thematic studies, theories, and methods are used. Therefore, objective conditions and experimental means require a level of knowledge as well as

qualified researchers from disciplines with different levels of difficulty, characteristics, development rules, conditions, and other factors. Moreover, the development stage and speed are inconsistent. More than one difference lead to varying distributions and writing capacities of authors. Therefore, the values of D reflected in different discipline personalities vary, as shown in Table 6.8.

Table 6.8 shows that the fractal dimension D increased in the order of natural sciences, technology, social sciences, and humanities. Therefore, basic natural sciences and technical sciences have higher levels of specialization than social sciences and humanities, i.e., the levels of knowledge and academic requirements of their research staff are higher. Moreover, laboratory equipment and instruments are required, such that the distribution of authors is centralized and disparity among individual writing abilities exists. Social sciences and humanities are broad and have lower quantitative level than natural sciences and technology; thus, the requirements and research conditions for researchers must be reduced. Therefore, research in these two areas is evenly distributed. In different areas, the values of D (i.e., the power exponent n) vary.

② Processing the statistics of collaborators

In previous verification processes of Lotka's law, two main methods were used to deal with collaborator data. The first method only counts the first author. The second method considers all the collaborators. The values of D and C differ in these two methods. D is higher in the first method than in the second method (Table 6.9). This condition is due to the unchanged self-similarity of author distribution when collaborators are considered. The original papers belonging to a single author now belongs to a few people. Moreover, the dispersion degree of authors and the values of D both increase. To truly reflect the phenomenon that intersects with contemporary disciplines as well as the penetration and increasing papers of collaborators, considering collaborators is more scientific.

In this table, NA indicates the number of authors, and NP denotes the number of papers.

Note: (single) means considering only a single author; (combined) means considering collaborators.

③ Deleting authors with high production.

When Lotka dealt with data, he deleted high-yielding authors; hence, data points did not fall into a regression line. Previous studies did not provide a convincing explanation for this treatment. The data in Fig. 6.4 can be obtained from the data of Lotka and Wang. In this figure, limited author data with high yield or low yield did not fall into the linear expression $\ln N(x) - \ln N$; hence, the intermediate region probably has a linear relationship. Fractal theory can provide a satisfactory explanation for this phenomenon. The double logarithmic graph shows a good linear relationship. The increasing portion is the self-similarity of the author distribution labeled as the number of papers

Table 6.9 Values of D in certain subjects

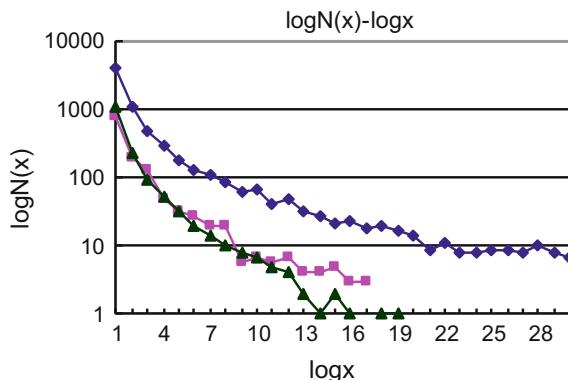
Subject or topic	NA	NP	D	C
Library science (single)	198	229	3.096	0.845
Library science (cooperation)	210	229	3.135	0.849
Forensic medicine (single)	997	1487	2.474	0.740
Forensic medicine (cooperation)	1010	1487	2.748	0.793
American revolution (single)	1316	2043	2.482	0.742
American revolution (cooperation)	1356	2043	2.546	0.755
Computer music (single)	458	970	2.168	0.662
Computer music (cooperation)	544	970	2.273	0.691
Ethnomusicology (single)	2269	4434	2.285	0.694
Ethnomusicology (cooperation)	2422	4655	2.291	0.696
History of technology (single)	164	231	2.616	0.769
History of technology (cooperation)	170	231	2.647	0.775
Information science (1968)	1666	2002	3.775	0.909
Information science (China)	1593	2883	2.251	0.687
Enflurane	432	262	2.539	0.754
Drosophila	826	3662	1.783	0.524
Physics	1325	3398	2.021	0.615
Chemistry (authors' names begin with letter A)	543	5355	1.898	0.571

and called the “non-scale area.” A small number of high-yielding or low-yielding authors are located in the upper and lower portions, respectively, of the non-scale area. Lotka's law and the D values are significant in the non-scale area. The most prolific authors are the core subjects, with their writing abilities belonging to a higher level. The writing ability of authors with low yields belongs to a lower level. The censored authors and high-yielding authors are represented by X_h and X_l , respectively, and $X_h \geq X \geq X_l$ represents the applicable scope of Lotka's law. When all partners are considered, X_1 may be less than 1. The values of X_h and X_l are retained throughout $\ln N(x) - \ln x$. The number of corresponding high or low yields can be obtained according to the values of X_h and X_l . The identification of the characteristics of the subject or topic and the analysis of the numbers of high-yielding and low-yielding authors are significant.

Note: In Fig. 6.4, the rhombus represents the chemistry data of Lotka, the square represents the physics data of Lotka, and the triangle represents the data of Wang.

(3) Author distribution of academic information on the Internet

Academic information is gradually developing toward the direction of networks. At present, the academic communities of networks include blogs, forums, and BBS. Compared with the traditional academic carrier, these Web 2.0-based academic communities with a network reflect many unique advantages, such as convenient

Fig. 6.4 $\log N(x)-\log x$ 

and free exchange of academic information as well as the worldwide spread of communication. Networks have gradually become one of the important channels for academic information exchange among scholars. Therefore, the author distribution network of academic studies has theoretical and practical significance. Qiu Junping and Yu Fan researched on author distribution on the Internet. They chose to study bloggers who signed up on scienonet.cn, and they analyzed 4546 blogger information from 2007 to 2010. A total of 8 subjects, including geoscience, engineering science, and comprehensive management, did not pass the K-S test. The research suggested that blog articles were easily published. Therefore, the distribution of authors frequently exhibited a growth phenomenon in a decreasing trend. The environment was the main factor that caused Lotka's fitting of author distribution to fail the K-S test.

Simultaneously, the difference between the distribution of bloggers and the distribution of paper authors was also determined. As illustrated in "Intelligence," the difference in the number of authors who published 1–10 papers from 1978 to 2007 was 629/86/22/10/4/6/3/3/1. In information science, the difference was 50/29/2/0/11/6/6/5/5. The results were attributed to differences in paper and blog articles. Papers were graded on a scale and could be mainly divided into authoritative journals, core journals, and general non-core journals. When the level is high, the quality of the paper is also high. Quality requirements include innovative points, words, and format. In general, when the level of a journal is high, the review period is long. Some journals take 1–2 years to be published. Blog articles do not have such rating. Anyone can publish blog articles any time, and not worry about quality issues. Blogs do not have a required number of characters and no reviewers. The relaxed environment of blogs has attracted more articles than journals, which also provide greater contribution to high-yielding authors. The per capita quantity of published articles in "Intelligence" from 1987 to 2007 was 1.29. However, more than 10 articles were on scienonet.cn. The authors with high yields in the author distribution for journals differ from those with high yields in the author distribution for scienonet.cn.

6.3.4 Research on Collaborators

With the rapid development of modern science, an increasing number of fast-paced and challenging interdisciplinary fields are being established. In particular, science and science effort presented a social trend after World War II. Organized scientific research and production became successful. Humans have entered the era of “big science,” as emphasized by Price. All of the aforementioned changes affect the cooperation and communication of scientists. Many research projects and tasks rely on collective disciplines to be completed. Individual research methods are changing to cooperative and collective approaches. Scientific studies are becoming increasingly complex. One of the most significant collaboration strategies is cooperative thesis. Cooperative dissertation can improve scientific labor organization, increase the number of scientific research, and improve the efficiency of scientific effort. Therefore, the study of the cooperation phenomenon will help analyze the distribution of literature authors.

Few studies have been conducted using Lotka's law as one of the distribution laws for authors in bibliometrics. Lotka only chose older collaborations that dealt with data from *CA* and *List of the History of Physics*. Kohl said that the results of the studies conducted by Lotka differed from the results of studies that counting all the statistics. The difference was due to the writings of the authors that were limited by scientific developments and communication means.

In 1987, D.K. Gupta used different data processing methods to study the distribution of authors in the field of entomology from 1900 to 1973. His statistical approach included: A. All authors, B. Only first authors, C. Non-collaborative, and D. Only co-authors. Statistics are shown in Table 6.9. The research conducted by Gupta showed that the values of n varied when adopting different data-processing methods. Moreover, the values of n had a direct relationship with the number of low-yielding authors; that is, the value of n would increase with the increase in the number of low-yielding authors. When the total number of authors was fixed, the value of n would be larger, and high-yielding authors would be less. Therefore, when the value of n was high, the capacity gap between authors was large.

The data in Table 6.10 show that the results are in the significant level of 0.01. The K-S test results are satisfactory (as shown in Table 6.11), thereby indicating that the distribution model of the author is related to statistical data processing.

In the table, NP indicates the number of papers, NA denotes the number of authors, and NA/T represents the percentage of NA composed of the total number of authors.

DS indicates data statistics, TA represents the total number of authors, and U0.01 denotes a significance level of 0.01.

(1) Distribution model for collaborations

The cooperation phenomenon has been extremely common in the academic community since the “invisible college” period. Price used the distribution of co-authors to study cooperation in the “academy.” He believed that many core authors would

Table 6.10 Distribution of authors in the field of entomology from 1900 to 1923

A			B		C		D	
NP	NA	NA/T%	NA	NA/T%	NA	NA/T%	NA	NA/T%
1	320	52.63	263	54.79	182	65.23	92	71.88
2	92	15.13	64	13.33	34	12.19	20	15.63
3	63	10.36	44	9.17	24	8.60	9	7.03
4	32	5.26	29	6.04	17	6.09	3	2.34
5	24	3.95	17	3.54	7	2.50	0	0.00
6	10	1.64	7	1.46	3	1.08	2	1.56
7	11	1.81	7	1.46	4	1.43	0	0.00
8	7	1.15	6	1.25	1	0.36	0	0.00
9	7	1.15	9	1.88	2	0.72	1	0.78
10	7	1.15	4	0.83	0	0.00	0	0.00
11	3	0.49	3	0.63	1	0.36	0	0.00
12	3	0.49	0	0.00	0	0.00	0	0.00
13	4	0.66	3	0.63	0	0.00	0	0.00
14	3	0.49	1	0.21	0	0.00	1	0.78
15	2	0.33	3	0.63	1	0.36		
16	1	0.16	3	0.63	1	0.36		
17	3	0.49	2	0.42	1	0.36		
18	3	0.49	2	0.42	1	0.36		
20	2	0.33	3	0.63				
22	1	0.16	0	0.00				
23	0	0.00	1	0.21				
24	1	0.16	2	0.42				
25	1	0.16	1	0.21				
26	1	0.16	0	0.00				
27	1	0.16	0	0.00				
28	0	0.00	1	0.21				
38	1	0.16	0	0.00				
39	0	0.00	1	0.21				
42	1	0.16	2	0.42				
43	1	0.16	0	0.00				
52	0	0.00	1	0.21				
53	1	0.16						
66	1	0.16						
Total	608		480		279		128	

be present in a group of academic scientists and some authors would be flowing around them. After flowing authors cooperated with core scientists once or twice, they would no longer appear. Simultaneously, Price found that a significant relationship existed between the number of papers and the average number of authors in each paper. Thus, Lotka's law cannot reflect the relationship between partners

Table 6.11 Results of the K-S test

Types	DS	TA	F(X)–S(X)	U0.01
	N	D	K-S test	Value of n
All authors	608	0.0468	0.0661	1.9
First authors	480	0.0388	0.0744	1.8
Single authors	279	0.0429	0.0976	2.2
Collaborative authors	128	0.0337	0.1441	2.4

and scientific papers. Price thought that a considerably appropriate variable should illustrate the relationship between partners and their paper.

The following equation can be obtained based on the idea of Price:

$$y = C/x^n. \quad (6.11)$$

Formula 6.11 is similar to the basic equation of Lotka's law. However, the difference between them is that x in this formula represents the average number of co-papers of authors, whereas x in the basic equation represents the number of papers written by authors. C and n are parameters, and x denotes factors that affect the cooperation thesis. Formula 6.11 can reflect the relationship between authors and their papers. Documents written under strong cooperation in the field of biotechnology were selected to verify Formula 6.11. A total of 50 papers, which involved 15 countries and 245 authors, were randomly selected from the field. Then, these papers were retrieved from the 1991 SCI. After the statistics were gathered, 100 authors without name repetition and doubt (i.e., 128 authors who did not publish this year and 17 authors who published 30–200 papers were not included) were considered. A total of 100 authors published 593 papers. Among which, 31 papers were from a single author, whereas the rest coauthor papers were 1495 times. The average cooperation per article was $(1495 + 100)/593 = 2.69$. That is, each article was written by 2.69 people, as shown in Table 6.12

In the table, NP represents the number of papers, NA indicates the number of authors, and TP denotes the total number of papers.

The deduction of Lotka's law shows the number of authors that published 1 paper at a certain time, which is 60% of all the authors. However, Table 6.12 shows that only 17% of the authors published 1 paper, whereas 16% of the authors published 1128 papers, thereby completing 50% of all the papers.

Table 6.13 shows the number of paper by collaborators. The data in the table reflect that authors with high yield account for 16% of the total, and cooperation accounts for 50%. The results suggest that authors with low yield frequently cooperate with authors with high yield to increase the quantities of their papers. Simultaneously, high-yield authors take the same approach to increase the number of their papers. Therefore, an increase in the number of co-authors results in an increase in the number of papers.

Table 6.12 Distribution statistics of papers

NP	NA		TP			
A	B	C = B/ΣB	D = Σ Ci	E	F = E/ΣE	G = Σ EI
1	17	0.17	0.17	17	0.0276	0.0276
2	19	0.19	0.36	38	0.0618	0.0894
3	12	0.12	0.48	36	0.0585	0.1479
4	7	0.07	0.55	28	0.0455	0.1934
5	10	0.10	0.65	50	0.0813	0.2747
6	5	0.05	0.70	30	0.0488	0.3235
7	5	0.05	0.75	35	0.0569	0.3804
8	4	0.04	0.79	32	0.0520	0.4324
9	4	0.04	0.83	36	0.0585	0.4909
10	1	0.01	0.84	10	0.0163	0.5072
11	3	0.03	0.87	33	0.0537	0.5609
13	2	0.02	0.89	26	0.0423	0.6032
15	2	0.02	0.91	30	0.0488	0.6520
16	1	0.01	0.92	16	0.0260	0.6780
19	1	0.01	0.93	19	0.0309	0.7089
20	2	0.02	0.95	40	0.0651	0.7740
21	1	0.01	0.96	21	0.0342	0.8082
22	1	0.01	0.97	44	0.0715	0.8797
23	2	0.02	0.99	46	0.0748	0.9545
28	1	0.01	1.00	28	0.0455	1.0000
Σ	100		1.00	593		

In this table, NP indicates the number of papers, NA denotes the number of authors, NC represents the number of cooperators, RC% indicates the relative percentage of C, CC% denotes the cumulative percentage of C, and AC represents the average number of cooperators.

In the preceding analysis, research on Lotka's law became optimistic after a new variable (i.e., the average number of collaborated papers of an author) was introduced. Data in Table 6.13 were deleted after the 11th row due to the volatility of authors with high yields. $n = 0.9641$ and $c = 0.781$, as obtained by the least squares method. To reflect cooperation, the thesis equation can be expressed as

$$y = 0.781/x^{0.9641}. \quad (6.12)$$

When the K-S test was used to verify Formula 6.12, consistency was under the significance level of 0.01. Therefore, the analysis is appropriate for research on the distribution of co-authors.

Table 6.13 Sorting of cooperation papers

NP	NA	NC	RC%	CC%	AC
A	B	C	D	$E = \sum Di$	$F = C/B$
1	17	62	0.0420	0.0420	3.65
2	19	92	0.0624	0.1044	4.84
3	12	78	0.0529	0.1573	6.50
4	7	49	0.0332	0.1905	7.00
5	10	134	0.0908	0.2813	13.40
6	5	56	0.0380	0.3196	11.20
7	5	83	0.0563	0.3756	16.80
8	4	84	0.0570	0.4326	21.00
9	4	88	0.0597	0.4923	22.00
13	2	54	0.0366	0.5980	27.00
15	2	83	0.0563	0.6543	41.50
16	1	15	0.0102	0.6645	15.00
19	1	77	0.0522	0.7167	77.00
20	2	116	0.0786	0.7953	58.00
21	1	23	0.0156	0.8109	23.00
22	1	74	0.0501	0.8610	74.00
23	2	132	0.0895	0.9505	66.00
28	1	73	0.0495	1.0000	73.00
	100	1493	1.0000		

(2) Hypotheses on and scale of collaborators

To facilitate research on collaborator problems, scholars have proposed the following hypotheses on collaborators. ① Different authors have varying capacities to write journal science papers, and these capacities are not uniform. Moreover, few papers are completed by only one author. ② Cooperation is mutual within the scientific collaborative organization. An author can have many partners, which can be standardized. ③ The more papers an author publishes, the more collaborators he/she has.

For the aforementioned hypothesis, we choose *Acta Physico-Chimica Sinica* in 1994–1995 as the data carrier and randomly selected 90 authors without cooperative relationship among them as samples. A conclusion was drawn after the data analysis and the K-S test. That is, relationships among collaborators exist within scientific collaboration, and more collaborators are connected to authors with high yield. The quantitative relationship among them can be generalized using statistical tools. Specific situations depend on the overall level of collected data, and a negative exponential relationship exists.

Table 6.14 Collaborative indicators of papers from different subject areas in the United States (1977)

Subject or field	Cooperation degree	Cooperation rate
Physical science	2.33	0.20 0.18
Mathematics	1.32	0.07 0.07
Computer science	1.49	0.05 0.04
Environmental sciences	1.55	0.19 0.22
Engineering science	2.06	0.03 0.04
Bioscience	2.01	0.40 0.44
Psychology	1.74	0.10 0.09
Social sciences	1.35	0.22 0.20
Other science fields	1.42	
Average	1.79	0.14 0.14

As research on collaborators deepens, scholars have prescribed new metric indicators, such as cooperation rate and cooperation degree, which reflect the intelligence level of cooperation among journal authors. When the value is high, cooperation intelligence is being fully played.

$$CD = (\text{total number of authors}/\text{total number of papers}) \times 100\%$$

Note: CD indicates the cooperation degree, i.e., the total number of authors divided by the total number of papers from a certain journal at a certain time.

$$CR = (\text{number of collaborative papers}/\text{number of papers}) \times 100\%$$

Note: CR indicates the cooperation rate, i.e., the total number of collaborative papers divided by the total number of papers from a certain journal at a certain time.

From the data of the survey undertaken in 1977 by King Research Inc., the CD and CR of different subject areas in the United States can be obtained, as shown in Table 6.14. The table indicates that most researchers in the United States currently focus on physics, engineering science, and bioscience. Scientific effort and research personnel that exhibit good scientific cooperation achieve the best results.

6.4 Price's Law and the Distribution of Other Authors

6.4.1 Price's Law

Famous laws in informetrics focus on the phenomenon of “outstanding person” in the literature. However, the meaning of “outstanding person” varies among different laws. For example, Lotka refers to authors with high yields, Bradford refers to the core journal with the highest number of papers, and Zipf refers to the widely applied frequency words. The verification of various laws will benefit from rationally disposing of the issue of “outstanding person.”

The data processing method of Lotka for “outstanding person” is relatively simple. He merely used a number of points and the fitting line model equation. The data points of high-yielding authors would not directly follow a straight line; thus, Lotka deleted high-yielding authors, which accounted for 1.3 and 1.02% of the papers from *CA* and *List of Physics History*. Lotka's law exhibits flaws that do not offset its advantages in the research results. However, these flaws have not prevented Lotka's law from becoming one of the three famous laws in informetricis. Lotka was an outstanding scholar with an excellent writing ability. He was a driving force in promoting the development of this discipline. Thus, an in-depth research on authors with high yield will facilitate the distribution of literature authors.

Price, the famous American historian of science, was the first to realize the significance of high-yielding authors. In his research, he found that only 75% of scientists published a paper during their lifetime, whereas 10% of the published papers of scientists accounted for half of all the papers published in their lifetime. Price pointed out that the number of high-yielding authors who wrote half of all the papers was equal to the square root of all the science authors in his book *Big Science, Small Science* published in 1969. Price's law can be described as

$$\sum_{m+1}^I n(x) = \sqrt{N}. \quad (6.13)$$

In Formula 6.13, $n(x)$ represents the number of authors writing x papers, $I = n_{\max}$ denotes the number of papers from the highest-yielding authors, and N indicates the total number of all the authors.

m can be ensured as follows:

$$\sum_1^m x \cdot n(x) = \sum_{m+1}^I x \cdot n(x). \quad (6.14)$$

Let $a(n)$ indicate the number of authors who published n papers. Then, the number of authors who published $n \leq N \leq n'$ papers is

$$A(n \rightarrow n') = a(n) + a(n+1) + \cdots + a(n') = \sum_{i=n}^{n'} a(i). \quad (6.15)$$

$a(n)$ represents authors who published papers cooperatively: $P(n) = n a(n)$.

Then, the total number of papers published by the authors whose amount of published papers is, i.e., $n \leq N \leq n'$

$$\begin{aligned} P(N) &= P(n \rightarrow n') \\ &= na(n) + (n+1)a(n+1) + \cdots + n'a(n'). \\ &= \sum_{i=n}^{n'} P(i) \end{aligned} \quad (6.16)$$

If $0.5 P(1 \rightarrow n_{\max}) = P(m \rightarrow n_{\max}) = P(1 \rightarrow m)$, then

$$A(1 \rightarrow n_{\max})^{0.5} = A(m \rightarrow n_{\max}). \quad (6.17)$$

From Lotka's law, $a(n) = C/n^2$, and $a(n)$ is the frequency.

Then,

$$p(1 \rightarrow n) = \sum_{i=1}^n i \cdot \frac{C}{i^2} = \sum_{i=1}^n \frac{C}{i} = C \sum_{i=1}^n \frac{1}{i}.$$

The harmonic series summation formula obtains the following:

$$P(1 \rightarrow n) = C(\ln n + 0.577 + \dots + \varepsilon n).$$

Then, the preceding formula can be transformed into:

$$0.5 C(\ln n_{\max} + 0.577 + \dots + \varepsilon n_{\max}) = C(\ln m + 0.577 + \dots + \varepsilon m).$$

$$\ln \frac{(n_{\max})^{\frac{1}{2}}}{m} = 0.289 + \dots + \frac{1}{2}\varepsilon m + \frac{1}{2}\varepsilon n$$

$\varepsilon m \leq 0.289$ is reasonable given that m will not be considerably less. Therefore, $0.5\varepsilon n_{\max} \leq \varepsilon m$. Thus,

$$m = 0.749(n_{\max})^{0.5}. \quad (6.18)$$

The preceding formula indicates that the total number of papers published by scientists was more than $0.749(n_{\max})^{0.5}$, which is equal to half of the total number of papers. Therefore, the number of papers from one of the lowest-yielding authors among these outstanding scientists was equal to 0.749 times the square root of the number of papers from the highest-yielding scientists.

The following are obtained from Formula 6.15:

$$\begin{aligned} A(1 \rightarrow n_{\max}) &= \sum_{i=1}^{n_{\max}} a(i) = \sum_{i=1}^{n_{\max}} \frac{C}{i^2} \\ &= C \left[\frac{\pi^2}{6} - \frac{1}{n_{\max}} + o\left(\frac{1}{n_{\max}^2}\right) \right], \end{aligned} \quad (6.19)$$

$$\begin{aligned} A(m \rightarrow n_{\max}) &= \sum_{i=m}^{n_{\max}} \frac{C}{i^2} \\ &= C \left[\frac{1}{m} - \frac{1}{n_{\max}} + o\left(\frac{1}{n_{\max}^2}\right) + o\left(\frac{1}{m^2}\right) \right]. \end{aligned} \quad (6.20)$$

Let $n_{\max} \geq (n_{\max})^{0.5}$, $n_{\max} \geq 6/\pi^2$, then

$$R = \frac{A(m \rightarrow n_{\max})}{A(1 \rightarrow n_{\max})} = \frac{\frac{6}{\pi^2}}{0.749(n_{\max})^{0.5}} = \frac{0.812}{(n_{\max})^{0.5}}. \quad (6.21)$$

Formula 6.21 indicates the proportional relationship between the total number of outstanding scientists and the total number of scientists, which is the formula for Price's law. This law states that if the total number of papers is n , then the proportion of low-yielding authors will be small. The number of authors is in line with the square root law due to the symmetry in this case. Therefore, Price's law is also known as the square root law. From its derivative results, Price's law is similar to Rousseau's law in the field of social sciences. Rousseau's law pointed out that the outstanding elite are $N^{0.5}$ in the overall size of N .

In general, the statement of Price on scientists (i.e., literature authors) remains a hypothesis. Although the scientific labor hypothesis is indispensable, it must develop from the original hypothesis to the theoretical one to finalize theory construction. Otherwise, the hypothesis is not only stagnant, but also hovering in the primary stage. Price's law does not belong to this category. Its role is to describe the relative relationship macroscopically and completely, as well as to guide in estimating the scale and writing ability of high-yielding authors. The application of Price's law is dissimilar to that of the law of precision; otherwise, research will be meaningless.

6.4.2 Distribution of Other Authors

Researchers have begun exploring a new distribution model for literature authors and conducting verification of Lotka's law. Shockley proposed the lognormal model. C.B. Willames discovered the relationship between biology and publishing models, as well as examined Fisher's logarithmic series of geometric model. Simon

proposed the function of Joel Distribution. The two influential author distribution models are as follows: (1) Function β of Simon

Simon proposed this mechanism because skew distribution has different confusing old and new concepts. In scientific publications, Simon's occurrences are as follows.

Assumption 1: The $(t + 1)$ th paper was published by a new author; its probability was α . No new author published the first t papers.

Assumption 2: The $(t + 1)$ th paper was published by the authors who published n papers. Its probability in proportion to n was $f(n,t)$, i.e., the proportion of the total number of papers to the n papers that the authors published. $f(n,t)$ is the number of authors who had published n papers in the original paper.

Let v be the investigated number of papers from different authors. The expected number of authors who published n papers given the two assumptions should be

$$f(n) = V p B(n, p + 1) n = 1, 2, 3, \dots \quad (6.22)$$

where $p = 1/(1 - \alpha)$ and $B(n, p + 1)$ are functions β with parameters n and $p + 1$. Yoel first derived the same equation in the biological problem for modern theory of stochastic processes; hence, Simon stated that the last equation should be Joel Distribution.

(2) Research conducted by Bookstein

The distribution of authors of Bookstein is as follows:

$$g(n) = g\left(\frac{n}{\lambda}\right) f(\lambda) d\lambda. \quad (6.23)$$

where $g(n)$ is the number of authors, λ is the writing ability of an author, $g(n/\lambda)$ is the proportion of authors who published n papers, and $f(\lambda)d\lambda$ denotes the total number of individuals who processed n within interval $(\lambda, \lambda+d)$.

Bookstein also analyzed the effect of social environment on scientists who write scientific papers. He believed that when the social environment was conducive for writing and promoted positive writing behavior, then the distribution of authors would be closer to Lotka's law.

In addition, Indian bibliometric metrologist Rao pointed out that in real life, the Matthew effect could not be neglected. Success generates success. He used the secondary negative distribution to describe the distribution of authors. In addition, he suggested that the negative binomial Poisson's distribution method could express bibliometrics, including collaborators, through translation transformation or probability conversion.

After the analysis of the aforementioned description and author distribution, the research conducted by Bookstein can be used to summarize the distribution of authors as follows:

$$f(x) = k/x^\alpha \quad (x = 1, 2, 3, \dots; k > 0; \alpha > 0), \quad (6.24)$$

where k and α are constant values, i.e., the proportion of the number of papers to the published x papers and $1/x$. Therefore, Formula 6.24 exhibits the following characteristics: A. the fitted curve presents an inverted V type; B. highly skewed; and C. the distribution has a “long tail,” which is infinitely close to the x -axis but never intersects with it.

6.5 Application of Lotka's Law

Since the 1970s, in the course of research and exploration, scholars have gradually transformed Lotka's law from a non-observability theory to a direction with practical applications. Research on Lotka's law has been applied to various disciplines.

6.5.1 Function of Lotka's Law

(1) Reflecting the condition of the achievement of technology labor

The number of published scientific papers is the main indicator for evaluating the work of researchers. Researchers exert effort to gain acceptance to society by publishing papers. This behavior frequently results in the inferior quality of technological papers. However, the passion of authors to publish and the number of authors are continuously increasing. Therefore, we studied the scientific research conditions under which scholars performed as technological literature for a certain period and a particular subject areaby applying Lotka's law.

For example, Chinese intelligence expert Wang studied the development of intelligence in China using Lotka's law. The results indicate that scientific intelligence teams have been formed and begun to take shape in 1985. However, few high-yielding authors belong to the current intelligence field of China. The use of the quantity method by Wang to describe the distribution of Chinese intelligence authors was impressive and achieved good results. Moreover, the method exhibits the advantages of deep penetration reasoning, strong recapitulation, accurate evaluation, and predicting capability.

(2) Scientific estimation of the labor scale

The scientific labor scale is an important content in social science. People increased labor productivity and developed the organizational structure of the labor scale by improving scientific labor organization and increasing scientific research

achievements. The main performers are the scientific researchers, and the implementors are the subject researchers. To achieve these objectives, the distribution of papers from researchers must be optimized. In modern science, scientific cooperation is indispensable and has become a major trend in promoting the development of subjects. In the overall development of scientific papers in 1920, the proportions of single authors were 95 and 82% in the fields of astronomy and chemistry, respectively. However, the ratings decreased to 68 and 30% in 1963. The proportions of collaborative research that received the Nobel Prize were 41% in 1910–1925 and increased to 79% in 1951–1972.

Chinese scholars studied the Chinese scientific labor scale according to four journals, including *ACTA MICROBIOLOGICA SINICA*, in 1976–1980 and to *SCIENCE CHINA* in 1982–1988.

Table 6.15 represents the scientific research level in China during the 1970s. Table 6.16 represents the scientific research level in China during the 1980s. The scale of scientific work of natural sciences was constantly expanding, which could be illustrated by the substantial increase in the cooperation degree of scientific papers.

(3) Mastering the scientific paper of a team of authors

The composition of a scientific paper is a complex and dynamic structure. We can understand the characteristics of scientific activities, grasp the rules of scientific development, and forecast development trends based on the statistics and measurement of the structure of the authors of scientific papers. Price examined authors in accordance with certain years and divided them into seven groups according to the number of papers they published and the numbers of continuous authors, core authors, and non-core authors. In this method, the types of authors are distinguished and author groups are understood. The following aspects can be considered for scholars who are studying scientific papers.

① Occupational structure of workers and research orientation

Let a_1 represent the number of theoretical authors, a_2 the number of practical authors, and A the proportion of a_1 to a_2 . Thus,

$$A = a_1/a_2.$$

If $A < 0.5$, then the number of theoretical authors are less than 1/3 of the total authors, and the applied authors account for more than 2/3 of the total. This result shows that subjects give priority to applied research. If $0.5 < A < 2$, then minimal difference is observed between the number of theoretical and applied authors, and subjects include both theoretical research and applied research. If $A > 2$, then the number of theoretical authors are more than twice the number of applied authors, which illustrates that the theoretical study of a subject is active.

Table 6.15 Statistics related to science papers from four journals

NA/per	Acta Microbiologica Sinica		Acta Chimica Sinica		Acta Physica Sinica		Acta Mathematica Sinica	
	NP	NA	NP	NA	NP	NA	NP	NA
1	14	14	24	24	138	138	181	181
2	42	84	53	106	76	152	35	70
3	50	150	30	90	68	204	6	18
4	37	148	20	80	28	112		
5	18	90	13	65	13	65		
6	14	84	7	42	7	42		
7	8	56			1	7		
8	5	40			3	24		
9	1	9	1	9	1	9		
10	2	20						
11								
12	1	12						
13								
14	1	14						
CA*	13		37		22			
Total	206	721	185	416	357	753	222	269
NA/per	3.37		2.81		2.24		1.21	

* Indicates the factory, academic community, school, and research unit. NA/per denotes the number of authors for each paper. NA indicates the number of authors. NP denotes the number of papers. CA represents the corporate authors

Table 6.16 Cooperation degree in related science subjects in *SCIENCE CHINA* in 1982–1988

Subject	Year							
	Cooperation degree							
	1982	1983	1984	1985	1986	1987	1988	Average
Mathematics	1.88	1.32	1.21	1.43	1.30	1.33	1.52	1.34
Physics	1.93	1.94	1.50	2.24	2.18	2.65	2.77	2.30
Astronomy	4.29	2.14	3.00	2.31	3.21	2.83	2.18	2.55
Technical science	364	1.91	2.53	2.29	3.00	2.43	2.23	2.34
Chemistry	2.00	3.39	4.24	4.10	3.75	3.81	3.72	3.99
Biology	4.25	3.61	3.67	3.97	3.77	3.68	3.16	3.65
Agronomy	2.81	3.30	4.00	2.00	2.40	2.50	2.00	2.46
Medicine	1.28	3.80	4.29	3.82	4.13	4.10	3.93	4.12
Geography	2.30	3.57	2.84	1.89	2.00	1.91	2.16	2.45

② Examination of the influence of author structure on journals from another aspect

Research shows that the authors of journal papers consist of two groups. The first group comprises authors who have already published papers. The second group is composed of new authors who are publishing papers for the first time. To facilitate comparison, the equation can be obtained by neglecting the influence of the number (N) of papers published on journals:

$$\frac{A}{N} = \frac{B}{N} + \frac{C}{N},$$

where A denotes the sum of new authors (B) and old authors (C); A/N indicates the author coefficient of the literature unit; B/N reflects the situation of the duplicable writings of authors; and C/N represents the emergence of new authors in a journal, which can be called the author increment of a literature unit. When the value of C/N is large and close to that of A/N, the journal frequently updates its authors and the published articles are mainly provided by new authors, thereby indicating the emergence of a large number of new authors. However, such case requires further analysis. For example, the result may be attributed to new subjects attracting more technical research and writing, or that a particular discipline, which is fraught with a major breakthrough in the period, has a large number of authors who published scientific papers. By contrast, the result indicates that the journal has several shortcomings, such as the decline in academic standards, changes in editorial policy, the emergence of new journals, and the substitution of old journals. The condition for authors from the *Journal of the Chinese Chemical Society* in 1952–1965 shows that 1465 total authors and 618 new authors are available over this 14-year period (Table 6.17).

CY indicates the corresponding years.

Table 6.17 Increase in the number of literature authors in the *Journal of the Chinese Chemical Society*

Roll	19	20	21	22	23	24	25	26	27	28	29	30	31
CY	1952 1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
A/N	1.84	1.62	2.17	2.46	3.37	2.57	2.38	2.26	2.26	2.60	2.48	2.24	2.80
C/N	1.56	1.06	1.22	1.06	1.15	1.06	1.21	1.21	1.15	1.21	1.14	1.10	1.20

③ Regional structure of authors and scientific activity centers

The Japanese scholar, Tangqianguangchao, identified the indicators for scientific activities according to the number of scientists and the achievements of each country, which accounted for the proportion of the total number of scientists and scientific results worldwide. He assumed that an activity center would be formed when its scientists and achievements accounted for over 1/4 of the total. Let c_i be the number of authors of the (i) country or region, N be the total number of authors, and C be the structural relative number of c_i and N. Then,

$$C = c_i/N \times 100\%.$$

From the perspective of Tangqianguangchao, the region will be a scientific activity center when $C > 25\%$.

④ New and old structures used by authors and the development stage of discipline

Let d_1 be the number of old authors, d_2 be the number of new authors, and D be the relative ratio of d_1 to d_2 within a certain period. Then,

$$D = d_1/d_2.$$

When $D < 1$, then less old authors and more new authors are available, and the frequent updating of authors indicates that the subject is in the start-up and development stages. As the value of D decreases, the subject becomes more immature. When $D > 1$, more old authors and less new authors are available. The research team is stable, and the discipline in the mature and aging stages. As the value of D increases, the subject ages more rapidly.

⑤ Cooperation structure of authors, research complexity, and discipline multidisciplinarity

Let G be the average number of authors of papers, H be the average number of authors of cooperative papers, and M be the number of the major authors of cooperative papers. The high values of G and H in a subject indicate that the research process and methods for the subject are complicated. By contrast, their low

values indicate that the research process and methods are relatively simple. The high values of G and M in a discipline indicate that the discipline is complicated. By contrast, the low values of G and M in a discipline indicate that the discipline is specialized.

6.5.2 Problems that Should Be Noticed During Application

Lotka's law can scientifically solve the distribution of authors in scientific bibliometrics. However, some problems arise in the scientific application of Lotka's law. Lotka excluded high-yielding authors and selected only the number of authors whose first names began with A and B in CA when gathering statistics. The statistical results of the signature of authors (including collaborators) show that author names that began with A–F accounted for 57%, G–M accounted for 30%, and N–Z accounted for 13%. Therefore, Lotka's law is not as accurate and rigorous as the laws of physics. In addition, Lotka's law is empirical, and thus, is influenced by many random and cultural factors. Lotka's law has different manifestations in varying times for various disciplines. The development of a discipline typically undergoes three stages: incubation, growth, and maturity. In these stages, the activities of scientific authors vary. Therefore, Lotka's law cannot fully describe the distribution of authors throughout the entire process of a subject. The capability to describe the distribution of authors during maturity is optimistic.

Chapter 7

Statistical Analysis Method for Literature Information

7.1 Significance and General Concept of Literature Information Statistics

The word “statistics” originates from the Latin term “status,” which means the “state and situation of various phenomena.” At present, “statistics” generally refers to statistical data, statistical work, and statistical science. These three concepts have similarities and differences. Statistical data refer to digital information that reflects the characteristics and regularity of a large number of phenomena. Statistical work refers to collection, collation, and analysis that make inferences to statistical data. Statistical data result from statistical work, and statistical science is the theoretical generalization and sublimation of the entire statistical work. Statistics is the knowledge of research and the principles and methods of counting. Its basic objective is to investigate, summarize, analyze, and interpret the data of a research object as well as to determine its quantitative characteristics and objective laws. In terms of nature, statistics belongs to applied disciplines or methodologies. In general, the function of statistical methods can be summarized in three aspects: ① an important tool for understanding and reforming things, ② a powerful implement for scientific management, and ③ a necessary means for macro control and micro adjustment.

7.1.1 Literature on Information Statistics and Its Significance

Every object is characterized by “quantity” and “quality,” and the development law of any object is presented through these two aspects. Statistics focuses on the quantity aspect from the dialectical unity of objects; thus, it has universal applicability. At present, statistics is widely applied in many disciplines and

departments. In the field of literature research and practice, statistical methods are common traditional methods. In human society, statistics has produced several problems in counting the number of documents since the emergence of literature. Literature statistics has started with a simple counting of literature.

Statistical activities on literature information have been introduced into the daily operations of libraries. With the gradual increase in the number of documents, literature statistics has become an urgent necessity. The task has become evident, and the scope has expanded. In recent years, given the requirements for the scientific management of libraries and information institutions, as well as the need for literature information works, many libraries and information institutions have shown interest in the statistical analysis of literature information. Literature information statistics has become an indispensable part of literature work, and even of library and information works. Literature information statistical analysis methods have been developed as important research methods in informetrics.

Literature information statistics refers to the measurement of a specific unit or its relevant characteristic information. The literature information statistical analysis method is the analysis of literature information using statistical methods. This method describes the quantity characteristics and variety rules of the literature using the data to achieve certain research purposes. Under the guidance of statistical principles, the method is gradually developed in the long-term practice of literature information counting.

In the informetrics research, the literature information statistical analysis method has extensive practical significance, which is mainly manifested in the following aspects.

- (1) Basis and conditions of literature information quantitative research. For each discipline and for various types of literature, information quantitative research is inseparable from literature information statistical data. Literature information statistics is the basic work in literature information research. Literature information data are necessary components to perform literature information quantitative research.
- (2) Determining the quantitative change law of literature information. Through the statistical analysis of literature information, the growth change, distribution characteristics, circulation conditions, utilization level of literature, and the law of literature number changes can be reflected.
- (3) Basis for library and information management and literature information management. Libraries and information centers must strengthen scientific management, which is an urgent requirement of an objective situation. Literature management is an important part of library and information management. Scientific management requires the decision of the manager to be based on the full understanding and thorough analysis of facts, which must be explained by timely and accurate data. Statistics is a powerful tool for presenting and analyzing numerical facts. Therefore, the literature information statistical analysis method, which is formed by the application of statistics to the field of literature information, is an important tool for management.

7.1.2 General Concept of Literature Information Statistics

A set of basic statistical concepts that can represent data features and demonstrate certain functions is established. These concepts are frequently used in literature information statistics. Therefore, choosing important parts to introduce is necessary.

- (1) Statistical population and population units. Statistical population refers to the entire statistical research object. Each unit that constitutes the statistical population is a population unit. For example, the statistical research object is Chinese journal, the total number of all Chinese periodicals (species) is the statistical population, and each Chinese periodical is a unit of population. These two concepts are relative and vary with special statistical research.
- (2) An absolute number is also known as an absolute indicator. It is the direct statistics of a certain indicator, such as the total number of journals and papers or the amount of relevant circulation. It is the basis for calculating the relative number, the average number, and all other statistical indicators.
- (3) A relative number is the quotient obtained from dividing two indicators. It is typically expressed as a percentage. The most significant advantages of the relative index include facilitating comparative analysis and easily showing the mutual connection and development degree of a statistical object.
- (4) The average is the total of the same types of objects divided by the population. It has two types: static average index and dynamic average index. The average index eliminates chances to a certain extent and represents the general level of all data to facilitate the comparison of the overall number of levels; thus, it is considered a critical class of statistical index. In literature statistics, the commonly used average index includes the following:
 - ① Arithmetic average. It is the ratio of the total sign gross to the total number of units, and is usually expressed in X . It differs in simple arithmetic average and weighted arithmetic average.
 - ② Geometric average. It is the N -th root of n numerical product and is expressed in M_g .

$$M_g = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Geometric average can be used to calculate a certain ratio of the average development speed to static state. It can also calculate variable values into a geometric series relationship between the values of the variables or the average number of asymmetric numerical distributions.

- ③ Median. It refers to the numerical value of the intermediate position in a variable sequence. It is unaffected by the extreme values of variables. Occasionally, it is a representative of the actual level instead of the average. Its calculation formula is as follows:

$$M_d = \begin{cases} X_{\frac{n+1}{2}}, & n \text{ is odd number} \\ \left(X_{\frac{n}{2}} + X_{\frac{n+1}{2}} \right) / 2, & n \text{ is even number} \end{cases}$$

where X_1 , X_2 , and X_n represent the number of n in a sequence.

- ④ Mode. It refers to the numerical value that appears the most times in a variable sequence, usually expressed in \hat{x} . The calculation of the average of certain phenomena has no practical significance. The mode, as a representative value, must be adopted. In cases with few variables, the mode can most directly occur in the number appeared the most times. If an excessive number of variables exists, then the statistical frequency is grouped, and the median value of the highest frequency group is regarded as the mode.
- (5) Statistical index. The comparative index that explains the changes in similar phenomena is called the statistical index. This index belongs to the relative number of indicators. It can indicate the direction and extent of the changes of research objects and reflect their levels at various periods. Therefore, the calculation of the index has two periods: the base period and several periods of changes. The total copies of periodicals in a certain information reference room were 1000 in 2011, which was the base period. The index of the base period is generally 100%. In 2012, the number of copies increased to 1300; hence, 2012 was a period of change. Its index should be

$$\frac{1300}{1000} \times 100\% = 130\%.$$

The statistical index can be divided into a fixed base index and a chain index given the different base periods for comparison. When the index series is in the same period as the base period, it is called a fixed base index. When each index in the index series is adjacent to the previous period as the base period, it is called a chain index. The index of literature statistics can explain the growth rate of literature and other variation trends. Table 7.1 presents the relevant data of the amount of borrowed books in a certain library in 2012–2016. The fixed base index or the chain index can show the change and growth in the number of borrowed books.

- (6) Mark variable extent. Mark refers to the unit of population characteristic. It is categorized into quality mark and quantity mark. Mark variable extent refers to the degree of difference in mark values (variable values). It is also known as the discrete degree or eccentric occlusion degree. Mark variable extent can be used to measure and evaluate the average index of a representative. In general, when mark variable extent is small, the average number of representative is

Table 7.1 Statistical index of the quantity of borrowed books in a certain library in 2012–2016

Index	Years				
	2012	2013	2014	2015	2016
Borrowed books quantity(Thousands of copies)	7.4	8.5	12.7	18.6	25.3
Fixed base index	100%	$\frac{8.5}{7.4} = 115\%$	$\frac{12.7}{7.4} = 172\%$	$\frac{18.6}{7.4} = 251\%$	$\frac{25.3}{7.4} = 342\%$
Chain index	100%	$\frac{8.5}{7.4} = 115\%$	$\frac{12.7}{8.5} = 150\%$	$\frac{18.6}{12.7} = 147\%$	$\frac{25.3}{18.6} = 136\%$

considerable, and vice versa. The calculation methods and indexes of mark variable extent mainly include the following:

- ① Range. It refers to the difference between the maximum and minimum values in a variable sequence.
- ② Mean difference. It is the average deviation in the variable sequence of each variable value and its arithmetic average or median. Its calculation formula is as follows:

$$\text{Mean Difference} = \frac{\sum |(x - \bar{x})|}{n}.$$

- ③ Standard deviation. It is the square root of the arithmetic mean of each variable value and its arithmetic average in the variable sequence. Its calculation formula is as follows:

$$\text{Standard Deviation } \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

- ④ Mark variable coefficient (dispersion coefficient). It is the ratio of the variable sequence of the standard deviation and its arithmetic average. It can truly reflect the degree of mark variation in different sequence levels. Its calculation formula is as follows:

$$V = \frac{\sigma}{\bar{x}}.$$

7.2 Principles and Indexes of Literature Information Statistics

7.2.1 Principle Requirement for Literature Information Statistics

When performing various literature information statistics, we must follow several basic principles, which have specific common requirements for data. This process is a focal point in considering and dealing with problems in literature information statistics. To achieve certain data standards and statistical purposes, we must have clear statistical subject range, time interval, literature types, and statistical tools (e.g., abstract, index). The principal requirements for literature information statistics and its data can be summarized as follows.

- (1) Pertinence. Literature information statistics is the work statistics of library and information units. Some of these works are monographic studies of the statistics for a certain purpose. Different specific statistical items have varying requirements, such as the content of statistics, the setting of indicators, and the methods for processing data. They should be targeted, not stereotyped.
- (2) Accuracy. The accuracy of data and results is the core requirement in literature information statistics. Only when literature data are accurate can a reliable analysis conclusion be obtained. In statistics, numerous works are used for the set goal, which is clearly defined as statistical bounds, the meaning of indicators, the representative of statistical tools, and the extensive data sources. All these works guarantee the accuracy of the data, thereby making them closer to reality.
- (3) Representative. In literature information statistics, some statistics are comprehensive total sample statistics. However, most are sampling statistics. Therefore, the question of whether a sample can represent the entire data is raised. Only a strong representation can ensure the reliability of statistical results and the credibility of an analysis conclusion. For example, when selecting a tool to study the literature of a certain subject, its representation must be considered.
- (4) Comparable. The statistical data of literature information are the bases for analysis and judgment; the comparison method is frequently used for analysis. Therefore, when setting the literature index and the statistical data, we must pay attention to the comparability of data. When facilitating a comparative analysis, convincing conclusions are drawn.
- (5) Cumulative. Cumulative data can be an absolute number or a relative number. For example, $R(n)$ in the study of Bradford's law refers to the cumulative number of related papers in the N species of journals. In determining core journals, cumulative percentage methods are used. When the cumulative time of literature data is long, the cumulative amount is considerable, and the accuracy of the literature statistics can be improved conductively.

These principles and requirements are interrelated and complementary. In literature information statistics, they must be integrated effectively to achieve better comprehensive results.

7.2.2 *Index System of Literature Information Statistics*

The statistical index refers to the numerical concepts and concrete numerical values that reflect overall phenomenon. It generally includes two parts: index name and index value. The statistical index system is formed by a series of interrelated statistical indicators. In literature statistics, these indicators include the number of journals, relevant papers and authors, book circulation and utilization, and reading rate. These statistical indicators are commonly used and their combinations can form a relatively complete index system. The index is an important means of implementing the statistical function and is the basis of all statistical research. Literature statistical research is mainly achieved through statistical indicators. The literature statistical index can reflect the scale, level, distribution proportion, and growth rate of literature information activities; thus, it can identify the characteristics and laws of literature.

The scientific literature system is a complex and evolving collection that does not only has an inherent connection among various characteristics, but also implements mutual restriction with many external factors. Each type of statistical index can only reflect a certain aspect of documentation and information. To describe comprehensiveness from different angles and to study inherent laws by comparing indexes, we must establish a series of statistical indicators to constitute a scientific index system. For a special research project on literature information, the setting of the statistical index can be considered from the actual requirements and flexible control. However, the connection and coordination among indicators must also be considered. For regular literature statistics in library and information institutions, a set of index systems that reflects the characteristics of various literature and the law of literature activities must be designed. The setting, meaning, scope, and calculation method of the literature statistical index must be a unified regulation that strives to achieve standardization. This aspect is one of the most important in literature statistics. It vigorously improve

(1) Types of statistical indicators. Different types of statistical indicators are available according to various classification methods.

- ① Statistical indicators can be divided according to the objects being described into collection indexes, borrowing indexes, citation indexes, author indexes, circulation indexes, and reader indexes. Collection indexes cover quantity, variety, quality, price of books, literature, and information. Borrowing indexes cover the quantity of all types of literature and the relationship between borrowers and readers. Citation indexes cover the quantity, source, and various distributions of citations. Reader indexes

cover the number and composition of readers. These indexes are the main content of literature statistics.

- ② On the basis of data form, statistical indicators have absolute and relative number indexes. An absolute index refers to the number of simple data, whereas a relative number index refers to the amount of proportional data.
 - ③ Statistical indicators have work evaluation indexes and work control indexes. Work evaluation indexes are used to evaluate the work conditions of systems and to compare among systems. Book and periodical circulation indexes reflect the lending workloads of libraries and the information institutions. When the amount of circulation is high, the workload of circulation is also high. Circulation rate indexes reflect the utilization situation of books and periodicals. When the circulation rate is high, the utilization situation is good. Work control indexes are used by managers to adjust and control systems. For example, the proportion of all types of library collection reflects the composition of a library collection. Therefore, managers can accordingly develop reasonable literature procurement policies and regulations for the collection structure. Certain indicators can be used for evaluation and control, such as the lend-denry rate, which has always been regarded as an evaluation index. At present, the queuing theory proves that the lend-denry rate is a good index for controlling collection quality.
- (2) Design of statistical indicators. In literature statistics, we must pay attention to the following points when setting up statistical indicators.
- ① Design principle of statistical indicators. The principal requirements of literature information statistics are also followed when setting indexes. Simultaneously, obtaining accurate and simple data should be considered as beneficial.
 - ② Concept of indicators must be clear. The actual definition and scope of statistical indicators should be clearly defined to obtain a unified standard. For example, collection quantity indicators, such as old and donated books, are not considered part of book circulation because of certain reasons. Hence, clear and reasonable rules must be established.
 - ③ Calculation method for indicators must be simple. A scientific and simple calculation method must be used to improve the quality of statistical data and the efficiency of statistical analysis.
 - ④ Index units must be clear. The quantity performance function of statistical indicators is achieved through certain measurement units. The units for all types of indicators should be clearly defined and unified before and after statistics are gathered. For example, the unit for the number of journals must be defined as “kind” or “volume.” The amount, of species and the volume of collections should be considered to fully reflect the collection

situation and the duplicate amount. Meanwhile, the measurement unit for statistical indicators aims at specific ranges and conditions.
s the efficiency of documentation and information quantitative research.

7.2.3 Statistical Indicators of Information Resource Management

In long-term library and information works, scholars have proposed various statistical indicators of literature information and document works. Statistical indicators quantitatively reflect the situation of library and information works from a different perspective. Most indicators are the common concepts of an index in bibliometrics. Some of these indicators are discussed in relevant chapters. In this chapter, we only introduce common statistical indicators of literature information.

Circulation indexes:

Circulation quantity typically refers to the total loaned documents within a certain time. Broadly speaking, however, any information exchange through library and information institutions can be called “circulation,” and its measurement is called the amount of circulation. At present, the amount of literature volumes is used as its measurement unit.

Circulation rate is a relative indicator that is generally defined as $p = \frac{N}{M} \times 100\%$, where N is the literature quantity of borrowing within a statistical period and M is the total number of literature in circulation.

Circulation speed refers to the number of circulation in unit time. It is the unit for volume/time.

These indicators can reflect the size of the circulation workload, as well as the speed and utilization efficiency of document delivery.

(1) Collection use efficiency can be expressed as

$$S = (\text{volume of annual lending}/\text{total number of library books}) \times 100\%.$$

(2) Book turnover rate refers to the average turnover time per book in a year; it can be expressed as

$$B = (\text{total number of times of annual book lending}/\text{volume of library books}) \times 100\%.$$

(3) Reader lending rate is the average number of readers who are borrowing documents; it can be expressed as

$$R = \frac{\text{The total number of lending documents}}{\text{The number of lending readers}} \times 100\%.$$

(4) Rejection rate is generally defined as

$$H = \frac{\text{The number of readers not borrowing literature}}{\text{The total number readers borrowing literature}} \times 100\%.$$

(5) Time difference coefficient is defined as

$$k = \frac{\text{The number of abstracts of current literature}}{\text{The total number of abstracts of this Year}}$$

Evidently, when the value of K is high, the time difference is less, and the reporting rate is high. The resulting values can be used as bases for secondary literature evaluation and purchase.

(6) Information absorption coefficient is generally defined as $I = \frac{N}{M}$, where M is the total number of documents published within the statistical span and N is the number of used documents. This coefficient can also be relative to the literature subject or the type of publication. It is used to measure the degree of information to be absorbed and used by society.

7.3 Types and Basic Steps of Literature Information Statistics

7.3.1 Main Types of Literature Information Statistics

The objects of literature information statistics are the literature and all the feature indicators related to the literature, such as readers, technical terms, books, periodical acquisition costs, and length of bookshelves. These relevant features of the literature can indirectly reflect the changes and utilization of the literature. Therefore, these features also belong to the scope of literature statistics. Literature statistics has formed a number of corresponding types because of the different objects, targets, or purposes of statistics. For example, overall statistics and sampling statistics are based on methods and scopes. Business statistics and special research statistics are based on the nature of literature. Direct and indirect statistics are in accordance with the manner of literature. In general, the types of literature statistics and data are grouped according to the objects and contents of statistics. They mainly consist of the following.

(1) Publication statistics. Publication is a broad concept that includes books, periodicals, scientific reports, patent literature, and other types of literature. It is the main object of literature statistics. The statistical analysis of publications is an important content and approach of bibliometrics. As early as 1917,

philologist F.T. Cole and N.B. Eales conducted statistical analyses of anatomy publications and proposed the initial measurement for the field of literature and information. Recently, a large number of bibliographic statistics also belong to this type of literature. In the field of bibliometrics, publication statistics and the most basic types of literature statistics are widely used. The following circumstances can be observed:

- ① Statistics on the amount of books, periodicals, and other literature according to country;
 - ② Statistics on the circulation of books and periodicals of each publisher (business);
 - ③ Statistics on the amount of books, periodicals, and other literature according to subject;
 - ④ Statistics on the amount of books, periodicals, and other literature according to language;
 - ⑤ Statistics on the amount of certain literature subjects published in a scientific journal;
 - ⑥ Statistics on the amount of monographs, papers, and patents published by a certain society and scientific research institution;
 - ⑦ Statistics on the amount of monographs, papers, and patents published by a certain author;
 - ⑧ Statistics on the number of publications in a certain region or period.
- (2) Author statistics. Statistics on the papers of authors are conducive to grasping the development level of science and technology. Simultaneously, they also provide data for studies on sciences. Statistics on the number of authors exhibit the following situations.
- ① Statistics on the amount of outstanding authors. Within a certain period, authors cited with high frequency were listed as outstanding authors. Some organizations abroad also produced an outstanding author index, which provided a basis for evaluating literature quality and studying talents in science.
 - ② Statistics on author ratio for each discipline or professional literature were gathered to analyze the peak research and development trends and provide the bases for studying science of science and futurology.
 - ③ Statistics on the authors of the literature for each discipline in different countries were gathered to analyze and compare the focus, development status, and trend of the disciplinary study of these countries.
- (3) Statistics of scientific and technical terms. The composition and quantity changes of technical terms reflect the mutual osmosis, as well as the rise and fall, of science and technology to a certain extent. Therefore, gathering the statistics of and analyzing scientific and technical terms are research methods in bibliometrics. In 1963, Price counted annual changes in technical terms. Tagliacozzo analyzed technical level through the statistics of technical terms.

- (4) Citation statistics refer to the number of cited literature attached to scientific papers or books or the use of citation index tools to count the number of related citations. Citation analysis based on citation data is an essential and commonly used quantitative research method. In early 1927, Gross conducted citation analysis in the history of philology. Recently, papers on bibliometric research are classified under citation analysis.
- (5) Statistics on the use of other relevant literature information mainly refers to counting the works of library and information institutions, including the collection statistics, circulation statistics, reader statistics, and quantity statistics of other documentation and information service items.

7.3.2 Basic Steps of Literature Information Statistical Analysis

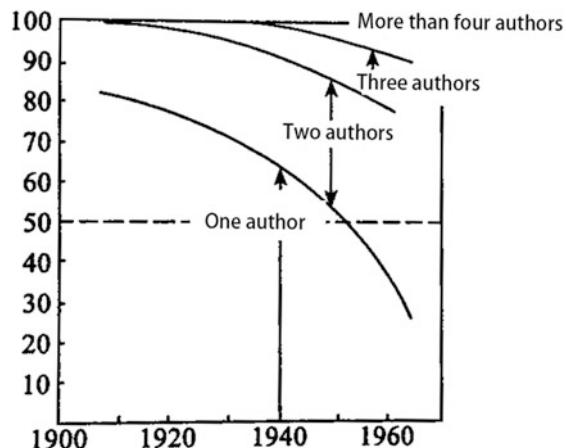
In literature information statistics, although the types of literature statistics differ, some basic steps are the same, which are discussed as follows.

- (1) Statistical survey refers to the collection of statistics, original data, and first-hand information of the research object. When acquiring raw data, the following points should be considered.
 - ① Determining the specific method. When the range of literature statistics is small, the comprehensive statistical investigation method is required, i.e., the statistics of all the objects should be collected. Therefore, the representative of the data is good and reliability is high. When the range is large-scale and covers a long period, part statistical methods are generally used, such as typical survey statistics, which focus on random sampling statistics. For statistics on different research purposes, requirements, and conditions, we can use direct statistics or indirect statistics. Direct statistics count various types of publications or directly count literature and citation quantity from original journals. Indirect statistics count different types of documents using retrieval tools, such as abstracts, catalogues, and indexes, or citation analysis tools.
 - ② Choosing statistical tools. When the statistics of the literature of a certain subject must be collected, we generally use retrieval tools (e.g., abstract, index, catalogue) as statistical objects. For example, the abstract method is appropriate to determine the core journals of analytical chemistry. In a particular case, we used the British “Analytical Abstracts” because its digest entry was focused and comprehensive compared with that of the American “CA,” which was incomplete, scattered, and difficult to count. Therefore, choosing an appropriate comprehensive and representative data source tool is one of the most important steps in literature statistical work.

- ③ Clearing index content. We should determine data indicators based on the requirements of statistical research. The content, scope, and unit of literature statistical indexes should be clear and relatively simple to easily obtain data.
 - ④ Statistical original data. Data should be accurate and reliable to avoid errors.
- (2) Statistical sorting is also called statistical summary, which involves processing and classifying original data. The content of the work includes the following.
- ① Calculation. We can calculate the average, cumulative size, percentage, and logarithm values based on the original data to expand the data items and provide adequate data for statistical analysis.
 - ② Sorting. We can sort data according to certain requirements and rules, such as ranking for Bradford-type data.
 - ③ Expression. Literature statistical data can be expressed in the form of statistical tables or charts to illustrate the relationship among statistical indicators. Statistical data results are expressed in a statistical table, which is eye-catching and clear, whereas a statistical image is more intuitive.
- (3) Statistical analysis is a key step in the entire literature statistical analysis method. It includes the conclusion analysis and error analysis of statistical data. Conclusion analysis determines the corresponding regularity conclusions from statistical data. Error analysis determines the accuracy and precision of the results. In general, the following methods can draw conclusions from the data.
- ① Analysis under theoretical guidance. We can analyze statistical data and draw some conclusions under the guidance of theories and laws of bibliometrics. For example, we can determine whether a set of statistical data is consistent with the distribution of Bradford by using Bradford's region analysis method.
 - ② Numerical analysis. Every object has an absolute value of certain aspects and the relative value of its share in the population. These values are important bases for the direct analysis of the characteristics of objects and their relationship with other objects, such as the comparison of objects during different periods (e.g., months, days), the comparison of many objects during the same period, and the comparison of many objects over the years. These are commonly used methods for literature statistical analysis.
 - ③ Image analysis. In bibliometrics, images are frequently used to describe statistical data. The images are then analyzed, and corresponding conclusions are drawn. Image analysis uses two methods: image function value analysis and slope analysis.

Image function value analysis. For example, we analyzed the changes in the number of authors in a paper and drew the appropriate conclusions. First, we counted the number of authors of each chemistry and chemical engineering literature in the American "CA" and obtained the original data. Second, we sorted

Fig. 7.1 Profile of authors of chemical engineering literature



through the original data and calculated the percentage of papers with one, two, three, or more than three authors annually per paper. Then, the relative number of authors was placed on the vertical axis, whereas age was placed on the horizontal axis to conduct image description (Fig. 7.1). Finally, we analyzed the result. As shown in the image, the relative number of literature with one author is decreasing annually, the relative number of literature with two authors is increasing annually, and the relative number of literature with three and more than three authors exhibits the most notable growth. The scientific results are reflected primarily through literature, and thus, the enhancement of coauthors in literature can reflect the strengthening of the collective nature of scientific labor. From the important premise and trend of the image, we can draw the conclusion that in the modern scientific era, the collectivization trend of scientific labor is significantly enhanced.

Image slope analysis. In mathematical analysis, the slope of an image is generally regarded as a geometric measure of the change rate of the curve. Therefore, the slope in image analysis is frequently the important basis for the research development speed and trend of objects. In general, when the slope of the image is considerable, the development speed of the research object is fast. Extrapolation in information analysis is an analytical method based on the aforementioned ideas.

7.4 Application of Literature Information Statistical Analysis

The application of the literature information statistical analysis method is extensive. It can provide statistical data and can be combined with other analysis methods. It can also be used on its own as a complete method for bibliometric research. We use the literature information statistical analysis method not only to determine the

quantitative change laws of literature, identify core journals, formulate literature collection and management strategies, and determine the size of retrieval tools and principle of selecting words, but also to conduct a series of research in the science of science and in forecast science. This section of the paper discusses only the main aspects.

7.4.1 Application to Information Resource Management

The scientific management of library and information must be based on literature information data and quantitative analysis. Through the literature statistical analysis method, we can obtain literature data, such as the amounts of collection, circulation, citation, annual growth of literature, and readers. These literature data and their changes reflect the actual level of literature work. They measure the scale of efficiency of literature work. They are also important bases for improving library and information management. Accordingly, we can formulate rational literature work policies, check the situations of various professional works, evaluate and adjust the progress of work in various departments, and forecast the prospect of library and information. In addition, we can also analyze the operating status of a certain library and information institution as a system. For example, we determine:

- ① whether purchasing books and periodicals is appropriate, whether it can basically satisfy the needs of readers, and whether cataloging a new book is rapid and timely;
- ② whether the job set is reasonable, whether the workload is appropriate, and whether the labor division of librarians is consistent with the principles of best use;
- ③ the quality of card making, card arraying, book shelving, and circulation management;
- ④ the effectiveness of directory usage and information consulting service;
- ⑤ reader analysis, including the analysis of reader composition, information needs, literature utilization, and core readers;
- ⑥ library automation effect analysis;
- ⑦ administrative management agency, particularly the analysis of the management level with curator optimal allocation and the use the entire labor force and equipment.

7.4.2 Application to Information Users and Literature Information Utilization Research

The literature information statistical analysis method is one of the basic methods for the research on information users (or readers). At present, reader statistics are extensively collected in libraries, and information centers have also gathered similar user statistics. Statistics are mainly based on data from the library cards of borrowers (or borrowing records from computers) and library registers. User statistical data are important materials for information user research. These statistics and data reflect the composition, quantity, and changes of users, as well as indicate the relationship between number of users and number of documents. We can determine the condition of users in information centers based on user (reader) statistics and simple calculations, and thus, we can conduct analysis and research on information users.

Analysis and research on the utilization and law of literature require the literature statistical analysis method. This method for literature utilization research mainly uses two aspects of statistical data: the literature data of the utilized situation and the citation data from library and information centers. Library and information centers provide literature information data to users through borrowing, reading, copying, translating, and solution consulting. The utilization statistics of literature are based on various service records, including the number of used literature data, literature disciplines, kinds of record, and types and distribution of the number of users utilizing literature. Statistics show that the number of journals that circulate simultaneously approximately constitute geometric progression, which is the formation of Bradford's distribution. We can establish each type of periodical circulation that affects the curve through monthly statistics based on the distribution of periodicals to predict the flow requirements for the following year, guide the collection of literature, and establish a reasonable collection. Therefore, we can determine the utilization rate of different documents and provide scientific basis for library and information management through literature statistical analysis.

7.4.3 Application to Literature Information Law Research

The literature information statistical analysis method has been widely applied to the study of literature laws, such as the laws of literature increase, aging, and distribution. The literature utilization and citation statistics of library and information institutions reflect the utilization of various types of literature in different disciplines and periods. We can depict the literature aging curve of different disciplines (types or languages) according to the annual statistical data of utilization literature. This curve indicates the laws of literature aging. In research on literature laws, the literature information statistical analysis method is frequently used with other methods.

7.4.4 Application to Discipline Development Law Research

The literature information statistical analysis method has been widely used in research on forecast science, science of science, and science and technology management. For example, through the statistical analysis of relevant literature, we can evaluate and predict the level of research and development characteristics, as well as the change trends of a certain subject or technical field. Moreover, we can determine the attractiveness of a certain subject and the accuracy of some types of topic selection. Furthermore, we can analyze the research achievements of a research staff or institution and the effect of a certain method.

We can analyze the development and trend of a certain discipline through the statistics of its literature and change situation. This approach is one of the basic methods for bibliometrics. The following graph can be obtained based on the statistics of the literature for a certain subject each year:

As shown in Fig. 7.2, (a) indicates that the subject is in the birth stage; (b) indicates that the subject has been formed and is in the rapid development period; (c) indicates that the subject will mature; (d) indicates that the subject has fully entered a mature stage, the amount of literature has reached a saturation state, and the curve is in the stationary section; (e) indicates that the subject may be a new branch of discipline or has completely matured and exhibits a fading trend. This graph intuitively shows the birth, development, maturation, and differentiation (or decline) of the entire historical process of a discipline, thereby providing a quantitative basis for the study of the characteristics and laws of subject development.

An example is provided to further illustrate the basic steps and application of the literature information statistical analysis method.

We analyze the development of information science research in China through literature quantity statistics.

- (1) Statistical survey: The researchers gathered literature statistics of information science in China. They chose 115 kinds of journals as statistical objects from 1956 to 1980, designed three statistical indicators (year, number of literature, and literature category or field), and gathered statistics of every journal.
- (2) Statistical processing: We cleared up by sub-entry according to the requirements based on the original data statistics.
 - ① Calculation: The cumulative total was 2800 articles. We calculated these articles according to three stages and the literature categories.
 - ② Arrangement: The literature data were arranged according to different years, the three stages, and the categories.
 - ③ Representation: The year and absolute amount of corresponding literature (articles) were mapped (Fig. 7.3) and listed in the three stages according to the literature categories (Table 7.2).
- (3) Statistical analysis: From the aforementioned statistical data of information science literature, we can observe the occurrence and development of information science and its future trend in China. An important premise is that

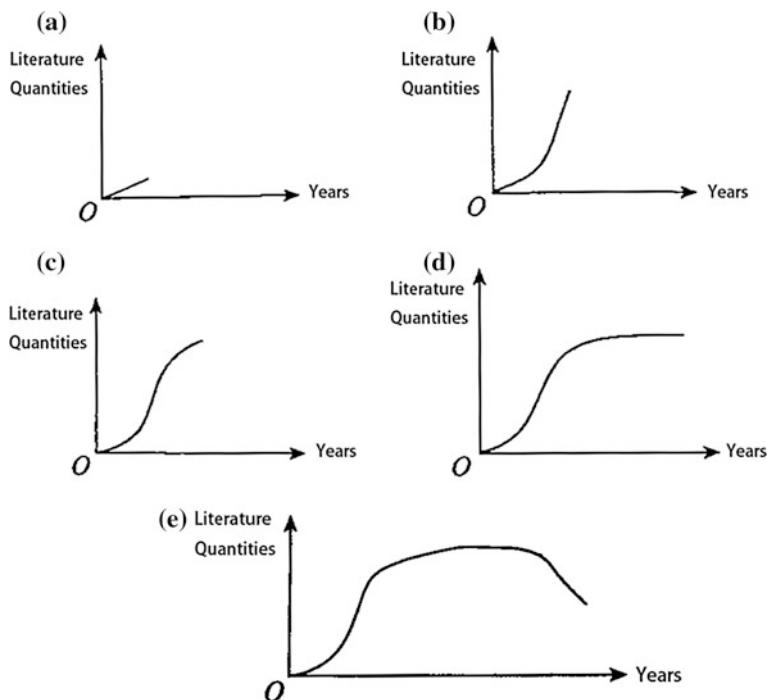
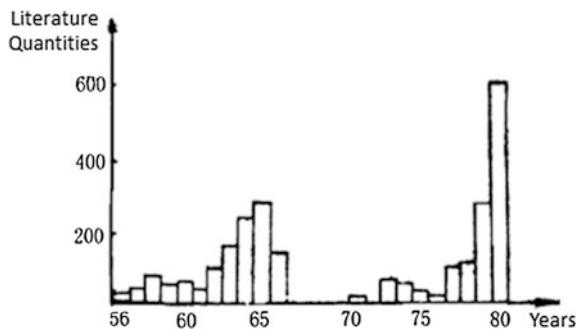


Fig. 7.2 Time distribution curve of the literature of a certain subject

Fig. 7.3 Distribution of the amount of literature in information science in China



literature growth is a significant indicator of subject development. From this premise and the documentation statistical data, we can draw the conclusion that the development of information science in China has occurred in three stages.

The first stage (1956–1966) is the generation and preliminary development stage of information science.

The second stage (1967–1976) is the stagnation stage.

The third stage (after 1977) is the comprehensive development and booming period of information science research. During this period, the number of literature

Table 7.2 Analysis of the amount of literature in information science (1956–1980)

Literature quantity (piece) Category	Stage			
	First stage 56–66	Second stage 67–76	Third stage 77–80	Subtotal
Information science theory	34	5	187	226
Work system and organization	114	4	162	280
Information data collection	155	75	165	395
Editing and publishing of information data	104	3	72	179
Information research	25		82	107
Information retrieval	243	39	235	517
Information service	131	20	35	186
China information science	209	15	77	301
Information science in the world	168	10	128	306
Information network	107	38	138	283
Total	1290	209	1281	2780

did not only rise sharply, but information science also infiltrated numerous areas of social sciences and natural sciences. A total of 99 journals published information science literature, including 29 professional journals, 32 related discipline journals, 22 natural science journals, and 16 social science journals. These figures showed that information science research exhibited large-scale and fast-paced development during this period.

7.5 Mathematical Statistical Method and Its Application

7.5.1 *Summary of Mathematical Statistical Method*

In general, traditional statistics is a method that uses figures to represent the real features of things. Scholars have emphasized the comprehensiveness and accuracy of the data obtained through this method. However, gathering the statistics of the whole population is difficult in most practices due to various objective reasons. Even when we collect statistics from the entire population, the acquired data will still be incomplete and inaccurate. Under this condition, researchers generally do not and cannot fully observe the entire population, but only select a random sample from the population. Then, certain methods are used to obtain information for analysis, form an opinion, and make overall conjectural judgments. Figure 7.4 illustrates the process and explains the relationship between a statistical population and a sample. The sample is an integral part of the total population according to the principle of dialectical materialism; hence, the characteristics of a partial sample reflect the overall characteristics to a certain extent. With this idea and given the

required objectives of statistical practices and the development of statistics, a new branch of mathematics, namely, mathematical statistics, was formed.

The mathematical statistical method is a new statistical method that adopts probability theory as its theoretical basis. This method is based on sample data obtained through experiment or observation, thereby providing a reasonable estimation and judgment for the overall objective laws of the study. The content of mathematical statistics is rich and includes random sample, sampling distribution, parameter estimation, hypothesis testing, regression analysis, and variance analysis. In summary, the study of mathematical statistics has two main aspects. The first aspect includes the process of extracting the sample from the population, the amount of sample, and the method of drawing the problem, i.e., the problem of sampling methodologies. The second aspect includes the process of sampling the results (sample data) for a reasonable analysis and making a scientific inference, which is the problem in statistical inference. Mathematical statistics are generated and developed to a certain extent because researchers cannot deal with too much data. This new method infers the whole from the sample and is suitable for dealing with random statistical objects.

7.5.2 Applications of the Mathematical Statistical Method

The **application of the mathematical statistical method** is common because obtaining sample data is easier than obtaining overall data and analyzing local data is simpler than analyzing overall data. In literature and information work, the application of the mathematical statistical method is not only necessary, but also possible. The applications and contents of this method include the following two main aspects.

- (1) Sampling statistics. In the statistics of the library and information fields, the sampling statistics method is frequently used to reduce statistical work. For example, to collect statistics on the annual readers of a library, the application of the mathematical statistical method is a relatively simple approach. The year is divided into a number of time units (such as weeks, days, or hours), which are drawn from a number of units and the statistics of the number of readers of the library. The average value is calculated and then multiplied by the total units of time. The result is the number of readers of the library for the entire

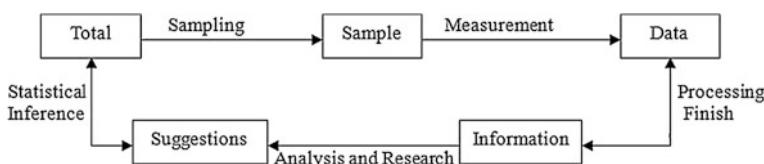


Fig. 7.4 The process of mathematical statistics

year. This simple method is a process of inferring the total from the sample. The theoretical basis and reliability of this method can be solved via theory of mathematical statistics.

- (2) Statistical prediction. The overall statistical data for future situations are impossible to obtain. We can infer a future situation using statistical data from past situations. This scenario is one of the statistical forecasting problems. For example, a new public library is created in a certain place. To determine the reasonable proportion of the scale for the arts and science reading rooms, we need to make a corresponding estimate based on the number of readers and the amount of data. The data reflect only the past situation; hence, inferences about the future can only be made based on the past sample. To guarantee reliability, we must use theories and methods of mathematical statistics.

Meanwhile, statistical objects in the fields of library and information science are mostly random events. The mathematical statistical method exhibits a unique advantage in dealing with such statistical objects and can provide effective and appropriate tools to solve measurement problems in the fields of library and information. Consequently, the mathematical statistical method plays a significant role in the fields of library and information.

The mathematical statistical method in the fields of library and information, which is considerably beyond the scope of the work of library and information units, is also widely used in the study of library and information science. For example, the x^2 test is used in bibliometrics to determine whether a literature measurement formula fits into the literature data of a discipline. The regression analysis method can determine the coefficients of certain formulas. The regression function model is also one of the commonly used mathematical models. In information research, regression analysis can be used to forecast information and the x^2 test can analyze the stability of the Delphi method. In the field of information retrieval, the sign test can be used to infer the literature sample under two types of search conditions to determine significant differences. In informetrics, the mathematical statistical method has important applications, such as a sampling method and for statistical inference, parameter estimation, hypothesis testing, and regression analysis. An example to illustrate the application of mathematical statistics is provided as follows.

Example: The average accuracy rate of the information retrieval system of a company (that was about to be sold) was 0.64 and the standard deviation (σ) was 0.0613269. A user who wanted to buy the system retrieval experiment applied 10 different questions. The results of the proportion of relevant literature were 0.69, 0.58, 0.62, 0.49, 0.59, 0.58, 0.62, 0.68, 0.61, and 0.71. From these data, the user could determine the measurement results of the company.

For this process, we can apply the mathematical statistical method through the following steps.

- ① The null hypothesis, $H_0 : \mu = 0.64$, is established.
 $H_1 : \mu \neq 0.64$

- ② The significance level, $\alpha = 0.05$, is determined. Look-up table:
 $|Z_{\alpha/2}| = 1.96$.
- ③ For the sample data, $\bar{X} = 0.617$ is calculated.
- ④ The calculation test statistics is $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{0.617-0.64}{0.0613269/\sqrt{10}} = -1.186$.
- ⑤ $|Z| < |Z_{\alpha/2}|$; therefore, the null hypothesis is accepted. That is, at a significance level of 0.05, the user can accept the measurement results of the company.

Chapter 8

Methods of Citation Analysis

The distribution of document information citation exhibits certain regularity, which is an important part of information theory of measurement. Moreover, a citation analysis method has a wide range of applications. Therefore, the discipline of literature information and citation analysis occupies a pivotal position and has important theoretical and practical effects.

8.1 Basic Concepts and Methods of Citation Analysis

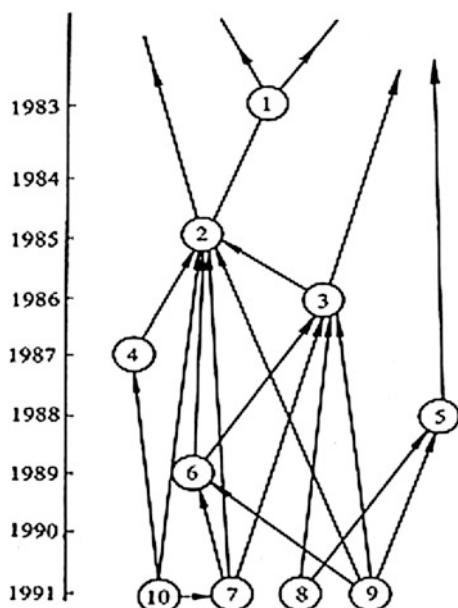
8.1.1 Basic Concepts of Citation Analysis

(1) References and citations

Scientific literature is not isolated, but is interrelated with one another. The relationship among scientific literature determines mutually cited documents. In general, an article or a book is required to refer to other relevant documents during preparation. A complete scientific paper consists of two parts: the body part and the references section. The references reflect the situations that authors absorb or take advantage of as ideas, methods, and techniques.

First, the distinction between references and citation must be established. Each cited reference is a reference for the citation provided by the authors. For example, a paper cites an article of Price, which has a reference; hence, the paper is a citation for Price. Therefore, a citation has varying meanings in the same document for different objects. The arrow that points to the rays cited in a literature and nocks a point to a citation can clearly demonstrate that the exchange of scientific literature links horizontal and vertical inheritance. The exchange situation is the mutual citation relationship structure called the citation network (Fig. 8.1). The citation network contains a wealth of useful information about document exchange, subject

Fig. 8.1 Citation network diagram



relation, and scientific development. It can trace back the history of scientific development, as well as the scale and trend of scientific development, through the statistics and analysis results of information data.

(2) Citation analysis

Authors frequently use a footnote or an endnote that lists the references or citations in published scientific papers or books. Such lists form the relationship between citation and cited of scientific literature. Citation literature is one of the basic natures of scientific literature. The citation relationship of scientific literature is the main basis of citation analysis.

Citation analysis uses various means, including mathematical, statistical, comparison, induction, abstraction, generalization, and logical methods. These methods are used to analyze a variety of scientific journals, papers, objects of citation, and cited phenomena to determine the characteristics of a quantity method and the inherent law of a bibliometric analysis method.

Citation analysis was introduced in the 1920s. In 1927, Gross et al. conducted the first citation analysis in the history of literature. They analyzed the references of articles in several chemical engineering periodicals and core periodicals in chemistry education.

An increasing number of papers about citation analysis have been presented in the field of bibliometrics. For example, Line and Sandison proposed literature obsolescence. Buckland discussed literature obsolescence and literature scattering, which were based on citation statistics. As a result of various disciplines, regions, and periods of scientific papers, a cited phenomenon frequently has its own

characteristics and laws. For example, we can infer a paper, a scientist, or the role of a scientific paper in the process of scientific development. Moreover, the connection and distinction among scientific papers, scientists and scientific journals, science and scientific disciplines can also be inferred. Therefore, the citation analysis method has extensive applications. This method can be effectively applied to many areas by practice, and it plays an increasingly important role.

8.1.2 *Citation Behavior and Motives*

The citation and cited scientific literature are manifestations of the scientific development of law, which embodies the scientific knowledge and intelligence content of accumulated and continuous inheritance. They also embody the principle of the unity of science and cross-penetration among multiple disciplines. Therefore, the records of scientific knowledge and scientific research directly infer that scientific literature cannot be isolated but is interrelated. An author of scientific literature inevitably cites the literature of other authors in writing scientific papers to learn from their experiences and results. Therefore, the citing behavior of scientific workers is a widespread phenomenon and is an indispensable part of scientific communication.

The citation analysis method has been applied extensively and has attracted widespread attention and in-depth research. Scholars have discussed this method in many aspects for a variety of cited motivations. The author of scientific literature generally does not cite literature that is completely unrelated to his/her topic. Various reasons are available for mutually cited literature. M. Weinstock pointed out the following 15 reasons for literature citation after systematic induction:

- (1) To respect the blazers
- (2) To demonstrate positive attitude about related works
- (3) To verify the methods and instruments used
- (4) To provide background materials
- (5) To correct their own work
- (6) To correct the work of others
- (7) To evaluate previous works
- (8) To seek full argument for their arguments
- (9) To provide existing works
- (10) To provide a guide for not being spread, rarely cited, or never cited literature
- (11) To validate data and physical constants
- (12) To check whether a particular idea or concept is discussed in the raw material
- (13) To check the cause of the characters in raw materials or other works of a concept or a noun
- (14) To deny the work or viewpoint of another
- (15) To object to the priority of others

These 15 citation motivations are normal for the scientific development of literature. Citation behavior constitutes a part of the scientific knowledge processes.

This behavior is generated by the citation and the cited literature from all angles and levels to reflect the scientific development of the current situation and the law.

Mutually cited literature is determined by the laws of the development of science itself and research activities. Scientific studies repeatedly show that scientific knowledge has significant cumulative inheritance. Scientific knowledge is derived or differentiated based on the original subject or any new subject or technology, i.e., the original subject or the development of technology. Therefore, the development of science and technology is continuous. Meanwhile, existing disciplines contact each other, intersect, and perform mutual penetration because of the unity of science principle. Therefore, any scientific study must be based on the achievements of predecessors, thereby absorbing the experiences of others. Accordingly, as a result of scientific knowledge and scientific research, scientific literature must also be interrelated. Authors inevitably cite other relevant literature to prove their own point of view and find bases for data when writing scientific papers. In published scientific works, authors list cited references to illustrate the origin of the citation data, to emphasize their reliability, to facilitate search, and to check for readers based on further research. A scientific works also show that an author pays attention to scientific morals and respect the labor of others. Thus, when scientific literature cites one another, the law of development is implemented, which is an inevitable phenomenon in scientific activities.

However, some citations are contrary to the purposes of motivation and behavior. F.C. Thome summarized these citations as follows:

- (1) To flatter someone for citing
- (2) Biased citation for mutual praising
- (3) To quote himself/herself
- (4) To support certain academic research interests of an improper citation
- (5) Under the pressure of the cited authority

Citation behavior cannot reflect the actual process of scientific development and communication; it can only cause confusion and contamination. Many experts and scholars have questioned the objectivity and accuracy of citation analysis. Therefore, we should study the mechanism of citation analysis and continuously improve the mathematical model for citation data processing to accurately reflect the essence of scientific development law.

8.1.3 Basic Types and Steps of Citation Analysis

- (1) Basic types of citation analysis

Technical methods are constantly enriched and perfected with the extensive application of citation analysis for decades. From the current situation, the three basic types of citation analysis are follows:

- ① Research on the number of citations. This type is mainly used for evaluating journals and papers.
- ② Study of the citation relationship or chain relationship. A citation network relationship exists among scientific papers, such as A by B, B by C, and C by A. The study of this relationship is mainly used to determine development and contact and to project future prospects.
- ③ Reflection of the theme of correlation research from the aspects of citation, which is mainly used to determine the structure of science and literature retrieval.

Many types of citation analysis can be derived when starting from the different characteristics of citations.. For example, citation analysis is based on language, country, age, and author.

(2) Basic steps of citation analysis

The basic steps of citation analysis are described as follows:

1) Selecting an object to gather statistics

We can choose the subject representative in an authoritative magazine according to the specific circumstances of the subject of study. Then, we can determine the number of phases and several related papers for statistics.

2) Citation data

In selecting papers, the number of citations is classified and counted, including publication age, language, type, and the quotation of the amount of self-citation of the author. Statistical projects can be identified according to the research purpose and requirements, such as concreteness and flexibility, and can be determined on their own. The relevant citation data are selected directly from SCI and used as bases for citation analysis.

3) Citation analysis

To acquire citation data, an analysis is performed based on various indicators or other perspectives according to the purpose of the study. Purposes include the theoretical analysis of citations, the concentration and discrete trend analysis of citation quantity, the citation quantity law with time of growth, and the major indexes of citation analysis, such as self-citation quantity, citation language, literature type, time, and country.

4) Conclusion

A corresponding analysis is performed according to the principle of citation analysis. Other general principles are applied to judge and predict the analysis. Citation statistics are key components of the basic steps of citation analysis. An analysis must be conducted based on citation statistics, regardless of the type of adopted citation analysis. Therefore, citation statistics are preconditions for citation analysis.

In collecting citation data for citation statistics, we must choose the standard statistical objects first, and the citation information source of literature can be provided. Many types of literature sources are available for citation statistics, such as review journals and other basic publications. A source can come directly from the original paper in the journal citation data.

At present, the most useful tools available for citation analysis are mainly SCI and JCR of the United States, and the Chinese SCI. In fact, only SCI can reflect the relationship between citations and cited scientific papers. Only JCR can reflect the relationship between citations and cited journals. In citation analysis, we can select appropriate tools for statistical analysis to realize the purpose and fulfill the requirements.

Scientific citation is presented by many authors and is obtained from different sources of journals or literature. It exhibits considerable randomness, which is influenced by article control factors to a large extent. However, science citation exhibits a certain distribution structure and regularity when we conduct a number of science citation statistical analyses. Research on distribution theory and the law of scientific citation is one of the important contents of literature.

8.2 Citation Analysis of the Main Tools

8.2.1 SCI

SCI is published by the Institute for Scientific Information. It was founded in 1961 and included 613 types of periodicals. It was not published on a regular basis from 1961 to 1965, and only three volumes were published during this period. Publishing was scheduled quarterly in 1966, and it became bimonthly since 1979. JCR is issued annually, whereas other annual journals are accumulated for 5 years.

In addition to the print edition, many versions of SCI are available in the United States with the development of computer technology, communication technology, and high-density storage technology, including the tape edition, CD-ROM version, and online version. SCI refers to the print and CD-ROM versions, whereas the tape and online versions are called SCI-Expanded.

SCI reports science and technology journals from nearly 50 countries and regions worldwide, including meeting records and academic books, and a small number of book reviews. The contents of 100 subjects are a type of cross-disciplinary international large-scale integrated retrieval periodicals, such as life science, physical chemistry, clinical medicine, agriculture, biology, engineering, and technology.

(1) Introduction to SCI structure

SCI is distributed bimonthly. Some versions are distributed annually, whereas others are accumulated for 5 years. The bimonthly issue currently has six volumes. A, B, and C are the booklets for citation index. D is the booklet for corporate index

and source index. E and F are the booklets for permuterm subject index. The structure of SCI for the year 2001 is illustrated as an example.

1) General introduction

This part briefly introduces the composition of SCI, its uses, functions, rules, and characteristics. A statistical comparison table is also provided. This table lists the sources of SCI from 1955 until the previous year, the statistical data of the source, and the references. The last part provides the source direction of publications, which is composed of a new control table and a new list of publications arranged in accordance with the abbreviation of publications.

2) Citation index

The citation index is the first part of the body of SCI, which reflects the literature cited in the past years. It can be divided into four types according to the different types of cited literature and the situation of authors. These four types are established by citations and cited literature. The description format and entry sequence of the citations for the four types are the same. The description of the project is as follows: author citation, abbreviation of publications, volume, start date year, literature type code, the same entry arrangement order and description of the project, and an alphabetical arrangement of author citations by publication acronyms. Citation is the same for the four types of index; hence, the following will be explained based only on the citation index of items.

① Author citation index

The author citation index is the personal description of the author in the cited literature retrieval. The search portal is cited as the first author. The cited item description is as follows: cite the name of the authors, year of publication, volume, and page. The entry sequence is the same as the description item order located on the cited item citation.

② Corporate author citation index

The description of the author is a group or organization cited in the literature according to the alphabetical arrangement of its publishing or distribution unit. The scope of collection includes the bible, contracts, reports, notices, books, and conference proceedings. The publications of the United States and other countries can be found in the corresponding publishing units. With the exception of the cited literature published in the year, which provides the full name at the end part, the other items are the same.

③ Anonymous citation index

It is a description of the unsigned cited literature by cited publications. The other components are the same as those of the citation index when the description of the cited authors of the project is short. In newspapers, magazines, and other volumes,

the pages of publications are not clear. The month and the day will be determined after publication.

④ Patent citation index

Patent literature cataloguing is cited strictly by patent number. The body has a patent country table

3) Corporate index

In the beginning of the fourth volume, an index based on the location and name of the author unit in the source index is located. It is then divided into regions and institutions. It introduces the volumes that may use the code, rule, list of source publications, source index, and group index. The guidelines, abbreviated institutions, and geographic name table in front of the literature index are used. The index text is found at the beginning.

① Geographic section

This section can be ranked based on the name of the author unit seat or the state. The part of a single table in the United States is located at the front according to state name.

② Organization section

The search portal is an organization, which lists the countries and regions of the institution.

4) Source index

It reflects the details of the literature source and the names of the authors alphabetically. The items that appear in the source index are the same as the corresponding items in the citation index because the index terms are derived from the citation index. The source index is divided into two: the anonymous source index and the source index with different author situations.

5) Permuterm subject index

As the traditional index, it selects words with practical significance from the paper title reported by source citation. These words can be regarded as the main and secondary words for different groups. The main figure follows Z in the beginning of the entry according to the arrangement of the the main words in alphabetical order. Before retrieval, an entire disable word table and a semi disable list exist; the former is not regarded as the main nor secondary word, and the semi disabled vocabulary words cannot serve as the main words but can be used as secondary words.

(2) Development and change in SCI printing in the United States

From its publication from 1961 to the present, SCI has been constantly developing and changing, as reflected in the following aspects.

First, a new Corporate Author Citation Index was added in the second phase of 1996. Second, various codes and their meanings were changed. In 1993, only 37 types of language codes existed before the fourth phase. From the beginning of the fourth phase, 48 types were added. With regard to the patent country code, 40 types of this code were available in second phase before 1995, and the code capital letters varied from 2 to 4. It increased to 53 types expressed with two capital letters from the beginning of the second phase in 1995. The literature and types of the scope of the collection have also been changed. For example, SCI only included 500 selected publications of conference abstracts in the beginning of 1989; before that, it included all sources of publication. Reprints was added in 1991 and News Items in 1996. The bibliography was classified as a review before 1989 and became independent in 1989. Discussions was included in the Editorial Material, and Notes was included in Articles in 1996. Afterward, the two types disappeared. Chronologies was no longer included in 1996. These code changes show that the scope of SCI gradually expanded. The description has become highly standardized, and the meaning of the code has become clear.

Third, the format of description has changed. In the citation index, the cited literature age is provided in full, such as 1997, whereas the original only provided the last two digits, such as 97. The request number of ISI is included in the source index but not the original. These changes in description format allow the user to discriminate the citation and the cited literature easily. As a result, searching for documents has become convenient, and many channels of original literature can be accessed.

Fourth, the volume and its content have changed. The volume of SCI changes with the expansion of the scope of the collection. Before the fifth issue in 1988, only four booklets (A, B, C, and D) were available. A and B are for the citation index, C is for the source index, and D is for the permuted index. From the fifth issue in 1988 to the second issue in 1996, it increased to five volumes (A, B, C, D, and E), A, B, and C are for the citation index, D is for the source index, and E is for the permuted index. Six volumes (A, B, C, D, E, and F) have been released from the third issue in 1996 to the present.

Fifth, other aspects have also changed. Each issue in the first part before 1990 was located in the booklet of the source index. Each issue was then located in C before the issue of 1988 and in D from the fifth issue of 1988 to the first phase of 1990; however, it was located in the first of each issue in A. In addition, the description of the footer changed from the middle of the third issue of the first booklet in 1966, and the footer was included with the SCI bimonthly author citation index, PG. 1 and SCI bimonthly source index, PG. 323. The latter description tells the user the page position and type, and the former does not. Users who are unfamiliar with it must turn to the start page of each part to know what the index is. These small changes brought great convenience to users.

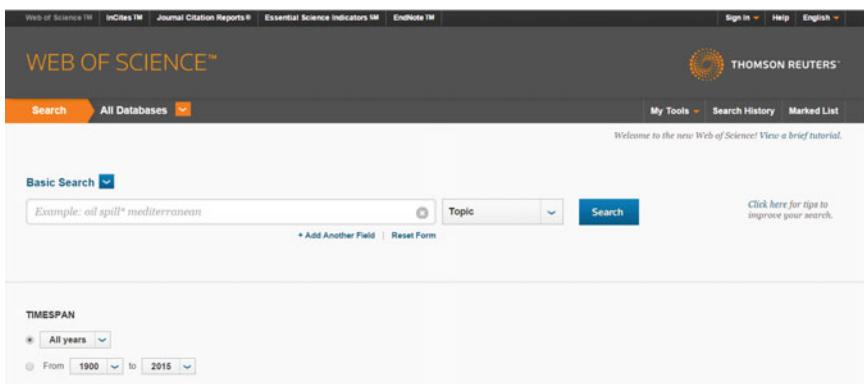


Fig. 8.2 ISI web of science

(3) Network version of SCI and its use

The SCI network version of the collection covers more than 6000 types of journals in a field of more than 150 subjects. It includes more than 2300 types of print and discs and all references. It is the authority in science and technology literature retrieval. It can be used to find the latest research results and to provide references to cited literature retrieval. The unique citation retrieval system is commonly used as an academic evaluation tool. The database of the SCI network version is updated weekly. The SCI network provides the latest literature retrieval and backtracking; it backs data to 1945 and provides English abstracts as far back as 1991. In addition to an increase in new records for about 17751 per week and the addition of about 362000 citation references and more than 700 patent citations, a total of more than 17 million articles (about 70% of which are in English) can be retrieved. The libraries of Beijing University, Tsinghua University, Wuhan University, and Chinese Academy of Sciences ordered the SCI network, which can be accessed for retrieval in the campus or area network (Fig. 8.2).

1) SCI network version retrieval

The SCI network can be retrieved in two ways: through the web version of ISI and online through an SCI search, which can be retrieved via an online dialog in the United States. The database provides two ways of searching, namely, Easy Search and Full Search (Figs. 8.3 and 8.4)

① Easy Search

Three retrieval approaches, TOPIC, PERSON, and PLACE, are available for Easy Search. The system can be retrieved through a series of simple prompts and questions.

For subject retrieval for vocabulary, one can input the title, abstract, keywords, and related vocabulary. The name search relates to the source author, citation author, and characters in the literature, thus allowing for the use of a truncation

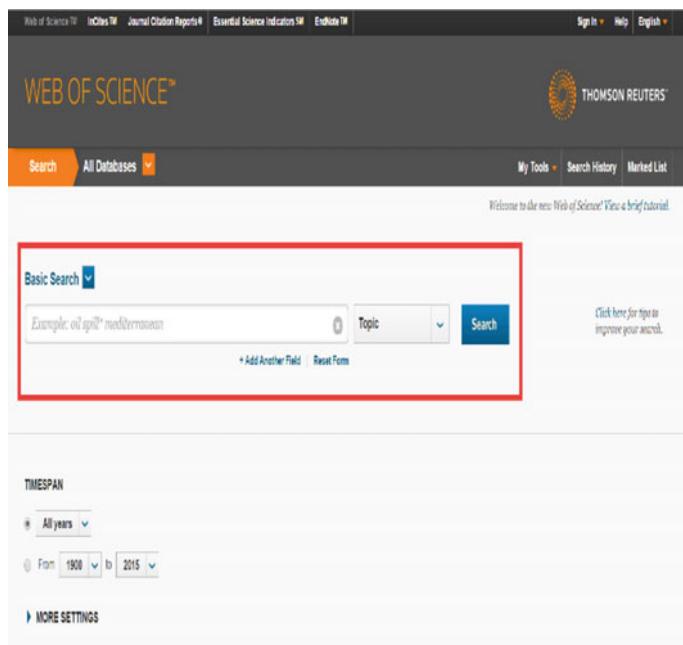


Fig. 8.3 Easy search

symbol to input the full name of the characters, space, initial word of the character's name, and the surname followed by a cut word operator. For place name retrieval, one can input the author name of an organization, country, city and zip code information, etc. The three approaches can be used with the logical operators And, Or, Not, Same, etc. The search results are displayed for the user and can be sorted by correlation degree or date. The search terms shown are limited to 100.

The method is generally convenient, fast, and suitable for beginners or users who require high novelty search rates or system response time.

② Full Search

Full Search, which is a more professional search method, can be set according to user needs. Different limits are set so that users can find the latest, most complete, and most accurate information. Four search options, namely, general search, cited reference structure search, advanced search, and open histories, are available.

A. General Search

General search is implemented by inputting the topic, authors, journals, and address of authors. In one or more retrieval approaches, Boolean Retrieval (AND, OR, NOT, and SAME) can be used to input key words or phrases. It can be used wildcards (* or ?) to expand the search scope. Different retrieval approaches are automatically connected with the operator of AND. The retrieval results can be

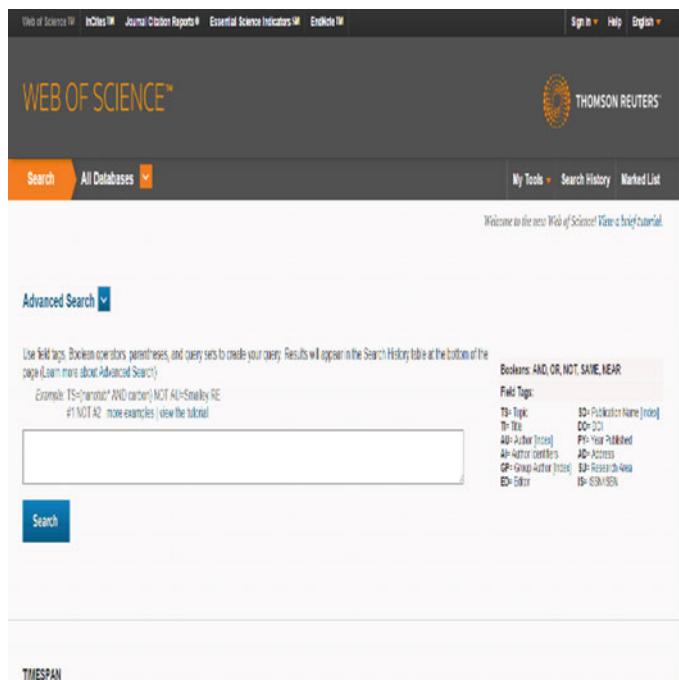


Fig. 8.4 Full search

restricted to the language, literature type, and, method of ranking. The searching result is displayed in the summary (Figs. 8.5 and 8.6).

- Topic

A retrieval term or phrase, which is defined in the topic of paper, abstract, and key words, can be entered. If retrieval topic is selected, the user search is limited to only the topic of the paper.

- Author

The first name of the author, the surname followed by a space, and the first letter of the surname (which can take five letters of the surname) are entered. If the name of the author cannot be fully determined, the truncation symbol as “” can be used together with a space to cut off the first letter. If the first letter of the name is unknown, the surname can be imputed.

- Source title

It can be retrieved from the records of published journals or from the list of journals when the name of the source journals are inputted in full or in part.

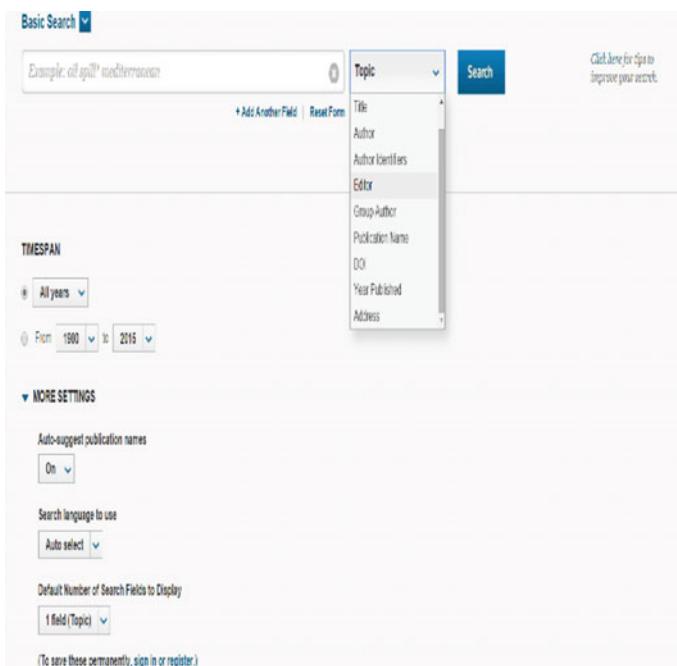


Fig. 8.5 General search table

- Address

Address terms, such as the name of the organization, city, and state or post code, can be inputted to retrieve articles published by an organization. The name and place of an agency are often abbreviated, and the address of the abbreviation table can be obtained by checking the online help system (Fig. 8.7).

The result of the retrieval is in two formats: summary and full record. The former provides the title of literature, author, and the name of the source journal and its volume and page number. Ten records are displayed per screen, and the full record format can be linked. The literature language, type, and number of references and citations are also provided in addition to the general information. The full record format can be selected from the literature by making a mark, clicking and “holding” the icon to be added to the collection of records, and clicking on the related records. The icon can be viewed with the literature citing the same information, and the function shows the situation of cross penetration in subjects.

B. Cited Reference Search

Cited literature search is a powerful search method for ISI Web of Science. Citation information retrieval can be implemented to find various references (patent, monographs, and conference literature) of literature. Citation retrieval is divided into two steps. First, according to user need, the references are selected in the

The screenshot shows the Web of Science search interface. At the top, there are tabs for 'Web of Science™', 'InCites™', 'Journal Citation Reports™', 'Essential Science Indicators™', and 'EndNote™'. On the right, there are links for 'Sign In', 'Help', and 'English'. Below the tabs, the 'WEB OF SCIENCE™' logo and the 'THOMSON REUTERS' logo are displayed. The main search bar has the word 'Search' and a magnifying glass icon. To the right of the search bar are 'My Tools', 'Search History', and 'Marked List' buttons. The search results summary on the left indicates 'Results: 165,394' from '16 databases' with 'Order of results by publication date'. A search term 'TOPIC: (knowledgemanagement) ...None' is shown. On the right, the results are listed with various filters and options like 'Save to EndNote online' and 'Add to Marked List'. The first result is titled 'Application of Systems Theory in Longitudinal Studies on the Origin and Progression of Alzheimer's Disease.' It includes author information (Lista, Simone; Khachaturian, Zaven S.; Rujescu, Dan; et al.), journal details ('Methods in molecular biology (Clifton, N.J.) Volume: 1303 Pages: 49-67 Published: 2010'), and citation metrics (Times Cited: 0). The second result is 'Alveolar echinoccosis: how knowledgeable are primary care physicians and pharmacists in the Franche-Comté region of France?' with similar details.

Fig. 8.6 Summary

This screenshot shows a detailed view of a specific article record. The top navigation bar is identical to Fig. 8.6. The main title of the article is 'Application of Systems Theory in Longitudinal Studies on the Origin and Progression of Alzheimer's Disease.' Below the title, the authors are listed as 'By: Lista, Simone; Khachaturian, Zaven S.; Rujescu, Dan; Garaci, Francesco; Dubois, Bruno; Hampel, Harald'. Publication details include 'Methods in molecular biology (Clifton, N.J.) Volume: 1303 Pages: 49-67 DOI: 10.1007/978-1-4619-2627-5_2 Published: 2010'. The 'Abstract' section begins with a note about the prevalence of an 'implicit' assumption in the field. The 'Citation Network' panel on the right shows '0 Times Cited' and '0 Cited References', with a link to 'Create Citation Alert'. It also includes a note '(data from Web of Science™ Core Collection)' and a link to 'View PubMed Related Articles'. The 'All Times Cited Counts' section shows '0 in All Databases'.

Fig. 8.7 Full records

retrieval results to find citations of all articles in the literature. As shown in the search page, a user can input the cited author name, cited journals (patent number, name of monographs known as cited research work), and the time of the cited literature. The scope of retrieval can be expanded by wildcards. The second step is

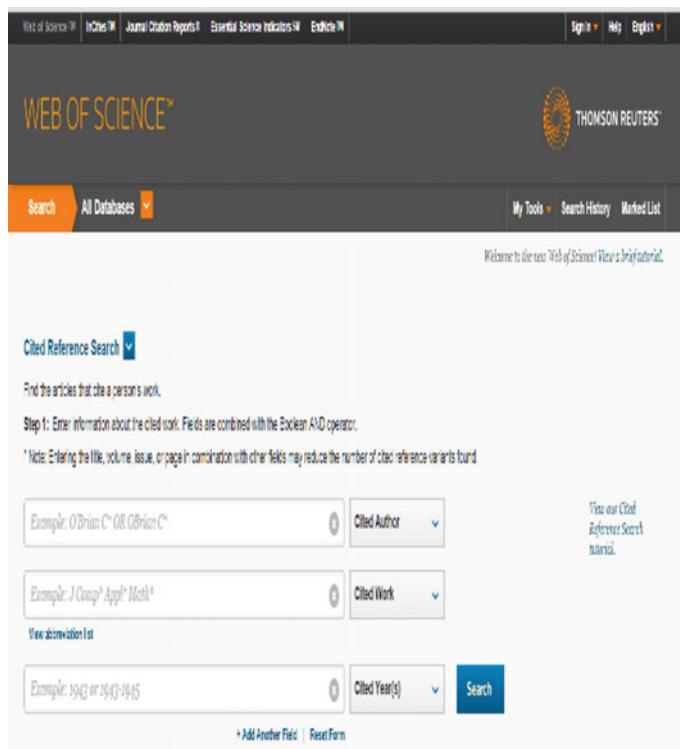


Fig. 8.8 Cited reference search—lookup

to find and display the citations, and several quotations will have different writing forms (Figs. 8.8 and 8.9).

The point of view of a cited literature or when an invention is to be confirmed, utilized, extended, and improved can be determined. Citation retrieval includes cited and cited works (books, periodicals, patents, etc.), age retrieval approach, or combined retrieval.

- Cited author

A user can input the first author's name of a cited literature. If the citation involved in the web of science includes literature sources of journal articles, a user can input the cited literature author's name. The longest surname of the author cannot exceed 15 characters. After 15 characters, the name is cut, followed by a space. Up to three initials of the name can be used to cut the word after the first letter.

- Cited work

A work can be cited by citation of a journal or a book title. Only 20 characters can be used in the extraction of the title, and the abbreviations and the title of a number

The screenshot shows a 'Cited Reference Search' interface. At the top, there's a note: 'Find the articles that cite a person's work.' Below it are instructions: 'Step 2: Select cited references and click "Finish Search."' and a hint: '+ Hint: Look for cited reference variants (sometimes different pages of the same article are cited or papers are cited incorrectly)'. On the right, there's a link 'View Cited Reference Search tutorial'. The main area is titled 'CITED REFERENCE INDEX' and shows 'References: 1-27 of 1'. Below that is a search bar with dropdowns for 'Select Page' (set to 1), 'Select All', 'Cite 40', and a 'Finish Search' button. The table has columns: 'Select', 'Cited Author', 'Cited Work (SHOW EXPANDED TITLES)', 'Year', 'Volume', 'Issue', 'Page', 'Identifier', 'Citing Article #', and 'View Record'. The data in the table is as follows:

Select	Cited Author	Cited Work (SHOW EXPANDED TITLES)	Year	Volume	Issue	Page	Identifier	Citing Article #	View Record
<input type="checkbox"/>	Arto F.A. Rousseeu S	Phys Rev Lett	2007	98		125001		1	
<input type="checkbox"/>	Bhattacharya C + [Show Authors]	MLQ-PHYS A	1999	65		641c	10.1002/1521-3775(199903)65:3<641::AID-MLQ641>3.0.CO;2-9	9	
<input type="checkbox"/>	Björk T., Rousseeu C	Journal of Chemical Education	1993	2		25-29		1	
<input type="checkbox"/>	Braine T., Rousseeu J + [Show Authors]	TRANSPORTATION	1994	18	1			4	
<input type="checkbox"/>	Gebra E.L., Rousseeu J.N + [Show Authors]	Computers & Operations Research	1994	2		125-133		1	
<input type="checkbox"/>	Hansen M., Rousseeu C	OPTIMIZATION	2000	4		xx		6	

Fig. 8.9 Summary

of important characters in the front can be inputted for books. A user can input a patent number in the field of publications cited for a patent.

- Cited year

For the retrieval of literature published by the publication of the cited literature, four digits of the publication year or the logical operator OR is inputted to connect a series of publishing year and patent release date.

C. Structure search

A structure search is used to retrieve the chemical reaction and compounds of a chemical structure.

D. Advanced search

Advanced search allows a user to use a highly sophisticated search and a combination of search terms.

2) Characteristics of the SCI network in the United States

The SCI network is mainly used to retrieve papers that were collected and cited similar to other versions. However, it is not a simple copy of the other version on the Internet regardless of the retrieval function, scope, update cycle, etc. It is much more powerful than other versions. More importantly, it combines the traditional citation index and advanced web technology not only to make all the information connected in the database, but also to link a number of other information databases that constitute a powerful, flexible, untraditional database. To sum up, the characteristics of the SCI network are mainly reflected in the following aspects:

① Characteristic link divided into internal and external links

- Internal links

These links include the number of citations, references, and the link of related records. For the source of the subject, one can find the number of times cited, cited reference, and related records in the full record. It displays the document list of the corresponding article. With a mouse click on three items, a user can view a paper that cited a list of other literature together with the paper that cited the same references in different years. According to the common references cited in the article, namely, relevance ranking, and the more times a literature is similarly cited in the current record, the closer the literature is to the subject. The position of this literature in the list is in the front. If no limitation exists, the retrieved records are all displayed. In general, the records in the list are underlined. If a user continues to click on a record, it will display a new document. Then, the full record of the text format can be viewed. This can be achieved by layers of in-depth excavation and retrieval of similar literature. With the update of the ISI database, all types of data change accordingly. However, some records in the list may not be underlined and cannot be viewed further because the records are not included in ISI or the year of the records is not within the period of the purchase year.

- External links

External links include the link to the original document, the library collection OPAC, and the other two or three information resource databases. When a user clicks on the full text of the literature record button, the original documents of the current record are displayed, but the premise is to buy the electronic version of the journal at the same time. Clicking on the button of holdings automatically accesses the OPAC library system to learn about its collection records (the link must be under the user's requirements; made by ISI for users). The two or three links of the literature information resource database include the following:

- A. The link to literature research can be carried out by the Web of Science to obtain more information on the relevant references and the full text of the literature with the link of the ISI Chemistry Server. More detailed chemical reaction information in the Science Web can be obtained to find literature by

- SM simply by clicking on the reaction button in the full record. One can learn about the chemical reaction in the past, present, and future.
- B. The links provide more than 40 patents issued by the patent literature information based on the web since 1963 with the Derwent Innovations Index of the Derwent patents citation index database.
 - C. The link to the Web of Science Proceedings includes two large databases: ISTP (Science and Technology Proceedings) and ISSHP (Social Science and Humanities Proceedings).
 - D. The link to BIOSIS Previews helps researchers quickly and comprehensively access life sciences and related fields of academic information.
 - E. The link to the NCBI Gen Bank is provided. Gen Bank is a gene and protein sequence database maintained by the National Institutes of Health in the United States. It includes all publicly available DNA and protein sequence information provided by NCBI, and the records include more than 4 million gene records.
- ② To make full use of the powerful force of the WWW, the SCI network version has completely changed the traditional literature retrieval method. Without the need to install any other software in the general browser interface and a new hypertext format, all information are interrelated.
- ③ The SCI network is updated weekly, and each update includes the entire system (number of citations, references, and related documents). The SCI disc version is updated with a summary each month. Without a quarterly summary update and the updated information and data for the current year, the past cannot be updated.
- ④ The SCI network allows for the retrieval of the entire database or the specified year. It can retrieve all cited authors rather than the first author when data are backed to 1945; it can cite based on maximal retrieval, which is based on the annual SCI disc version back in 1980. However, the retrieval results must be further ordered, and the first author can be retrieved.
- ⑤ The SCI network can be directly included in electronic journals, which is convenient to reflect relevant research results. However, the performance of electronic journals, such as PDF and HTML, includes a variety of versions (manuscript, preprint, submit the draft, revised, finalized, etc.), and several purely electronic journals are not formal. Sometimes, it is difficult to make a correct judgment of its content, which affects the accuracy of the link. Uniform standards and rules need to be developed.
- (4) Comparison of various versions of SCI (Table 8.1).

8.2.2 *Essential Science Indicators (ESI)*

Essential Science Indicators (ESI) is a world-famous academic information publication agency of the Institute for Scientific Information (ISI). Its research service

Table 8.1 Comparison of various versions of SCI

Index	Network edition	CD with abstract	CD	Online	Printing
Collection scope	6000 kinds of journals, including the newest journal and CD	3500 kinds of journal	3500 kinds of journal	More than 6000 kinds of journals, including CD version and the latest journal periodicals	3500 kinds of journals
Publication cycle	Weekly updates, back data to 1945	Monthly update, data back to 1991, with annual accumulated in two discs	Quarterly update, back to 1980, with the annual accumulated in one disc	Monthly update, in the system of Dialog, DIMDI STN trace the data to 1947, and trace data to 1980 in Data star	Published bimonthly, back data to 1961
Retrieval method	Window type, provide hyperlinks entrance including shortcut and complete search,	Menu style, the interface provides a shortcut key, only one DOS interface, searching software requires a special study, provides help online	Menu style, the interface provides a shortcut key, only one DOS interface, searching software requires a special study, provides help online	Command help, which cannot provide convenient information portal on retrieval interface, retrieval software requires specialized learning, provides help online	Book-style, hand-searching, fixed retrieval entrance, retrieval methods require specialized learning
Information giving	Providing complete bibliographic information and the abstract of the author, proving the electronic mail and the website of journal publisher since 1997	Bibliographic information, abstract, references, and the number of relevant literature (up to 20 articles, not activated), which is the retrieval tool of the abstract	Bibliographic information, number of references and relevant documents (up to 20 articles, not activated), which is the bibliographic retrieval tool	Bibliographic information, which is a bibliographic retrieval tool	
Output format	Results can be arranged in accordance with the relevance degree, the date, or the name of the first author, download limited to 10 articles	the order is fixed, which is decided by the retrieval software, and the amount of download is not limited			

(continued)

Table 8.1 (continued)

Index	Network edition	CD with abstract	CD	Online	Printing
Recall	Fuzzy search function	Precise retrieval	Precise retrieval		
Retrieval time	There are two choices of Internet and Intranet, which are influenced by the network's export and transmission speed	There are two ways of single and area networks, the speed is faster	There are two ways of single and area networks, the speed is faster		
User fees	The purchase price is more expensive than the CD version (\$60000 in 1999)	The purchase price is lower than the online version (\$23000 in 1999, and the multi-client version is \$45000)			
System maintenance	Has nothing to do with the end user	The end user needs to download the software to support area network, the network of the disc should be maintained, and the workload is large	The end user needs to download the software to support area network, the network of the disc should be maintained, and the workload is large		
Others	Support maximal retrieval, support first author retrieval, online registration can be directly filled in from the full text delivery service	It can retrieve based on maximal retrieval, but the results must be arranged again	It can retrieve based on maximal retrieval, but the results must be arranged again		

group, which is the analysis and evaluation tool, was established in 2001 to measure scientific research performance and track the trend of scientific development. It is a metrological analysis database based on SCI (science citation index) from ISI and SSCI (social science citation index) that collects more than 85,000 academic journals that includes more than 1,000 million literature records. It provides services by the Web of Science of ISI and is an important part of the integrated service platform. ESI carries out statistical analysis and sorting from the perspective of citation analysis for 22 professional fields, including national research institutions, journals, papers, and scientists. The main indicators include the number of papers,

citation frequency, and average citation for papers. Users can learn from the database about the development and influence of a subject for a certain range of scientists, research institutions, national (city), and academic journals for determining the key scientific discovery, assessment of research performance, and mastering the trend of scientific development. It can systematically target and analyze academic literature by ESI. As a part of the ISI Web of Knowledge, ESI provides a dynamic, comprehensive research and analysis environment based on the network.

(1) Structure of ESI

The main contents of ESI include Citation Rankings, Most Cited Papers, and Citations Analysis. Citation Rankings include scientists, institutions (universities, enterprises, government departments or academic research institutions, etc.), countries, and the range table of journals. Most Cited Papers include highly cited papers, hot papers, baselines, and research fronts. Citation Rankings and Cited Papers provide links to the top paper pages and time sequence diagram. Aside from the three main modules, ESI also provides a comment for various forms and data, which include In-Cites, Special Topics, and Science Watch. Figure 8.10 shows the ESI page structure that reflects the relationship among the various pages. It introduces the modules as follows (Fig. 8.10):

1) Scientists ranking

Citation frequency is a form of peer recognition that often reflects the dependency of scientific research groups on the degree of scientists. It can even said that the essence of the scientific community is composed by the scientist of the most cited

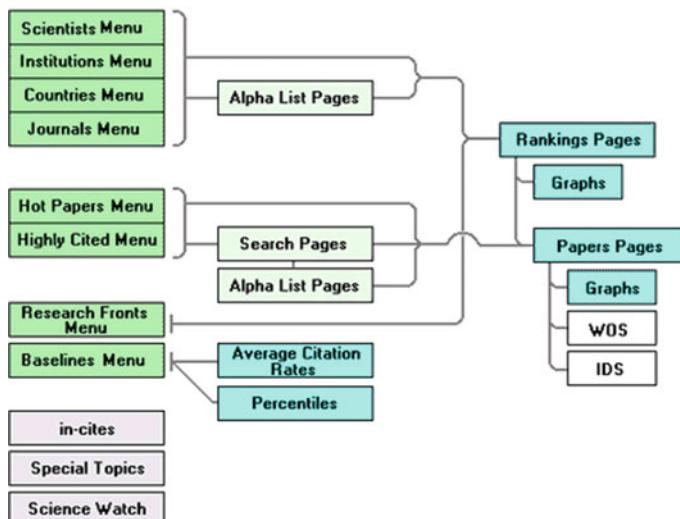


Fig. 8.10 Flow chart of the ESI page structure

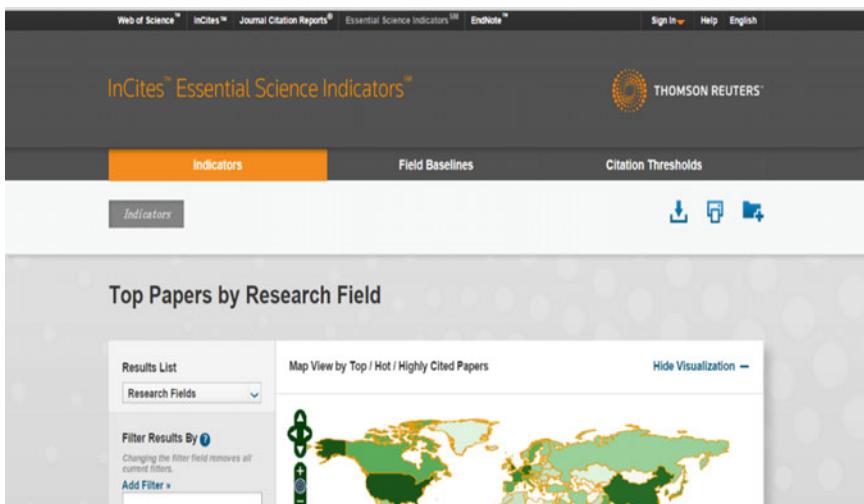


Fig. 8.11 Essential Science Indicators(ESI)

paper. Many highly cited scientists are recognized through an honorary award. ESI ranks scientists according to the 10 years of the sum of the papers cited by them. The ISI citation index database contains about 300 million scientists, of which only 5 million are included in ESI. The number of these scientists cited in 10 years is the top 1% of all scientists.

2) Institutional ranking

Scientific research is carried out in scientific research institutions, so the study of institutions related to the reorganization of scientists is reflected in the reputation of research institutions. Publication and citation frequency at the organization level can be measured by the total research organization. ESI ranks institutions according to the 10 years of the cited frequency of the institution papers. ISI includes about 100 million research institutions, and 3,000 of these represent the top 1% of all institutions.

3) National rankings

As measured by publication and citation, the level of scientific research achievements in various countries is unbalanced, and the distribution of scientific research achievements is uneven. The level of scientific research activity in a country is generally associated with the capacity of the country's GNP or other related economic output ability. The comparison of countries in a single paper is helpful to correct the differences in the national scale and the paper output. It is to rank according to the sum of the cited frequency of the papers of various countries. The ranking of a country belongs to the top 10% according to the total cited frequency.

4) Journal rankings

Similarly, there are differences in reputation and influence as reflected in the citation frequency. ESI provides a long-term journal citation ranking. The Journal Citation Report and short-term citation behavior are compared by the query of ISI. The ranking journals belong to the top 50% by the total cited frequency.

5) Highly cited papers

According to the cited frequency of a paper, ESI selects the top 1% within the scope to form the highly cited papers list. Generally, the peak cited frequency of the papers appear in the paper published in 2–4 years, and several papers are cited for many years. A few studies have delayed cognition. The difference is very large, which is related to the nature of the paper, the field, and the discovery of the report. For example, the cited frequency of papers is reported to increase very fast. With the further elaboration of the discovery, the cited frequency decreases rapidly. The cited frequency of the reported method and the technology gradually increases with the transmission of the method and technology and the application of the technology. ESI sets a relatively specific field and the year of different frequency standards to ensure that for the selected papers in the corresponding areas and years, the cited frequency is within the range of 1%.

6) Hot papers

Hot papers refer to the same field and other papers published in the year compared with the publication of a high frequency of citation. The choice of hot papers is also based on a certain condition, but the choice of the time period is relatively short. In other words, the publication age cannot exceed two years and is cited within the current two months. This means that the paper must be paid close attention to in a very close period of time. Every field and time period have been set under the selected conditions, and 0.1% of papers are selected according to the corresponding conditions.

7) Baselines

Baseline is the large measure of accumulation across paper group cited frequency. This large paper group has a certain citation frequency. The baseline is composed of average citation frequency and percentage two tables.

① Citations per paper

ESI calculates the average citation frequency of each year from 1992 to 2002 (10 years) based on the amount of the cited papers published from a particular year to the present. The average value is equal to the single citation frequency divided by the number of papers. The average number of 10 years is given in the “all years” column. The citation frequency of each discipline and all subjects is given. Agricultural science had 7.87 average citation frequency in 1994, which means agricultural science published from 1994 to the present was cited 7.87 times. In various areas or all areas of 10 years, averages can be used as scientists, institutions,

countries, and journal rankings given by the paper cited values of the baseline. The average value of a discipline in the year of independence can be used to compare the papers published in that year to determine whether this literature belongs to the highly cited papers list in ESI or are from the papers of the journal science citation index. (Table 8.2)

② Percentile

The percentage point is expressed as a benchmark, in which the fixed proportion or higher than the benchmark begins to fall. The percentage is typically 1%, which indicates the fixed proportion of the top paper and is sorted by the citation frequency. The percentage of ESI and year number is 0.01, 0.1, 1, and 10%. Citation frequency is highly skewed because many papers are not frequently cited. The highly cited papers account for only a small part, which is probably a set of distribution. One of the methods for selection is to order by citation frequency and then select a certain proportion of the previous papers. The percentage table shows the selection conditions for the selected frequency of the four different percentages of the various fields, the entire field, and each year. Table 8.1 lists the percentage of academic papers for each year from 1994 to 2004 and presents the citation rates of the papers before eligibility, including 0.01, 0.1, 1, and 10%. For example, the citation frequency of agriculture science papers under the selected conditions was 281 in 1994; the selected papers published accounted for 0.01% of the total number of papers.

8) Research fronts

Scientific research areas, especially those on the frontier of subject development, are characterized by the close contacts of scientists. Various patterns, including formal and informal, exist. The most outstanding of these is a scientist's reference to the work of another scientist. It reflects the process of how scientists achieve their goals on the basis of other people's work in the form of a reference. The citation network structure can be based on the result of several pieces of the original core status in the field of literature to describe a particular study. In the essential science indicators database, achieving the work goal involves "RESEARCH FRONT ANALYSIS." The research frontier is a set of natural and social sciences that is determined by the high frequency of core literature and the recent citation of these core papers. The core literature represents a series of contemporary literature, which is based on the combination of SCI and SSCI database updates once a year. The research frontiers of documents from 1994 to 2004 have been set up. The statistical indicators of each of the leading papers includes the subject of the research frontiers, a series of keywords about the theme (which compose the research front of highly cited core literature), the number of core literature in recent years, the total number of papers cited in this group, the citation frequency, and the average publication year. The development trend of a discipline or subject can be tracked or predicted through research frontiers.

Table 8.2 22 Subjects' citation frequency from 1994 to 2004

Citation frequency of each course from 1994 to 2004												
Subject	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	All Years
All areas	15.32	14.38	13.10	12.25	10.90	9.60	7.95	5.97	3.83	1.61	0.24	8.46
Agricultural sciences	7.87	7.53	7.22	6.55	6.10	5.41	4.60	3.28	1.96	0.83	0.11	4.56
Biology and biochemistry	26.87	24.41	22.25	21.51	18.66	16.27	13.78	10.30	6.56	2.83	0.39	14.97
Chemistry	13.49	12.51	11.70	10.86	10.04	8.78	7.45	5.50	3.81	1.63	0.24	7.64
Clinical medicine	18.00	17.54	15.63	14.38	12.80	11.27	9.40	7.07	4.59	1.95	0.26	10.09
Computer science	4.47	4.48	4.09	4.03	3.80	3.20	2.38	2.02	1.44	0.45	0.08	2.35
Economic management science	8.95	8.11	6.49	5.99	4.96	3.96	3.05	2.07	1.28	0.45	0.09	3.99
engineering	5.30	4.93	4.61	4.52	3.87	3.40	2.83	2.18	1.29	0.52	0.08	2.97
Environmental ecology	14.33	13.11	12.23	10.94	10.20	8.46	7.07	4.82	2.94	1.12	0.20	7.29
Geography	14.78	13.69	12.59	11.43	10.33	8.39	6.61	4.89	2.75	1.23	0.22	7.45
Immunology	33.62	31.65	28.71	25.85	24.25	20.61	17.76	13.58	8.58	4.10	0.52	19.06
Materials science	7.60	6.91	6.50	5.80	5.40	4.76	4.11	3.10	1.92	0.78	0.10	3.97
Mathematics	4.97	4.63	4.32	3.77	3.27	2.92	2.22	1.57	1.01	0.41	0.06	2.55
Microbiology	23.31	22.27	20.39	19.55	17.54	15.23	12.64	9.50	6.03	2.53	0.32	13.53
Molecular biology and genetics	42.50	39.20	35.70	34.00	31.17	27.80	22.94	17.45	11.30	4.79	0.72	24.11
Cross discipline	2.52	3.19	2.97	3.64	4.09	5.27	4.75	5.95	5.94	3.63	0.71	3.76
Neuroscience and behavioral science	30.00	27.66	24.52	22.60	20.28	17.87	14.72	11.15	6.83	2.67	0.32	15.89
Pharmacology and toxicology	15.51	14.93	13.26	12.72	11.13	10.30	8.67	6.66	4.46	1.74	0.23	9.01
Physics	11.87	11.37	10.70	9.60	8.81	7.85	6.78	5.10	3.44	1.59	0.27	6.91
Botany and zoology	10.73	10.23	9.47	8.51	7.40	6.43	5.26	3.91	2.38	0.97	0.18	5.89
Mental disorders/psychology	15.86	13.64	12.11	11.32	9.87	8.75	6.63	4.93	2.80	1.09	0.19	7.87
Introduction to sociology	5.59	5.42	5.12	4.75	4.35	3.66	2.96	2.08	1.29	0.51	0.11	3.26
Space science	17.46	17.36	16.22	16.20	14.17	14.51	10.37	9.60	5.72	3.23	0.58	11.09

9) Top papers

A top paper is ranked in the top 1% cited papers, which involve top scientists, institutions, countries, and journals in specific areas. In the list of scientists, institutions, countries, and journals that are cited in the table, if an object has a top paper standard, the records of “VIEW” have a link to the top list. As shown in Fig. 8.12, rank is done in a descending order in the top list of papers cited frequency. Another way of ranking is to sort from the first page via the drop-down menu. Each record of top papers includes citation frequency, time sequence diagrams, title, author, source journals, and author address. When a certain number of ESI papers cited by the authors, institutions, countries, and periodicals of a corresponding link is reached, the ranking of the object can be viewed by the link. If a user wants to view an article including its abstract, the citation information, and other complete records, he/she can click on “GO TO WEB OF SCIENCE” in the right corner; however, this feature is only for users who subscribe to “ISI WEB OF SCIENCE”

10) Chart page of ESI

Two types of ESI chart page are available to provide readers a more intuitive understanding. The format of the two charts is listed as follows. One is the top of the charts and highly cited paper chart, which is from the individual at the top of the paper and highly cited papers. For example, as shown in Fig. 8.12, each paper record of the first “citation frequency” has an icon on the right; one can click on “enter” to access the cited frequency chart page, which has only one figure displaying the papers cited frequency along with the trend of the time change cited by time on the horizontal axis and by year on the vertical axis (Fig. 8.13).

The screenshot shows a search results page titled "Papers by Research Field". On the left, there's a sidebar with filters: "Citation Trends" (selected), "Documents", "Filter Results By" (with an "Add Filter" button), and "Include Results For" set to "Top Papers". Below these are "Clear" and "Save Criteria" buttons. The main area displays three search results:

- 1 GLOBAL CANCER STATISTICS** (Times Cited: 10,844)
 - By: JEMAL, A.; BRAY, F.; CENTER, MM; et.al
 - Source: CA-A CANCER J CLIN 61 (2): 69-90 MAR-APR 2011
 - Research Fields: CLINICAL MEDICINE
- 2 GLOBAL CANCER STATISTICS, 2002** (Times Cited: 8,661)
 - By: PARKIN, DM; BRAY, F.; FERLAY, J; et.al
 - Source: CA-A CANCER J CLIN 55 (2): 74-108 MAR-APR 2005
 - Research Fields: CLINICAL MEDICINE
- 3 CANCER STATISTICS, 2010** (Times Cited: 6,963)
 - By: JEMAL, A.; SIEGEL, R.; XU, JQ; et.al
 - Source: CA-A CANCER J CLIN 60 (5): 277-300 SEP-OCT 2010
 - Research Fields: CLINICAL MEDICINE

At the top of the main area, there are buttons for "Sort By" (set to "Citations"), "Customize Documents", and a page navigation bar showing "1 - 10 of 23,891".

Fig. 8.12 Top paper view

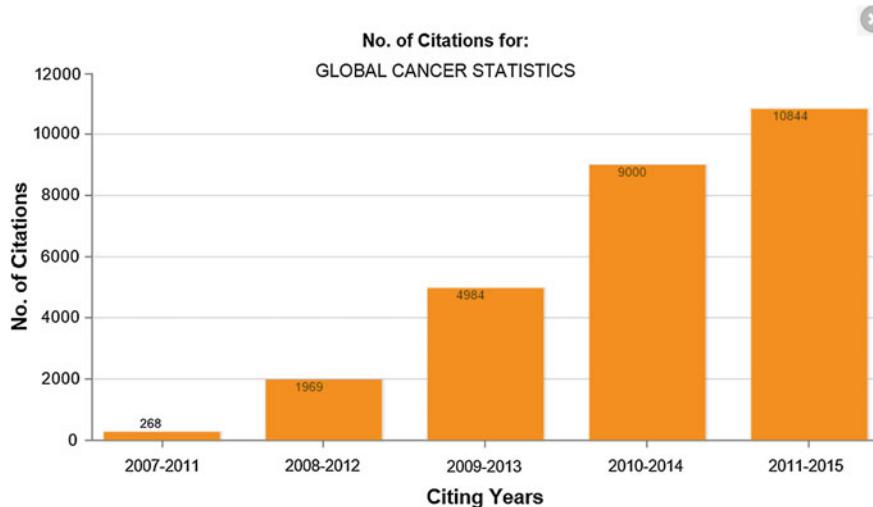


Fig. 8.13 Top papers cited chart

The other is a highly cited paper time sequence diagram from scientists, institutions, countries, and periodical journals. It includes three kinds of graphics, as shown in Fig. 8.14.

- ① The number of papers for 5 years by a period of time displaying the past 10 years was divided into 5 years for a period of time series published in the papers (papers, reviews, conference papers, and research notes) for the quantitative development trend.
- ② The number of papers for 5 years by a period of time displaying the total number of papers published in the past 10 years was divided into 5 years as a period of time for the development trend of quantity.

Normalized	2005-2009	2006-2010	2007-2011	2008-2012	2009-2013	2010-2014	2011-2015
Total Papers	100%	106%	112%	118%	124%	130%	119%
Total Citations	100%	105%	111%	119%	125%	131%	106%
Total Citations per Paper	100%	99%	99%	100%	101%	101%	89%
Top Papers	100%	106%	112%	119%	124%	130%	123%
Citations to Top	100%	91%	82%	71%	57%	42%	28%
Citations per Top	100%	86%	73%	60%	46%	33%	23%

Fig. 8.14 High citation time series graph

- ③ With 5 years as a period of time of average citations per paper, each column graphic display is a 5-year average frequency of reference or the ratio of the number of papers to citation times. Each calculation includes only the number of papers within 5 years and the number of literature cited.

Each figure is cited in time on the horizontal axis, and the past decade is partitioned into overlapping sections with 5 years as a range of time series. This is mapped out on the vertical axis of the histogram with regard to the number, citation frequency, and citation times.

11) In-Cites

In addition to various ranking and other measurement data, ESI also provides the material editor comments for regular updates. It includes In-Cites, a type of data about the corpus with a review collected and indexed by ESI. It is specially written for ESI to the corpus and the original review of the column. Scientists narrate about their highly cited papers behind the scenes and about their field of application prospect as well as the future development of the review. In addition, feature articles point out the emerging disciplines, highly cited research institutions, and different countries of the research situation, with a high impact factor for journals and other topics.

12) Special Topics

Special topics based on the analysis of the selected topic areas within the field provide an in-depth survey. Clicking on a given topic calls out the data on this topic. These data include field growth in the field of highly cited papers, scientists, research institutions, and countries. Special topics represent a relatively narrow standard of literature. ESI uses a combination of lexical retrieval and research frontier analysis to determine the subject.

13) Science Watch

ESI contains the material of “science observation” for a year ago. Scientific observation in the news weekly reflects the scientific progress and achievements of news weekly in 1990 launched by ISI. The features are based on the recent and most comprehensive analysis of the citation. The typical content is that the selected areas of highly cited scientists and institutions are ranked, the hot or emerging areas are reviewed, and national or international research trends are tracked. Another feature is the publication of the world’s top scientists. Each of the 10 major papers published in the last two years is included in biology, medicine, physics, and chemistry, with a high frequency of citation in the last months. Every list is accompanied by an expert review. The target readers are science universities, policy makers, enterprise management personnel, scientific journal editors, and any person who wants to thoroughly understand contemporary scientific research progress. A few recently published scientific articles discuss the influence of seven western countries. It also includes universities in chemistry, immunology, clinical medicine,

electrical engineering, and other areas with the citation impact of ranking, the discussion of academic leaders, and the world's most accomplished scientists.

(2) Function of ESI

1) Search function of ESI

Retrieval methods differ because of their different retrieval object and requirement, and the retrieval results are also different. The following provides an introduction of the ranking of citations, analysis of highly cited literature, and part of the retrieval citation analysis. The author added many pictures to illustrate.

① Citation ranking

To present ranking information, ESI mainly provides two ways: “by field” and “by name.” The search interface and the principle of scientists, institutions, journals, and countries are basically the same. The following is a case study of the process about institutions and scientists. General users can log into the platform (need an account) of ISI web of knowledge from two web addresses (<http://www.isiwebofknowledge.com> or <http://www.isinet.com>). This case, which provides the trial account login, is from Wuhan University’s library. As shown in Fig. 8.15, for the login interface, one needs to choose ISI Essential Science Indicators in the drop-down menu and click on the button “GO” to enter the home page of the ISS Essential Science Indicators (Fig. 8.16). Clicking on the appropriate link will retrieve scientists, institutions, countries, and journals from the citation order. With the institution as an example, one needs to select the “institutions” to enter the page of the ranking menu, as shown in Fig. 8.17. The retrieval approach is divided into two columns: “by field” and “by name.”

A. Citation by field

The drop-down menu of the “by field” column lists 22 disciplines contained in ESI. If one wants to retrieve a specific research area of highly cited institutions, he/she can select the field from the drop-down column of “by field.”

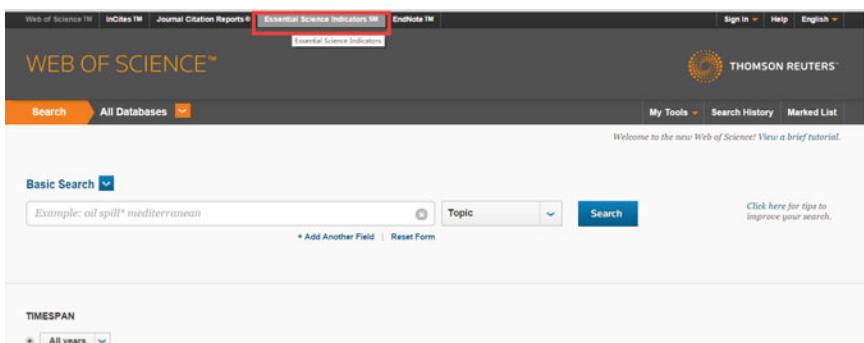


Fig. 8.15 ISI web of knowledge SM



Fig. 8.16 ISI essential science indicators

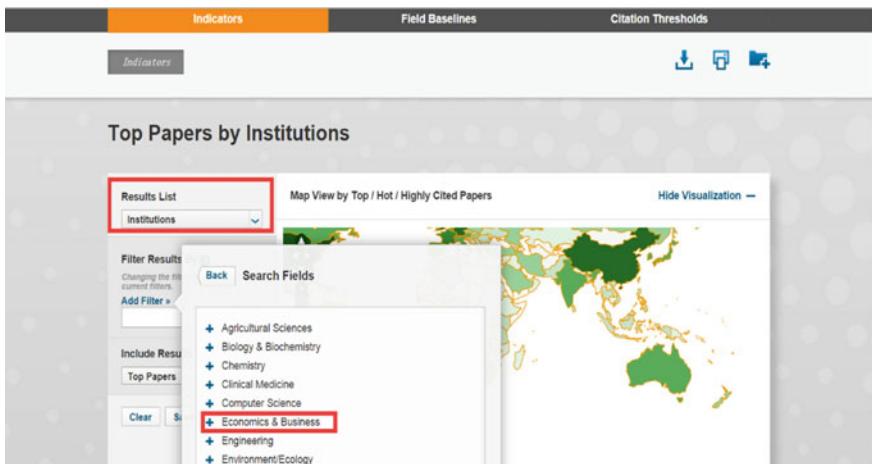


Fig. 8.17 Agency ranking menu page

Retrieval results: With “Economics & Business” as an example, after selecting the field, the ranking results within the field of organization are revealed, as shown in Fig. 8.17. The default order is the total number of citations. The drop-down



Fig. 8.18 Scientists ranking search page

menu in the “sorted by” column next to the total number of papers and citation frequency can also be used. The reordering of organization names is alphabetical.

The sorting table is divided into five columns. The first column is the sequence number of institution ranking. The second column “VIEW” shows the organization in the field of “Economics & Business” in the list of the top 1% (Fig. 8.12). The paper in the top 1% and the following “papers” (the total number of collected papers) are different in the columns. By clicking the “VIEW” icon in the column chart, one can see the graphical explanation of the citation data. As mentioned previously, there is no need to repeat them. The third column is the organization’s name, which can be linked to the institution’s various academic papers cited frequency sorting table. The last three columns are the total number of papers published in the last 10 years, the total number of times cited, and the average citation per paper.

B. Citation by name

With the sorting by scientists as an example, the citation order menu is accessed to select “scientists” in the scientists ordering page.

In step 1, the name alphabetical list can be used to search for a name or input name directly for each data element, scientist, periodical, country, and institution (Fig. 8.18). For example, if Hendrickson WA is inputted, the page shows all persons ranked in the top 1% of papers about the research field in the sorting table, as shown in Fig. 8.19.

In step 2, with the name Hendrickson as an example, the name is cited in the field of BIOLOGY & BIOCHEMISTRY for 2297 times. Clicking on the

Journals By Rank		Categories By Rank						
All Journal Categories ranked by Number of Journals				Show Visualization +				
1 - 25 of 232				Customize Indicators				
Category	Edition	#Journals	Total Cites	Median Impact Factor	Aggregate Impact Factor			
MATHEMATICS	SCIE	312	319,429	0.001	0.141			
BIOCHEMISTRY & MOLECULAR BIOLOGY	SCIE	290	3,273,847	2.672	4.149			
MATERIALS SCIENCE, MULTIDISCIPLINARY	SCIE	260	2,208,680	1.567	3.673			
MATHEMATICS, APPLIED	SCIE	257	407,510	0.828	1.097			
PHARMACOLOGY & PHARMACY	SCIE	255	1,285,250	2.362	3.030			
NEUROSCIENCES	SCIE	252	1,987,755	2.791	4.010			
ENGINEERING, ELECTRICAL & ELECTRONIC	SCIE	249	980,001	1.235	1.798			
EDUCATION & EDUCATIONAL								

Fig. 8.19 Hendrickson WA

Map View by Top / Hot / Highly Cited Papers				Show Visualization +	
Report View by Selection				Customize	
Total:	Authors	Web of Science Documents	Cites	Cites/Paper	Top Papers
12917	HENDRICKSON, WR	61	2,495	40.90	5
	HENDLER, T	100	1,812	18.12	0
	HENDLISZ, A	46	1,795	39.02	2
	HENDON, HH	52	1,414	27.19	3
	HENDRICKSON, CL	63	2,207	35.03	0
	HENDRICKSON, WA	62	1,958	31.58	2
	HENDRICKX, F	57	1,039	18.23	1
	HENDRICKY, M	167	2,766	16.56	5
	HENDRICKA, MC	168	1,195	7.11	0
	HENDRIE, H	17	2,455	144.41	2
	HENDRIKS, AJ	103	1,132	10.99	0
	HENDRIKS, J	1	875	875.00	1

Fig. 8.20 Ranking of Hendrickson in the field of BIOLOGY & BIOCHEMISTRY

underlined BIOLOGY & BIOCHEMISTRY in the table shows the ranking of the person in the field, as shown in Fig. 8.20.

Step 3 uses the navigation buttons to find the page containing the Hendrickson name. Hendrickson is cited 2297 times in 31 articles from the field of BIOLOGY &

BIOCHEMISTRY. The average citation for each paper is 74.10 times. According to the total number of citation times in the field of scientists ranking, Hendrickson ranks 605th.

② Retrieval of highly cited papers

The retrieval methods and results are basically the same between highly cited papers and hot papers. Two ways can be adopted: by field and by search. The highly cited papers are selected from the last 10 years within the month, and the hot papers are published in the 2 years within the last two months. The retrieval of highly cited papers is taken as an example.

A. Retrieval by field

The process proceeds from the home page of ESI to the highly cited papers page, as shown in Fig. 8.21. The highly cited papers menu is divided into two parts; one part is for by field, and the other part is for by searching. If the user wants to see the top 1% of the list in a certain area, he/she can select from the drop-down menu in the field and then click “GO.”

B. Retrieval by searching

ESI also provides five ways to retrieve information from the drop-down menu in the column of “by searching,” as shown in Fig. 8.22. These five ways include the subject, name, organization name, state, and journal name. The user can input the content in the display search page and then click the button “search.” The highly cited papers will be displayed (Fig. 8.23).

If the user wants to know the frequency of papers in the field of the subject within the scope of the cited papers, as shown in Fig. 8.26, the following operations can be carried out.

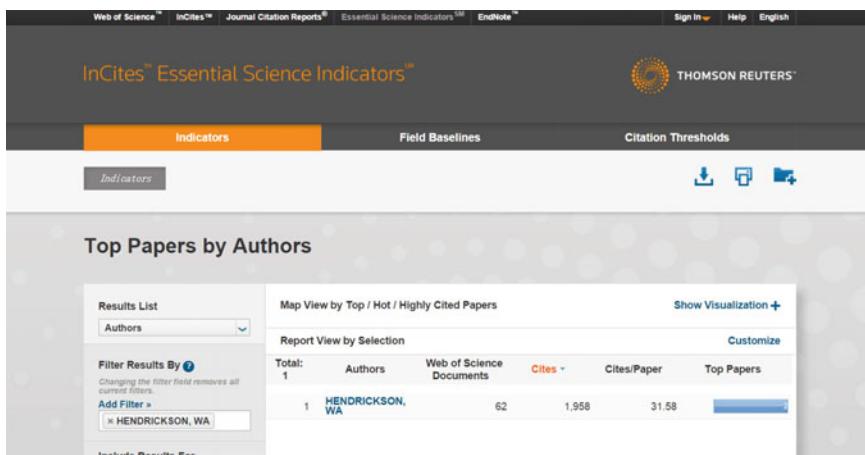


Fig. 8.21 Highly cited papers search page

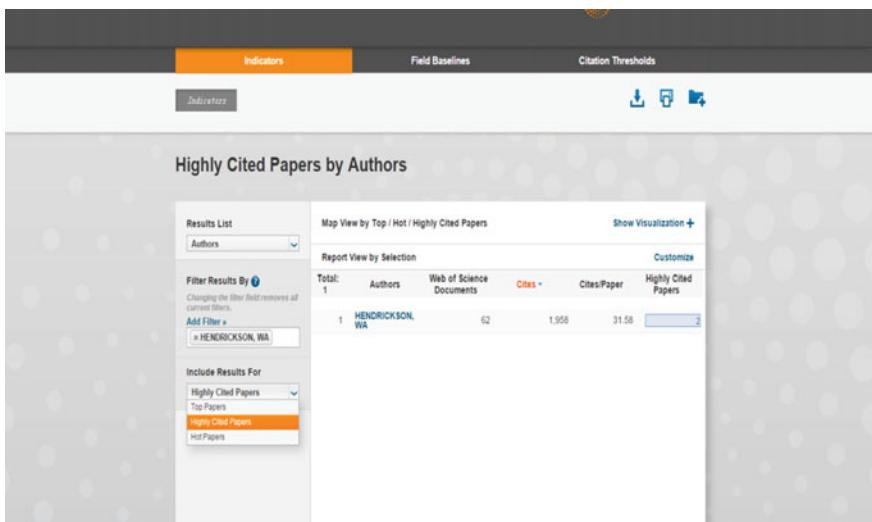


Fig. 8.22 Highly cited papers by searching



Fig. 8.23 Highly cited papers by topic

Step 1 pays attention to the citation frequency of the paper. If the subject areas of the paper are known, it can be returned to the highly cited papers menu in the “by field” column to select the subject, that is, the subject areas of highly cited papers ranking. According to the citations for the page, the user can know the top ranking.

In step 2, if the user cannot determine the subject areas of the paper, he/she can find the journal source in the paper record in the page shown in Fig. 8.26. Clicking the name of the journal with an underline would show the journal science ranking table (Fig. 8.27).

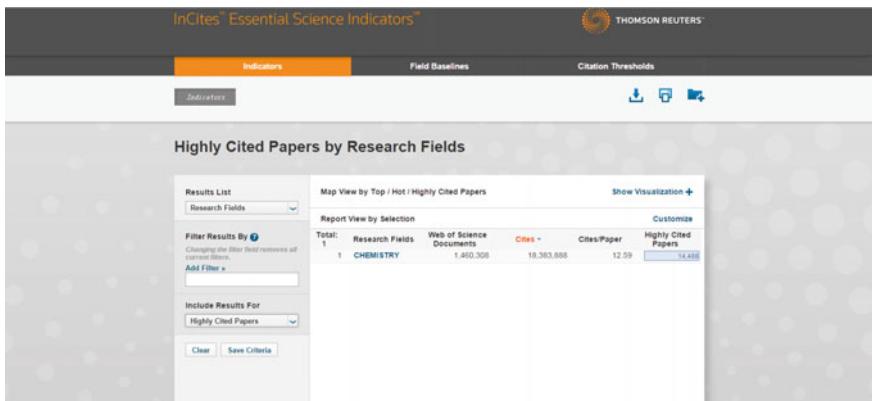


Fig. 8.24 Highly cited papers ranking in the chemical field

Step 3 involves going back to the menu of highly cited papers (Fig. 8.19) and selecting the subject in the “by field” column, which shows the list of subjects in the highly cited papers. The navigation buttons can be used according to the total number of citations to find the location of the paper. The frequency of papers is the cited ranking in the field of chemistry. If the citation frequency of the paper is 169 times, which can be found in the navigation buttons on the page as shown in Fig. 8.24, it is ranked in the field of chemistry for 918.

- Research frontiers

The research frontiers of the retrieval method and citation ranking are roughly the same, but the searching result in the column “FRONTS” is a row from the title of the core of the lottery jargon, as shown in Fig. 8.25. By clicking on the icon, a user can see the list of the core articles that compose the frontier.

④ Retrieval rules

ESI provides logical operators for natural language queries against the subject and key words. A user can input a single word, phrase, or ending with the wildcard “*”. The character does not distinguish between the sizes of the writing, but if non-letters are used, the value of returns is empty (e.g., title retrieval of highly cited papers).

“Apoptosis” is used to retrieve papers whose title contains “apoptosis.”

“Appto” is used to retrieve papers whose title contains apoptosis, apoptotic, etc.

“Monoclonal antibod” is used to retrieve papers whose title contains monoclonal antibody or monoclonal antibodies, etc.

Detailed search rules and names are not introduced one by one.

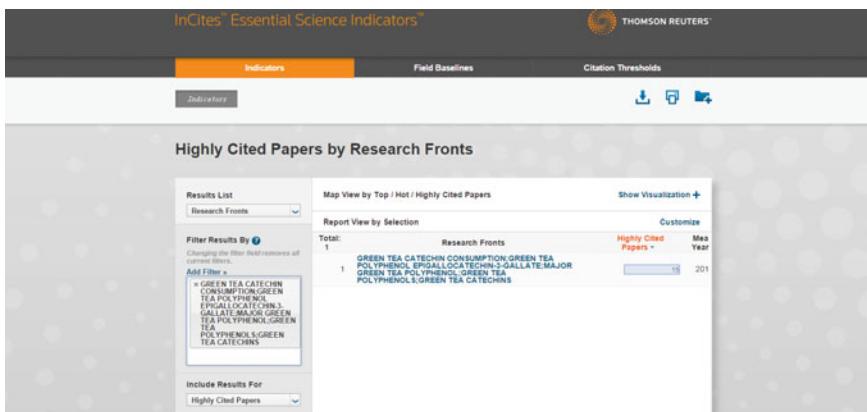


Fig. 8.25 Research fronts of GREEN TEA

2) Evaluation function of ESI

ISI Essential Science Indicators is a comprehensive scientific literature retrieval tool that is based on the network, and it also has a powerful citation analysis and scientific evaluation capability. As a basic scientific measurement and analysis tool, ESI has the following evaluation functions.

- Analysis of the scientific research performance of a company, research institution, state, and periodical
- Tracking the development trend of research in the field of Natural Science and Social Science
- Analyze and evaluate the ability of employees, partners, critics, and competitors
- Determination of output and influence of scientific research results in a professional research field

In addition, ISI Essential Science Indicators can also carry on research on the discipline structure, discipline relations, and science history. ESI provides a link to the article, which is from the scope of journals within the subject, and the correlation of the subject reflected by cited literature. The structure of the subject can then be understood. It reflects the contact between different subjects from different disciplines and the literature periodical net and chain relation. It can also show the occurrence and development of a project or event by the research frontier and suggests ideas or methods to improve, expand, and correct so as to understand the frontier problems in various disciplines, obtain a complete history of scientific development, and predict the direction of future development and hot problems.

ISI Essential Science Indicators can provide an important data source for scientific evaluation to help answer various questions, such as “which papers are cited the most in the field of immunology,” “what new research areas are emerging in the field of agricultural science,” “which countries are the most influential in the field of chemistry research,” “who are the most cited researchers in the field of molecular biology,” and

“what are the most influential journals in the field of Earth Science.” The answer to these questions is undoubtedly one of the important indicators of scientific evaluation.

The evaluation functions provided by ESI are for those who are mainly engaged in scientific and technological or economic policy in government departments, management personnel engaged in scientific research projects in universities, companies, government departments, private laboratories engaged in scientific research project management, research experts engaged in monitoring and evaluation of scientific, and scientific publishing department staff, university, or company recruiters. The database of ESI not only provides much practical scientific data for scientific evaluation, but also can realize the link among ISI Web of Knowledge, ISI Document Solution, and Science Watch.

In conclusion, the evaluation function of ESI is mainly reflected in the following five aspects: evaluation of science research achievements, evaluation of scientific and technological talents, evaluation of scientific research institutions, evaluation of scientific publications, and evaluation of scientific discipline.

(3) Characteristics of ESI

ISI Essential Science Indicators helps researchers determine main scientific discoveries, assess research performance, and grasp the trend of scientific developments. The international academic literature can be systematically targeted and analyzed by ISI Essential Science Indicators because it can provide comprehensive quantitative data, clear and accurate statistics, and other information resources and valuable connections.

The information system of ISI Essential Science Indicators is characterized as follows:

1) Basic analytical tools

ISI Essential Science Indicators can provide a comprehensive analysis of scientific literature, namely, the research achievements of companies, research institutions, and national and academic journals are analyzed to (1) determine the basic trend and direction of scientific and social science research, (2) understand the most advanced countries, periodicals, scientists, papers, and research institutions in various fields, (3) determine the research output and impact in specific research areas, and (4) evaluate potential collaborators, critics, peers, and employees.

2) Wide range of service objects

ISI Essential Science Indicators can provide a timely answer to the complex questions faced by experts in various fields, which include government officials, to formulate scientific, technological, or economic policies. It also serves research management in universities, enterprises, government departments or private research institutions, an analyst or expert in an organization of a government department, enterprises, academic publishing units or academic foundations, an expert on the observation and evaluation of academic studies, and personnel of a university or enterprise.

3) Comprehensive data, papers, and projects

ISI Essential Science Indicators collects and analyzes high-quality academic journals published in research literature by ISI Science Citation Index Expanded TM, ISI Social Science Citation Index. It analyzes data from more than 1,000 million articles included in more than 8,500 kinds of journals in the world and the use of citation analysis in scientometrics research methods and tools. According to the following classification, it provides relevant data for scientists, universities, enterprises, government research institutions, countries, journals, and high citation rates and includes the evaluation of the effect of the benchmark. With the guidance of experts, comments from scientists and researchers are edited. High citation rate of papers, hot topics, research frontier, and Science Watch is also provided.

4) Connection and integration of other information resources

ISI Web of Knowledge SM provides power and support for ISI Essential Science Indicators SM. ISI Web of Knowledge is a new generation of academic information resource integration system based on the Internet (WWW). ISI Web of Knowledge uses the “one stop” information service to build design ideas. By virtue of its unique citation retrieval mechanism and strong cross retrieval function, it can integrate information resources together, which includes academic journals, patent technology, conference proceedings, chemical reactions, research funds, Internet resources, academic analyses and evaluation tools (ISI Essential Science Indicators), and the academic community and its high-quality information resources. It provides high-quality, reliable academic information on natural science, engineering technology, biomedical science, social science, arts, and humanities through knowledge retrieval, extraction, management, function analysis, and evaluation, thereby greatly expanding and deepening the breadth and depth of information retrieval. The users of ISI Essential Science Indicators can be directly connected to various information resources and a worldwide variety of academic fields, including origin, change, dynamics, trends, and applications through the platforms of ISI Web of Knowledge, such as ISI Web of Science, ISI Proceedings, Derwent Innovations Index, BIOSIS Previews, ISI Current Contents Connect, ISI Highly cited.com, ISI Chemistry, ISI eSearch, ISI Document Solutions, and the NCBI GenBank database.

Given that ESI was born 40 years later than SCI, the company of ISI has provided a more perfect data service by intelligence experts and computer experts through 40 years, so it is more convenient and simple to use the evaluation function of the company. ESI has the general evaluation index of SCI, such as the number of papers and independent variable. It has a wider range of evaluation and services provided by the in-depth processing of products, not just like SCI in which the user retrieves the data and then analyzes the data, which contain the service. This clearly indicates that ESI is the integration of the retrieval function and the evaluation function based on the results of the evaluation of the basis of the presentation to evaluate the fundamental purpose.

The publication and distribution of this large index tool, to a certain extent, provide a large amount of data required for citation analysis, which is powerful tool

of quantitative evaluation. Other countries and regions have also made use of quantitative indicators to carry out university rankings, scientific research evaluations, and other activities. In fact, quantitative evaluation of scientific research has become the common practice and general trend worldwide. Under the international trend, China has gradually attached importance to the application of quantitative indicators in scientific evaluation and has carried out bold explorations and useful attempts. In recent years, SCI has been widely used in China, and it has become one of the important indexes in the evaluation of scientific projects and scientific research. Its influence continues to expand and play a guiding role in scientific research and teaching.

8.2.3 Main Tools of Domestic Citation Analysis

Five documents and information units are used to develop and publish citation database products issued in the form of printed publications or databases, etc.

(1) Chinese Science Citation Database

The Chinese Science Citation Database (CSCD) is funded by the Chinese Academy of Sciences and the National Natural Science Foundation of China. It was founded in 1989 by the Chinese Academy of Sciences, and its websites was built in 2002. It exerts a huge impact in China. CSCD has been appointed as the database of the specific query library of the Chinese Academy of Sciences, the fourth Young Scientist Award Committee, the National Natural Science Foundation of China, the National Key Laboratory of Science, and the University of Science and Research. Thus, it has authority and is known as “China’s SCI.”

CSCD, the latest version of which was released in 2002, contains in 1,000 Chinese and English scientific and technological core journals and outstanding periodicals in the fields of mathematics, physics, chemistry, astronomy, geography, biology, agriculture and forestry science, medicine, health, engineering, environmental science, and management science. The core library course journals have 670 types, and the expansion of library periodicals have 378 kinds, which have accumulated nearly 10 million paper records from 1989 to the present and citations of nearly 400 million. It is rich in content, scientific structure, and accurate data. The system provides a new index, the citation index, so that the user can quickly query the details of a science and technology literature cited from millions of citations by using the function besides the general retrieval function. A user can also retrieve relevant literature from early important documents by using the name of the author, which has a very important reference value for cross discipline and research development. The scientometrics indicators database in China, which was derived from the Chinese Science Citation Database 1, is a powerful tool for scientific literature measurement and citation analysis.

CSCD can query the conditions of monographs, journal articles, conference papers, patents, and other informal publications; the citation of scientific journals and published information papers as well as the special subject literature can be queried.

The network version of CSCD was developed in 2002. This version provides a unified service with the China Science Literature Database and Chinese Scientific Literature Catalogue Database integrated for China's science literature database service system. The network of CSCD provides two ways of searching, namely, source literature and citation search.

CSCD and the China Academic Journal Electronic Magazine of Tsinghua University established the Chinese Academy of Scientific Measurement of Evaluation in June 1999.

The database is currently available online and provides free query service. It can be used to query an author and a journal cited by other journal papers, and the results can be used to generate the citation frequency and impact factor of the journals by CSCD. The website is <http://sdb.csdl.ac.cn/index.jsp>.

(2) Chinese Science and Technology Paper and Citation Database

The Chinese Science and Technology Paper and Citation Database (CSTPC) is a database with special functions. It is based on the statistical analysis of scientific and technological papers in China and was created and developed by China Science and Technology Information Institute. The scope of collection contains Chinese Journal of Natural Science statistics and the main social science and relates to all majors in the fields of natural science. It is an important basis for scientific and technical personnel to discover relevant references and is an important tool for scientific and technological management at all levels and in scientific research institutions, universities, and colleges to understand the national units and departments.

The main functions of the database are (1) to identify the important scientific papers published domestically, (2) to learn about the statistical analysis and ranking of scientific papers in China over the years, (3) to know about the details of papers published in various regions, departments, units, authors, various disciplines, and funds, and (4) to carry out a citation analysis of scientific and technological papers.

(3) Chinese Social Sciences Citation Index

The Chinese Social Sciences Citation Index (CSSCI) was built in 1998 by Nanjing University of Chinese Academy of Social Sciences Research Evaluation Center and Hong Kong University of Science and Technology. Its network version was built in the same year; it is an important research project of the state and Ministry of Education and an important tool for information inquiry and evaluation of the main documents and information in China. It addresses the blanks in the Social Science Citation Index in China. CSSCI has selected 419 kinds of humanities and social science academic journals, 17 kinds of journals published in China, and more than 5,700 papers, which can be retrieved and cited. The annual update of disc data began in 1998. Data in 2001 was accessed on the Internet in November 2003, which refers to the practice of SCI and CSCD.

The main functions of CSSCI are as follows:

- Study of humanities and social science research by CSSCI

CSSCI provides information to users from two aspects: the source of the document and the cited literature. It also provides relevant situations of certain papers and convenience for personnel research.

- Evaluation and management of social science research by CSSCI

The collected journals are ranked according to their impact factor and based on the qualitative evaluation of famous domestic experts.

- Evaluation and management of human and social science periodicals by using CSSCI

The system of CSSCI can provide a variety of quantitative data, which can be used to evaluate the academic influence and status of journals.

(4) Chinese Humanities and Social Science Citation Database

The Chinese Humanities and Social Science Citation Database developed by the Chinese Academy of Social Sciences, which collected 34 million records of academic literature from 1999 to 2001 via its database version 2002, contains 1.2 million citation records that cover all fields of humanities and social science research. It collects more than 600 kinds of source journals and updates annually.

(5) Chinese Citation Database

The Chinese Citation Database (CCD) is published by the China Academic Journal. Its collection is from Chinese academic journals of the electronic magazine published by the source database products in literature and references. Examples include the world's largest continuous dynamic update of CNKI, China journal full text database, China's outstanding master's degree thesis database, Chinese important conference papers, China's important newspaper database, Chinese important dissertations full-text database, and China's yearbook full-text database. It realizes the links between literature citation and cited document of CNKI, which is about periodicals, books, papers, magazines. It has reached more than 36 million articles. With the expansion of digital resources, CCD's literature type and quantity in CNKI are expected to increase, and the various types of links and cited references will also increase accordingly. The data in CCD of the CNKI network are updated daily.

8.3 Distribution Law of Citation and Key Indicators of Analysis

Although a science citation is cited by many authors, it is derived from different journals or different literature and is largely influenced by human factors; thus, it exhibits high randomness. However, a scientific citation has a certain distribution

structure and regularity. It is an important content of informetrics to study the distribution theory and law of scientific citation.

8.3.1 Citation Structure and Its Significance

Scientific literature is the objective record of science and technology development and is a reflection of the structure of a scientific and technological system. Scientific literature citation reflects the mutual connection between science and technology and the connection of disciplines from the perspective of the literature used. Owing to the impact of science and technology system structure, the number of various types of individual scientific citations also forms the corresponding distribution structure. Many studies have shown that citation and the cited phenomenon between scientific papers quantitatively embody the scientific inheritance and development of longitudinal and transverse relations in the differences between various disciplines and contacts. The formed citation relations between scientific paper citation link and citation network make science traverse across time and space, and a complete network system is formed from small pieces of specific research topic related to the discipline. The system structure of scientific citation can be described from different elements. Science citation has several basic elements, such as the type of citation literature, subject content, language, age, and citation source. The distribution of science citation basically has the following types according to the different elements and indexes to be described.

- (1) According to the frequency distribution of citation quantity
- (2) Distribution of citations according to age
- (3) Distribution of citation by subject or topic
- (4) Distribution of citation by literature type
- (5) Distribution of citation by language
- (6) Distribution of citation by country
- (7) Distribution of citation by author
- (8) Distribution of citation by journals

8.3.2 Distribution Law of Citation Quantity

Citation quantity is an object that contains the number of references. It is one of the basic characteristics of the citation chain. It not only reveals the mutual contact of the citation and the cited, but can also reflect the contact strength of the object from the quantitative point of view. If a large number of citations exist in two papers or two journals, the intensity of the citations between them reveals that the contact is closer. Therefore, it is an important content and way to reveal the law of scientific

literature citation. From the present study, the distribution of citation quantity can be analyzed from the following aspects.

(1) Theoretical distribution of citation

Analyzing and comparing the citation data of a certain number of papers would reveal that the variation law is displayed by average as the midpoint. The closer to the midpoint frequency, the larger the decrease is in the average frequency, and the normal distribution of the intermediate height and low polarization is formed. If the frequency distribution is asymmetric, the theoretical distribution is asymmetrical. For example, the frequency distribution of citations in agricultural scientific research papers has a positive distribution. If the average number of citations in a research object is difficult to obtain directly from the statistic, the mathematical statistics method can be adopted to achieve the desired purpose by using the average number of samples to assess the overall average within a reliable range.

(2) Citation number of distribution

Citation number of distribution is the distribution of the average number of references cited by each research paper. It not only reflects the breadth and depth of the citation of the author's paper, but also shows the intensity of the relationship between the citation and cited literature. Therefore, the distribution of citations is an important aspect of the distribution structure. When a citation analysis is carried out, the analysis of citation data distribution is also generally carried out.

The citation analysis shows that a research paper's citation of 5–15 has the highest frequency; the distribution of citation below 5 or more than 15 is gradually reduced. According to statistics of foreign countries, about 90% of journal articles listed the cited literature, and each paper was cited by an average of 15 references, of which about 12 are articles from journals.

The number of citations in a scientific research paper depends on the principle of "the need to set the lead." Too many citations may not be able to make the new science and technology information outstanding and too few citations cannot provide sufficient citation clues. It exerts a great influence on the number of citations for the information content and redundancy of literature. The distribution of citation number is influenced by many factors, which are mainly influenced by the following factors.

- 1) The number distribution of citation is associated with the discipline nature of the paper.

Through the citation analysis, it can be known that the number of citations in scientific literature is greater than that of research papers in our country, as shown in Tables 8.3 and 8.4. "Chinese Science Bulletin" has an average citation quantity of 6.3 in the statistics of 7 years, and 17 kinds of industry, such as forging machinery and professional technical journal, have an average citation of 2.8. The former is 2.3 times the latter. This shows that the theory construction of research literature in China is stronger. The scientific and technical personnel engaged in this kind of

Table 8.3 Average citation amount of papers in Chinese Science Bulletin

Project	1973	1974	1975	1976	1977	1978	1979	Total
Paper quantity(article)	51	64	82	53	7.9	153	299	781
Citation quantity(article)	216	473	504	324	573	1153	1738	4945
Average citation(article/paper)	4.2	7.3	6.1	6.1	7.2	5.8	5.8	6.3

Table 8.4 Forging machinery and 17 other kinds of journal articles average citation quantity

Project	1973	1974	1975	1976	1977	1978	1979	Total
Paper quantity(article)	518	460	584	570	513	570	567	3782
Citation quantity(article)	729	1248	1360	1444	1482	1950	2540	10753
Average citation(article/paper)	1.4	2.7	2.3	2.5	2.8	3.4	4.4	2.8

research are relatively good at using literature, and the quantity of theoretical research papers is basically stable (generally between 6 and 7 in each paper). The quantity of technical journal citation is relatively less, but there is an increasing trend, especially rapid growth in the past 2 years. The increase in the average citation amount of scientific journals in China reflects the circulation and utilization efficiency of science and technology literature in China. It also shows that the number of scientific and technological personnel engaged in the development and application of research to strengthen the concept of information, improve the awareness of intelligence, and use science and technology literature is large.

2) The distribution of citation is related to the language of the paper.

In general, the average citation quantity in foreign literature is higher than the number of each paper citation in Chinese. For example, the current sampling of agricultural scientific research paper's citation statistics shows that 100 foreign papers have a citation of 1604; the most citation has 42 articles, and the least has 3 articles. The average citation of each paper amount is 16 articles, of which 13–17 articles account for 35% and 18–22 articles account for 22%. In 100 Chinese papers, the citation is 1055; the most has 39 articles, and the least has 2 articles. The average citation of each paper amount is 10 articles, of which 7–11 articles account for 37% and 2–6 articles account for 31%. Obviously, the number of foreign citations per paper is more than 6 articles of the domestic paper on the average. This reflects the gap between domestic and foreign authors in terms of literature use.

3) The distribution of citation is related to human factors.

Some scientific and technical personnel have access to some scientific and technological literature in a research, and they can obtain inspiration from them. However, it is not reflected in the citation literature in writing papers but appears in the reference literature of the number of citations (generally only 1/3–1/2 of the total literature).

8.3.3 *Garfield's Law of Citation Concentration*

Many studies have shown that the distribution of science citation has the characteristics of centralization and discrete. The concentration and dispersion of citation distribution are relative to several measures. The distribution of citation is the trend of concentration and dispersion by year, language, literature, and so on. We mainly discuss citations by the distribution of journals' source and the frequency on the center with the average distribution of concentration and discrete.

(1) Garfield's law of citation concentration

The citation of Garfield in the United States shows that 75% of all references in the database of Scientific Citation Index are from less than 1000 kinds of cited journals; 70% of the citations published in 500 journals are cited in SCI.

Half of the 3,850,000 citations published in some years are only published in 250 journals. By contrast, the other half are scattered in more than 2,000 kinds of journals. Garfield found that a discipline of non-journals is largely composed of other disciplines of core journals. He thought that virtually all disciplines of core journals would not be more than 1,000 kinds, perhaps even less than 500. This is the law of Garfield's citation concentration.

Analysis of Chinese Science Citation also draws the same result. The distribution of citations in our country is not only in accordance with the law, but is also more obvious. Statistics show that the Chinese Science Citation Index is less than 3% of the total number of citations, and 25% of cited journals account for 90% of the total citation quantity; in addition, 75% of the total number of citations is 72 cited journals.

(2) Concentration and dispersion of citation in different journals

According to Brad Ford's law, there are three sources for a scientist to refer and read a journal article. A third of articles comes from a set of core journals of the discipline, 1/3 is from another set of journals, which primarily consist of the discipline of non-core journals, and the remaining 1/3 is from another group of journals in the field of this subject. If the three kinds of journals are called class A, class B, and class C, it is obvious that the citation of a subject mainly comes from a few of the journals, and the other part is distributed in the number of other journals. At the same time, the distribution of class C journals can be reflected from the citation relationship among the subjects for a subject. The distribution range of a subject in class C journals can be determined through interdisciplinary journal citation times and the citation coefficient of each subject. In general, the natural science interdisciplinary journal of class C is mainly composed of comprehensive and technical journals.

(3) Concentration and discrete trend of citation

The citation of each paper held by the determination of the citation mean and standard deviation (S) can reflect a subject's average citation amount of

concentration and discrete tendency. For example, a sample survey of citations in the current agricultural science paper shows that the average number of Chinese citations is 10.55 per article. The average number of foreign citations is 16.04, which represents the trend of group information to a certain extent. However, the average number of such gains is the real representative concentration of the group. Further determining its discrete trends is necessary. The discrete trend is smaller, and the average central tendency is more accurate. The discrete trend is the size of the dispersion, generally with standard deviation (S).

In practical work, the average number (X) and standard deviation (S) of the simplest method use the frequency distribution table by calculating the grade difference. In Chinese journals ($x = 10.4$ and $S = 6.75$). In foreign periodicals, ($x = 16.2$ and $S = 6.82$). S is expressed absolutely from the size of the potential, and it has a relationship with the size of the average. If the average is different, the relative dispersion needs to be calculated. The relative size of the potential variation coefficient is expressed by V . Chinese papers' citation volume $V = s/\bar{x} \times 100\% = 64.56\%$. Furthermore, the amount of citations in foreign citation is $V = 42.1\%$.

From the viewpoint of standard deviation, Chinese and foreign periodicals' paper citation dispersion size is consistent, and the average number of representatives is good in both Chinese and foreign citations. However, from the viewpoint of the variation coefficient of V , the quantity of Chinese is more than English, which means foreign paper citation quantity is relatively far from that of Chinese papers. This is because the amount of citation in foreign languages is larger than that in Chinese. To a certain extent, it reflects the level of foreign papers in research work and can provide the amount of documents, which expresses a certain advantage to the Chinese language.

8.3.4 Analysis of the Main Index of Citation Measures

Scientific citation index analysis is very significant in improving literature information work and management and increasing the level of literature information quantitative research. The citation index analysis includes the date of quotation, the language of the quotation, the type of citation, the country of the quotation, the author (especially the master), and the citation analysis of classic work. Analysis of several major indexes is conducted in the following:

(1) Analysis of citation age

Analysis of the distribution law of citation according to time is one of the main contents of citation analysis. It can reflect publication, dissemination, and utilization, especially in literature aging and research on this field. The analysis of the distribution of citation time is widely used and is an effective method.

As early as 1965, Price put forward the “maximum citation years” from a large number of statistical analyses and pointed out that the peak of a cited article is in the second year after the article is published. That is, most of the citation of literature published in a year is from the previous two years. The maximum number of citations reflects the most active and vital period of scientific literature. It not only has a significant effect on the determination of important parameters in theory research on literature information, but also determines effectively the period to remove old literature. To make the rate of literature utilization reach the best value, the determination of important parameters should play a guiding role for the publication and distribution of literature. Therefore, this research topic has attracted many scholars, who put forward some amendments. For example, the former Soviet Union scholar GuoKeTatekefu proposed that the literature cited peak is about 2–4 years after publication. In different periods and different academic environments, the discipline of literature of “maximum age of citation” also differs. We should use the development point of view to observe the conclusion of Price.

Many studies showed that the distribution of citation presents certain regularity over time. In general, the citation quantity has increased annually from the past to the present; the time is closer, and the cited literature is more.

A citation time distribution curve can be obtained if the citation time is placed on the horizontal axis, the amount of citation on the vertical axis, and the data points of each year are depicted and then connected by a line. Through an analysis of the reasons that cause the change in curve curvature, we can not only understand the spread of literature utilization, but can also study the process and law of scientific development.

Through an analysis of the distribution curve of citations, we can roughly determine the period of using the cited literature. The average time of using Chinese cited literature from publication is about half a year, whereas that for foreign literature is about 2 years. The best time to be cited in Chinese scientific literature is roughly 2–5 years after publication, whereas that for foreign language literature is about 3–8 years. Most of the citations used by scientific workers are published in the last 10–20 years, and the literature of 20 years ago is rarely used.

According to the above conclusion, we can use the best years of literature to determine the literature service mode and save the fixed number of year so as to provide a quantitative basis for scientific management of literature. It also reflects the level of scientific development from one side. This is because the citation quantity is largely restricted by the literature source, and the amount of scientific papers published reflects the status of the science itself from a point of view. Therefore, studying the publication and exchange of scientific literature and the progress of research by analyzing the distribution of citation is important.

The distribution of citation time is influenced by many factors, such as the discipline nature, type of literature, literature language, quality of literature service, and several human social factors. On the whole, the average citation of a newly emerging discipline and a marginal discipline is longer than that of the subjects, that of theoretical literature is longer than that of scientific literature, and the original

Table 8.5 Distribution of science citation time

Group	Annual interval	Frequency of Chinese	Frequency (%)	Frequency of foreign	Frequency
1	<1930	10	0.9	9	0.6
2	1930–34	4	0.4	7	0.4
3	1934–39	7	0.7	13	0.8
4	1940–44	9	0.9	18	1.1
5	1945–49	1	1.0	36	2.3
6	1950–54	24	2.3	39	2.4
7	1955–59	98	9.3	64	4.0
8	1960–64	64	15.5	148	9.3
9	1965–69	121	11.5	258	16.1
10	1970–74	185	17.5	424	26.4
11	1975–79	415	39.3	568	35.4
12	>1979	7	0.7	20	1.2

research literature is longer than the others. The citation time of the monograph is long, and the journal papers are more easily out of date (Table 8.5).

In the distribution of citations, the difference between domestic and foreign literature is a notable feature. In short, Chinese literature in the citation is concentrated in recent years and account for 70–80% of literature from nearly 5 years. Foreign literature is relatively concentrated, but the decline is relatively stable. The proportion of Chinese literature was very small 20 years ago, and foreign language still has a considerable proportion. The citation used to determine the half-life of several subjects in Chinese literature is about 2 years, whereas that for foreign literature is 7–8 years; the gap is very large. The distribution of citation has a guiding significance for literature collection, removing old literature, determination of the service mode of foreign literature, etc. (Table 8.6)

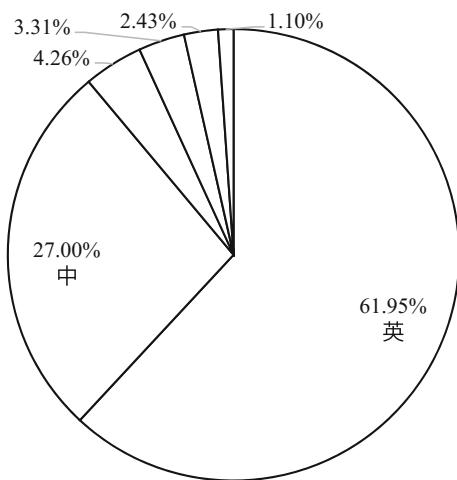
(2) Analysis of citation language

The cited literature is composed of different languages. When more than one language of literature is cited, the more important and commonly used the cited literature is. It is valuable to investigate and analyze the distribution of citation literature, especially for people to introduce foreign literature, the selected topic, the education of foreign language, etc. A statistical analysis is conducted on 3054

Table 8.6 Comparison of three kinds of Chinese and foreign literature

Language	Physics	Automation technology and computer science	Chinese Science
Chinese	1.8	2.0	2.2
Foreign language	7	7.3	8.5
total	6.1	6.3	7.1

Fig. 8.26 Use of language situation in Chinese chemists



英文61.95%，中文27.00%，俄文4.26%
日文3.13%，德文2.43%，其他文种1.1%

citations of 10 kinds of journals published in 1979 for 367 articles, which are from “atmosphere science” and “metal heat treatment.” It shows that Chinese literature accounts for 27% of the total citations, and English literature accounts for 58%. This means that it is sufficient to access the required 85% of documents as long as they are master English and Chinese. The percentage of English citations is particularly high. On the one hand, a number of academic papers are published in English; on the other hand, it shows that most scientific workers use English to retrieve foreign information as a tool in our country. In addition, the Department of Library and Information subscribes to English books more than books in other languages.

Statistics show that a large proportion is occupied by English, Chinese, and Russian in the cited literature. The language of English is still the most used language of scientific workers in our country and will become the dominant trend (Fig. 8.26).

At the same time, the distribution of citation language is not the same for different subjects and majors. For example, there is no Japanese literature in the 435 citations of 61 papers from the Journal of Mathematics. By contrast, the citation of German and French occupies a certain proportion. The largest proportion is held by Chinese (as much as 47%) in the 442 citations of 41 articles from Environmental Science.

(3) Analyzing the type of citation literature

The literature cited in scientific research is very broad and includes journals, books, and special literature. Science Bulletin is a comprehensive academic journal in China. It is published in various disciplines or professional papers in theory research with a wide representation. Therefore, it is appropriate to analyze the

Table 8.7 Distribution of citation type in the Science Bulletin

Document type	1973	1974	1975	1976	1977	1978	1979	Total	Percent
Books	37	72	69	61	77	206	354	876	18
Journals	170	351	408	256	454	816	1292	3807	76.8
Special documents	3	1	4	5	16	10	24	63	1.2
Chinese internal data	6	12	23	2	29	61	65	199	4

citation of literature in the Science Bulletin. The special literature that appears in the Science Bulletin includes patent specification, scientific and technical reports, conference documents, technical standards, product samples and catalogues, and degree papers. The distribution of citation type is shown in Table 8.7.

The figures in the table show that the journal offers abundant information and reflects the dynamic of scientific research in a timely manner. In addition, the journal papers account for 76.8% of the entire citation. Certainly, the journal will always be the first source of scientific and technical workers in China in the present condition of our country and in the next period of time. The citation rate of special literature is very low, which shows that the publicity and service about this kind of literature are weak. Actually, it shows that content update should be faster for conference proceedings and technological reports to strengthen the collection and service of this kind of literature as the information department, which is a very important job.

In general, the proportion of the journal paper is the largest, followed by books. It is cited in recent years by the rising trend for patent specification, scientific reports, technical reports, technical standards, product samples, and thesis.

(4) Citation analysis for the country

As a result of the need for scientific research, scientific and technological workers who are in any country will inevitably cite the scientific literature of other countries. The distribution of citation in the country is formed. The country of citation analysis, especially the statistical analysis of the mutual citation of the literature, can prove the status of each country and determine the quantity and direction of the international literature exchange. It is important for us to study the level of scientific development and technical strength of various countries, formulate reasonable technology import policies, and improve our comprehensive competitiveness.

For citation analysis based on country, we can determine the literature exchange ratio and the cited deviation value of reference to carry out an in-depth study for literature of the world. The definition of the exchange ratio between A and B is

a = the total number of citation in country A that cited the references of country

B/the total number of citation in country B that cited the references of country A

(8.1)

If $a > 1.0$, then country A cites more literature of country B. If $a = 1.0$, then countries A and B cite the literature of each other in equal. If $a < 1.0$, then country B cites more literature of country A. The exchange ratio is the relative measure index of different countries in citation with each other.

The total number of articles that are from scientific and technological workers in a period from a country can measure the size of a country's scientific research ability to a certain extent. If scientific research in this country can be determined by the number of citations alone, it can be inferred that the number of documents cited in this country is as large as the number of references cited in other countries.

(5) Citation through analysis of the author

The distribution of citations is an important basis for understanding and evaluating a certain subject of professional technology staff performance. It has an important reference value for objective evaluation institutions and the academic level of scientific research personnel. The distribution curves of Bradford and Lotka are very similar to the distribution of verification in informetrics.

The distribution of citation can be determined for discipline, institutions, etc., which is similar to the methods listed above.

8.3.5 Self-citation Analysis of Scientific Literature

It is sometimes cited as a paper or a book of other authors and sometimes as previously published literature in the author's citation literature. A citation that is limited to use by itself is called a self-citation. Self-citation is one of the most important and common forms of citation. The reason for self-citation is that the author hopes to present work associated with a previous work, which is the performance of the research results and the inheritance. Therefore, self-citation is a literature phenomenon and is one of the basic attributes of scientific literature communication. Similarly, self-citation analysis is an important component of citation analysis. It can reveal the relationship among various countries, various disciplines, all kinds of professionals, various groups, languages, and all kinds of periodicals through the analysis of the special law of the self-citation process, which reflects the progress and dynamics of scientific research and illustrates several trends and laws in the scientific community.

In self-citation analysis, two measures are often used: self-citing rate and self-cited rate. The former refers to the number of certain types of the self-citation proportion in the total citation frequency; the latter refers to the number of certain types of self-cited proportion in the total citation frequency. The specific calculation formula is combined with the type of self-citation.

(1) Self-citation of the subject

A subject or a discipline in the field of literature cited in this discipline or professional literature is called the subject self-citation. The relative stability and absorption ability of the discipline can be evaluated by a statistical analysis of the subject self-citation rate. The self-citation rate of a subject can be expressed by the following formula:

$$\text{Self-citing rate} = \frac{\text{the number of times cited in the subject}}{\text{total number of citations}} \times 100\% \quad (8.2)$$

The magnitude of the self-citation rate can be used as the measure of the degree of scientific independence, stability, and the degree of interaction index between disciplines. It has an important reference value for the collection and management of scientific research.

(2) Self-citation of a nation

A country or a region in the published literature cited in national or regional literature is known as self-citation of the same country or region. National self-cited is to cite the country's literature. The self-cited rate is expressed as follows:

$$\text{Self-cited rate} = \frac{\text{the number of times cited in national literature}}{\text{total number of citation}} \times 100\% \quad (8.3)$$

The status of scholars in this field is analyzed by examining the statistics of the self-citation of a professional document in the same language and the literature from the same country or region. In general, if a country is in the leading position in a certain field, the self-cited rates of the same language literature are higher for the same country.

Regional self-citation and institutional self-citation are similar to self-citation. This measure can display the characteristics of a region and reflects the level of the subject and the continuity of scientific research work.

(3) Journal self-citation

When a paper cites the phenomenon of papers published in the same journal, the source of the journal cited itself is called journal self-citation. The self-citation rate is one of the important indexes to evaluate the quality of a journal.

The self-cited rate of periodicals is calculated in the American Journal Citation Report as follows:

$$\text{Self-citation rate} = \frac{\text{the number of times cited by itself}}{\text{the total number of citation}} \times 100\% \quad (8.4)$$

(4) Author self-citation

An author who cites his previous published papers or co-authored papers is called author self-citation. The self-cited rate is as follows:

$$\text{Self-citation rate} = \frac{\text{the number of papers the author cited his own paper or co-authored paper}}{\text{total number of citation}} \times 100\% \quad (8.5)$$

The author self-citation rate can be used to illustrate the stability of a subject in the field of experts' team and the future trend of development of the discipline.

(5) Self-citation of the same language

A literature that cites the language used by itself is called self-citation of the same language. The formula of the self-citation rate is as follows:

$$\text{Self-citation of the same language} = \frac{\text{the number of citations in the same language}}{\text{the total times of citation}} \times 100\% \quad (8.6)$$

8.4 Citation Analysis of Scientific Journals

A scientific journal reports a large number of papers and their references. The published time of a scientific journal is short, and the content of a scientific journal is new. It can fully reflect the current situation and trend of the development of technology and the exchange of literature. Thus, research on the utilized situation and related influence of scientific journals is an important part of citation analysis. The enormous success of the *Science Citation Index*, which is based on the statistical analysis of journal literature, promoted the publication of Journal Citation Reports (JCR), which is an important complement of SCI. This section discusses the issue of citation analysis of scientific journals.

8.4.1 Decentralization and Centralization Law of Periodical Literature

The previous chapter introduced the bibliometric classic distribution, Bradford's law of scattering, in detail. Bradford's law describes the distribution of journal articles. According to Bradford's law, a scientist needs to refer to papers from three parts: one-third of papers are from a small number of core journals of his discipline, one-third of papers are from a large number of non-core journals of his discipline, and one-third of papers are from a large number of journals of other disciplines. From the viewpoint of published papers, it indicates the distribution trend of scientific literature of scientific journals.

Professor Garfield, an American citation analysis expert, used the data of *Journal Citation Reports* (JCR) as reference and formed the distribution curve of the cumulative amount of citations according to the quantity of cited journals (shown in Fig. 8.27).

Garfield carried out research on the measurement of citations. It indicated that in 1969, 73% of citations in the *Science Citation Index* database were from less than 1,000 kinds of journals; 500 kinds of journals provided 70% of references in SCI. The papers of journals (less than 200 kinds) accounted for more than 50% of cited citation amount in the *Science Citation Index* database. The papers in 2,000 kinds of journals accounted for about 84% of total citations. This scenario indicates that the cited literature are highly concentrated in a small number of core journals. This central tendency is more apparent than the distribution of papers described by Bradford's law. Moreover, Bradford's law often reveals the statistical results of a single discipline or a particular specialty, whereas the citation index reveals the macro result of cited journal literature of various disciplines in all natural sciences. That is, in a broader sense, the citation index confirms Bradford's Law. A similar result was obtained through a citation analysis of Chinese scientific and technical journals (shown in Fig. 8.28).

Half of the citations recorded by the *Chinese Science Citation Index* (CSCI) are only from 3% of total cited journals. Twenty-five percent of total cited journals

Fig. 8.27 Distribution of the cumulative amount of citations according to the quantity of cited journals

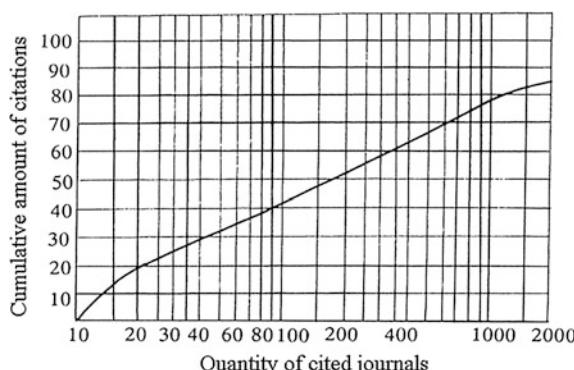
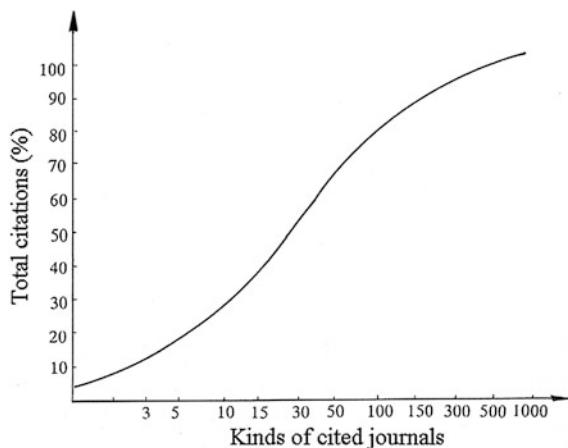


Fig. 8.28 Distribution of Chinese science journal citation



account for 90% of citations, and 75% of total citations are only related to 72 kinds of cited journals.

According to statistical data of many years, Garfield conducted an in-depth research and eventually introduced his famous law, Garfield's law of concentration.

- (1) For the whole of natural science, the sum of core journals of various disciplines is not more than 1,000 kinds and may even be only 500 kinds. The concentration degree for different disciplines may not be the same.

From the distribution curve of cited journals provided by Garfield (shown in Fig. 8.27), the former 1,000 kinds of journals contain approximately 75–80% of the total cited literature, whereas 500 kinds of journals contain nearly 70% of the total cited literature. It shows a very high concentration degree. The main reason for this concentrated trend is that for selecting references, it highlights the willingness and behavior of users compared with the option of collected papers in a journal. Besides, the artificially controlled factor is stronger, and the choice is more stringent such that the result literature become more centralized.

- (2) Any discipline needs tail journals, i.e., non-professional scientific journals described by Bradford's law. Most of them constitute core journals of other disciplines. Given that Bradford's law usually discusses the given professional papers, it can divide journals into core journal area, general journal area, and non-professional journal area according to the rate of published papers. However, Garfield's law discusses the statistics of cited journals in various disciplines of the whole natural science. The journal tail of some specialties is actually included in other professional core journal areas and forms the core and tail stacked result of various discipline journals. Thus, it makes multidisciplinary journal literature more highly concentrated at the macro level. At the same time, it makes core journals play a more significant role and have a larger application range.

Considering that evaluating journals has greater objectivity and practicality according to user feedback information, when determining the core journals, any library and information department has more obvious advantages when Garfield's law is used instead of Bradford's law.

8.4.2 Main Indices of Evaluating Journals

Scientific and technical journals play a very important role in scientific and technical activities and are the main tool of scientific communication. They occupy a very prominent place in scientific literature and provide more than 70% of total scientific and technical information for scientists and experts. Therefore, they are hailed as the most successful scientific information carrier in any place in the entire history of science and are a formal, open, and orderly exchange medium among scientists. For this reason, an objective and comprehensive evaluation is often needed to evaluate the role of scientific and technical journals in the scientific activities and literature exchange and the pros and cons of their quality. Thus, the direction and approach of evaluating journals will be improved and perfected. The citation analysis method is one of the most effective means to achieve this purpose. The method can comprehensively and objectively evaluate scientific and technical journals through various measurable indicators. A brief introduction of several commonly used measurable indicators of evaluating journals is provided below.

(1) Number of articles published

Number of articles published is one of basic indicators to describe the capacity of journals to produce papers. It is defined as the total number of papers published in a journal within a given time. This is an absolute index. According to the statistical need, it may be the total number of papers published in journals, and it could also be the number of papers in a particular discipline or specialty.

(2) Citing rate of journal

Citing rate of journal is one of basic indexes to describe the capacity of journals to absorb external literature. It is defined as the total number of references cited by a journal within a given time. According to the statistical need, it can separately add up the number of multi-discipline cited by journals, the number of multilingual references, and the proportional relation among them to reflect the capability of this journal to comprehensively absorb the information of literature.

(3) Cited rate of journal

Cited rate of journal is one of basic indexes to evaluate journals from the angle of users. It has a distinctive feature of citation analysis. It is defined as the total number of cited journals within a given time. This index directly reflects the role of journals in the development of science and exchange of literature based on the objective use

number. Generally, the greater its value is, the more important the role it plays. However, because this index uses an absolute number, sometimes it is difficult to evaluate accurately. For example, journals with many published papers often easily obtain a higher citation number than journals with a few published papers. To avoid this bias, the relative number of index is often introduced to unify the evaluation criteria.

(4) Mean citation rate

The mean citation rate reflects the statistics of journals' mean citation level to correct the bias caused by the different amounts of published papers. According to the angle of citation journal and cited journal, there are two different mean citation rates.

- 1) Mean citation rate. Within a given time, the number of references cited by journals is divided by the corresponding number of articles published. It represents the mean level of each paper citing references. The mean citation rate generally reflects the capacity of journals to absorb information, academic level, and scientific exchange degree. It can not only evaluate journals, but can also evaluate authors, disciplines, countries, regions, institutions, and so on.
- 2) Mean cited rate. Within a given time, the number of articles in a journal cited by other journals is divided by the total number of articles published in this journal. The higher the mean cited rate is, the higher the academic level it generally has. However, in this definition, the selection of time interval is not clearly defined. As time goes on, the cited rate of literature declines. How to select a time period is the most accurately reflected cited level of journals that is to be discussed with two important measurement indexes, i.e., impact factor and immediacy index.

(5) Impact factor

In 1972, Garfield proposed the impact factor (IF). IF is an important index of evaluating journals. The index is a relative number of indicators and mainly used to adjust and amend the situation that large and old journals have the advantage of the absolute number of published papers in cited journals. At the same time, it selects the number peak of a cited journal to calculate its mean cited rate. It can suitably reflect the real situation of journal utilization. Price proposed that scientific papers will be accepted and understood and will reach the cited peak phase when they are published 1–2 years later. Garfield defined the calculation formula of IF based on this idea.

$$\text{Impact factor} = \frac{\text{Cited times in this year of papers published two years ago in a journal}}{\text{Total number of papers published two years ago in this journal}} \quad (8.7)$$

For example, the IF of a journal in 2005 is equivalent to the citation number of papers in a journal (published from 2003 to 2004) cited by all the source journals in

Science Citation Index divided by the total number of papers published in this journal from 2003 to 2004. Generally, if IF is large, then this journal plays a large role and exerts much influence on the development of science and literature exchange. This journal also has better quality.

The *Journal Citation Reports* compiled by the American Social Science Citation Index regularly publishes the impact index of more than 3,800 papers collected by SCI to provide convenience for people to evaluate the above journals. In recent years, China conducted a related research. The Institute of Scientific and Technical Information of China began publishing a variety of measurable indexes of Chinese scientific and technological papers' source journals from 1997. It provides the conditions of evaluating journals for Chinese research management institutions at all levels (shown in Table 8.9).

In recent years, with in-depth citation analysis, researchers have begun to realize that the citation peak area of many scientific and technical journals is not always concentrated in 1–2 years after the papers have been published. The citation peak area of several disciplines or regional journals may be extended to the third year, the fourth year, etc. Thus, several researchers proposed a more general definition of IF.

$$\text{Impact factor} = \frac{\sum_{i=1}^m n_{k-i}}{\sum_{i=1}^m N_{k-i}} (m = 1, 2, 3\dots), \quad (8.8)$$

where n_{k-i} represents the number of citations of published papers in k-i year and N_{k-i} represents the number of published papers in k-i year. According to this formula, the definition of Garfield's IF can be expressed as

$$\text{Impact factor} = n_{k-1} + n_{k-2}/N_{k-1} + N_{k-2}. \quad (8.9)$$

Rousseau, an informetrics scientist, randomly selected 107 kinds of journals in different disciplines or specialties and compared and analyzed the impact index based on $m = 2$, $m = 3$, and $m = 4$. The results showed that if $m = 3$ and $m = 4$, the most calculated IFs are greater than the IFs defined by Garfield ($m = 2$). At the same time, it produces different orders of journal queue sequence. Table 8.8 presents Rousseau's analysis results of several mathematics journals whose m was 2 and 4 in 1985. Readers can see the difference between the two IF calculation formulas.

Therefore, the issue of determining the value of m should be further discussed. In general evaluation, it is appropriate to use Garfield's IF formula.

(6) Immediacy index

The utilization speed or time difference of journals used by users is also an important index to evaluate journals. Users often read high-quality journals and excellent papers and absorb and utilize the knowledge as soon as possible. To

Table 8.8 Sequence of mathematics journals' impact factors according to $m = 2$ and $m = 4$

A	B	C	D
Commun Algebra	1	3	2
P K Ned Akad A Math	2	14	12
Discrete Math	3	8.5	5.5
Nagoya Math J	4	12	8
Math Scand	5	8.5	3.5
B Sci Math	6	7	1
J Math Soc Jpn	7.5	5	2.5
P Am Math Soc	7.5	13	5.5
B Soc Math Fr	9	1	8
J Number Theory	10.5	6	3.5
Q J Math	10.5	2	8.5
Ann Sci Ecole Norm S	12	11	1
Math USSR SB	13	15	2
Can J Math	14	10	4
Stud Math	15	4	11

describe the tendency of using this aspect, Garfield also introduced a new evaluation criterion, i.e., the immediacy index. The immediacy index (cited index during that time) is used to measure the utilization speed of journals. It is also a basis for measuring the importance of journals. It is generally the average number of cited papers during that year of the published papers in a year. With n_k as the total number of cited papers during that year of the published papers in k year, N_k as the total number of published papers in k year, and IMI as the immediacy index, the calculation formula of IMI is

$$IMI_k = n_k/N_k, \quad (8.10)$$

i.e.,

$$\text{Immediacy index} = \frac{\text{the number of cited papers during that year of published papers in a year}}{\text{the number of published paper during that time}}.$$

Simply speaking, the immediacy index is the number of published papers of a journal that are cited during that time divided by the number of published papers during that time. It can be used to characterize the immediate reaction rate of journals. Therefore, sometimes, the immediacy index is also called the current year cited index. Table 8.9 shows the immediacy index of physics journals of Chinese scientific and technical papers of statistical source journals, i.e., the immediacy index and its measurable indexes.

Table 8.9 Chinese scientific and technical journal measurable indexes (fragment)—physics journals

Journal title	Number of papers in 1994	Number of papers in 1995	Paper cited frequency and 1995	Total cited frequency	Number of papers in 1996	Paper cited frequency in 1996	Impact factor	Immediacy index
Progress in Physics	14	20	25	98	44	0	0.7353	0.0000
Applied Laser	86	106	78	161	98	9	0.4063	0.0918
Chinese Journal of Acoustics	55	65	41	205	120	18	0.3417	0.1500
Chinese Journal of Luminescence	64	70	42	94	78	2	0.3134	0.0256
Spectroscopy and Spectral Analysis	147	146	89	271	149	5	0.3038	0.0336
Chinese Journal of Physics	290	277	170	744	295	15	0.2998	0.0508
Chinese Journal of High Pressure Physics	46	47	26	89	48	5	0.2796	0.1042
Acta Optica Sinica	116	108	60	175	141	3	0.2679	0.0213
Chinese Journal of Astronomy and Astrophysics	283	352	170	521	381	40	0.2677	0.1050
High Energy Physics and Nuclear Physics	54	52	28	75	62	5	0.2642	0.0806
Journal of Engineering Thermophysics	169	164	81	218	163	16	0.2432	0.0982
Journal of Applied Acoustics	95	111	48	192	109	6	0.2330	0.0550
Nuclear Fusion and Plasma Physics	57	50	23	90	61	3	0.2150	0.0492
	35	40	16	45	41	1	0.2133	0.0244

(continued)

Table 8.9 (continued)

Journal title	Number of papers in 1994	Number of papers in 1995	Paper cited frequency and 1995	Total cited frequency	Number of papers in 1996	Paper cited frequency in 1996	Impact factor	Immediacy index
High Power Laser and Particle Beams	92	101	41	122	107	7	0.2124	0.0654
Technical Acoustics	25	42	14	49	45	1	0.2090	0.0222
Journal of Atomic and Molecular Physics	74	74	30	76	87	5	0.2027	0.0575
Chinese Journal of Magnetic Resonance	61	78	28	79	82	6	0.2014	0.0732
Chinese Journal of Chemical Physics	91	94	31	98	101	7	0.1676	0.0693
Chinese Journal of Computational Physics	79	89	28	121	81	1	0.1667	0.0123
Acta Photonica Sinica	102	114	35	78	219	17	0.1620	0.0776
Cryogenics & Superconductivity	47	47	15	28	55	1	0.1596	0.0182
CHIN J NUCLEAR PHYSICS	67	69	20	29	42	1	0.1471	0.0238
Chinese Journal of Quantum Electronics	51	69	16	63	80	2	0.1333	0.0250
Journal of Chinese Mass Spectrometry Society	48	59	14	33	54	1	0.1308	0.0185
Acta Mathematica Scientia	79	67	17	82	66	0	0.1164	0.0000
Chinese Journal of Low Temperature Physics	85	78	18	62	74	2	0.1104	0.0270
CHINESE PHYSICS LETTERS	207	200	33	66	256	3	0.0811	0.0117

(continued)

Table 8.9 (continued)

Journal title	Number of papers in 1994	Number of papers in 1995	Paper cited frequency in 1994 and 1995	Total cited frequency	Number of papers in 1996	Paper cited frequency in 1996	Impact factor	Immediacy index
Physical Testing and Chemical Analysis (Part A: Physical Testing)	102	99	16	57	94	0	0.0796	0.0000
Commun Theor Phys	147	155	21	42	166	2	0.0695	0.0120
Image Technology	36	36	4	5	52	0	0.0556	0.0000
Physics Examination and Testing	73	78	4	19	76	0	0.0265	0.0000

(7) Self-citing rate

The self-citing rate refers to the proportion of the number of self-citing and the total number of references.

$$\text{Self-citing rate} = \frac{\text{The number of self-citing}}{\text{The total number of references}} \quad (8.11)$$

(8) Self-cited rate

The self-cited rate refers to the proportion of the number of self-cited and the total number of cited. It is defined as

$$\text{Self-cited rate} = \frac{\text{The number of self-cited}}{\text{The total number of cited}}. \quad (8.12)$$

For example, the cited number of *Amer.J.phys* in 1990 is 1,090, among which the self-citing is 151. Thus, the self-cited rate is 13.8%.

The numerators of the two definitions above are the same, and the denominators are different. Therefore, they describe the situation of self-citing from two angles, i.e., the situation of self-citing and the situation of cited by other journals.

The self-citing rate and self-cited rate can be used to evaluate and judge journals from many aspects.

In general, if the self-citing rate and the self-cited rate are high, then this journal has less professional exchanges and the academic environment is closed. Hence, this journal is professional and independent. If the self-citing rate is high and the self-cited rate is low, then this journal has a high academic status and the selected material range is stable. This journal is often cited by other journals and exchanges sufficiently. Meanwhile, if the self-citing rate and the self-cited rate are low, then the reported range of this journal is large, the content is rich, and this journal is widely cited by other journals. At the same time, using the degree of concentration and decentralization of the self-citing rate and self-cited rate can determine professional journals and general journals.

In conclusion, the statistical analysis of self-citing rate and self-cited rate can help us learn and evaluate the multi-faceted nature of journals from the aspects of disciplinary nature, professional direction, selection scope, exchange degree, and so on. Thus, it is convenient for collecting, managing, and using journals.

The above listed are several measurement indexes that are frequently used in scientific journal citation analysis. Through a statistical analysis, one can evaluate various journals and determine the function, nature, and role of scientific journals to achieve the purpose of scientific evaluation and scientific management.

8.4.3 *Journal Citation Reports (JCR)*

(1) Publishing overview of JCR

After the compiled and published *Science Citation Index* (SCI), the American Institute for Scientific Information (ISI) continuously publishes *Journal Citation Reports* or JCR for short, a new journal citation analysis tool.

JCR was compiled and published in 1975. It is a new part of the annual index of SCI, which was gradually separated into a book and published yearly. JCR is a by-product of SCI. Based on the accumulation of the SCI database in many years, JCR uses the automatic process of a computer to compile the citing and cited relations among specialized journals that have been system classified, collated, analyzed. Thus far, it is an authoritative journal evaluation tool recognized internationally. The first publication of JCR in 1975 was based on SCI data in 1974 and reflected the citing and cited relations among 2,400 kinds of source journals, 40,000 source literature, and 4,248,065 references.

The edition cases of JCR are as follows. First, according to the range of subjects, it can be divided into JCR Science Edition and JCR Social Sciences Edition. JCR Science Edition contains citation analysis evaluation information of more than 5,000 kinds of journals in scientific and technical areas. JCR Social Sciences Edition contains citation analysis evaluation information of more than 1,600 kinds of journals in social scientific areas. Second, according to the form of the carrier, it can be divided into printed edition, CD-ROM edition, and web edition.

JCR is the most important tool to evaluate journals and analyze their citations. By using the statistical data provided by JCR, one will clearly learn the situation of the citing journal and the cited journal, citation frequency, citation network, and self-citing. JCR provides a reliable basis to objectively evaluate scientific journals, easily and quantitatively evaluate the mutual influence and interaction of journals, properly assess the role and status of several journals in the scientific exchange system, determine the core journal group, and so on. Therefore, JCR uses the citation analysis method to conduct research on bibliometrics and scientometrics; it is also a powerful and convenient tool for technical management.

(2) JCR printing plates' layout structure

JCR is composed of five major parts: journal rankings, source data listing, journal half-life listing, citing journal listing, and cited journal listing.

1) Journal rankings

This part is composed of nine small parts. The first part, which is called the journal classification table, is the most important. It is composed of 13 contents. Based on the alphabetical sequence of the abbreviation of the source journal title, journals form the journal classification table (shown in Table 8.10).

The first column is the rank number sequenced by the abbreviation of the journal title. The second column is the abbreviation of the journal title. The third, fourth,

Table 8.10 Fragment of journal rankings (from JCR, Garfield, 1988)

Rank number	Journal title	Number of citations in 1985 and 1986				Number of published papers in 1985 and 1986		Impact index	Number of cited papers in 1987	Number of published papers in 1987	Immediacy index
		Cumulative number of each year	1986	1985	86 + 85	1986	1985				
1	A Van Leeuw JMICROB	809	47	34	81	75	40	115	0.704	2	0.05
2	AAPG Bull	3233	128	235	363	100	131	231	1.571	19	0.112

fifth, and sixth columns sequentially list the situation of papers cited by SCI, SSCI, and A & HCI source literature in a related year. Among them, the third column is the sum of journal cited time in each year. The fourth and fifth columns are the number of citations in this journal 1–2 years ago. The sixth column is the sum of citations in the fourth and fifth columns. The seventh to ninth columns list the situation of published papers, i.e., the produced information. Among them, the seventh and eighth columns are the number of published papers in this journal 1–2 years ago. The ninth column is the sum of papers in the seventh and eighth columns. Based on these data, the tenth column calculates this journal's IF in the current year. The eleventh and twelfth columns are the citing situation of this paper in the current year. Specifically, the eleventh column is the cited number of this paper in the current year, and the twelfth column is the number of published papers in this journal in the current year. Based on this, the thirteenth column calculates the immediacy index of this journal.

The content information of the second part to the sixth part is similar to that of the first part. However, the sequence arrangement is different so as to investigate the situation of journals' sequence from different angles. The second part is sorted by the total number of cited journal over the years, the third part is sorted by IF, the fourth part is sorted by the immediacy index, the fifth part is sorted by the number of cited papers published in the current year, and the sixth part is sorted by the number of cited papers published 1–2 years ago.

The seventh part is the social science journal classification table and is sorted by the abbreviation of the journal title. The content of this part is similar to that of the first part.

The eighth part is the classification of SCI source journals based on subject category. Each category is sorted by the value of IF and gives the cited half-life of journals (according to the definition of Burton's half-life).

The ninth part is the alphabetical table of all source journals and the subject category table used in the eighth part. According to the table of this part, users can quickly determine the category of journals.

2) Source data listing

The second part of JCR is source data listing. The sequence of source data listing is based on the source journal title. It lists the related data of published papers, including:

- ① Number of published papers in each journal in a statistical year,
- ② Number of citing references in the above papers,
- Average number of citing references per paper.

The above data of review papers, non-review papers, and the sum of both are listed in Table 8.11.

Table 8.11 Data fragment of source journal (from JCR, Garfield, 1988)

Journal title	Non-review papers			Review papers			Sum of non-review and review papers		
	Source paper number(S)	Reference number(R)	R/S	Source paper number(S)	Reference number(R)	R/S	Source paper number(S)	Reference number(R)	R/S
J Antimicrob Chemoth	245	4192	17.1	5	259	51.8	250	4451	17.8

Table 8.12 Journal half-life listing (the first part) (from JCR, Garfield, 1988)

Citing half life	Citing journal	1987	1986	1985	1984	1983
4.9	Laser Surg Med	0.84	8.09	24.13	38.16	50.11
1982	1981	1980	1979	1978		
57.82	64.22	68.46	72.08	75.94		

Table 8.13 Journal half-life listing (the second part) (from JCR, Garfield, 1988)

Cited half life	Cited journal	1987	1986	1985	1984	1983
2.9	Laser Surg Med	4.72	24.99	50.44	73.64	89.18
1982	1981	1980	1979	1978		
94.81	96.61	100.0	100.0	100.0		

3) Journal half-life listing

The third part of JCR is journal half-life listing. It includes three parts.

The first part lists journals based on the alphabetical sequence of the abbreviation of the source journal title (Table 8.12).

- ① Cumulative distribution percentage data of the citing number of citing journal
- ② Citing half-life (defined by Burton) (Table 8.13).

The second part provides the related data of cited journals.

- ① Cumulative distribution percentage data of the cited number in the current year of papers published over the past 10 years
- ② Journal cited half-life (defined by Burton)

The third part lists the cited journal table based on the descending order of the cited half-life.

4) Citing journal listing

Citing journals are listed by sequence of the abbreviation of the journal title in the cited journal listing. It, in turn, provides the impact factor, journal title (abbreviation), total number of citing other journal papers, number of citing other journal papers of each year from 1978–1987, and the number of citing other journal papers in 1978. In the citing journal, it lists cited journals (in descending order of the number of cited) (shown in Table 8.14).

5) Cited journal listing

Cited journals are listed by the sequence of the abbreviation of the journal title in the cited journal listing. The format of cited journal listing is similar to that of citing journal listing. It, in turn, gives the impact factor, abbreviation of the journal title, total cited, and the distribution of total cited of many years. In the cited journal, it lists citing journals based on the descending order of the number of citations (shown in Table 8.15).

Table 8.14 Citing journal listing fragment (from JCR, Garfield, 1988)

IPF citing journal IPF cited journal	Total of citing, cited	Amount of citing, cited per year									Total of citing, cited before 1978
		1987	1986	1985	1984	1983	1982	1981	1980	1979	
1.15 J Am Stat Assoc	2416	49	145	203	200	161	165	157	118	96	97
1.15 J Am Stat Assoc	391	12	29	42	37	26	22	34	17	17	140
1.19 Ann Stat	170	5	9	18	22	12	17	18	8	10	8
1.10 Biometrika ...	153	1	9	8	8	13	7	4	6	5	8
						...					84

Table 8.15 Cited journal listing fragment (from JCR, Garfield, 1988)

IPF cited journal IPF citing journal	Total of cited, citing	Amount of cited, citing per year									Total of cited, citing before 1978
		1987	1986	1985	1984	1983	1982	1981	1980	1979	
0.53 J Chem Eng Data	1691	27	76	95	103	76	97	84	65	60	77
0.53 J Chem Eng Data	247	12	26	28	23	15	18	16	5	6	5
0.86 Fluid Phase Equilibr	172	0	5	16	7	11	16	12	7	7	80
0.82 J Chem Thermodyn ...	95	0	1	13	4	4	9	4	2	1	3
						54

(3) Introduction of JCR® Web

JCR® Web has two editions: JCR Science Edition and JCR Social Sciences Edition. JCR Science Edition provides the citation analysis assessment information of more than 5,000 scientific and technical journals collected by *Science Citation Index Expanded™*, whereas JCR Social Sciences Edition provides the citation analysis assessment information of more than 1,600 social scientific journals collected by *Sciences Citation Index®*.

JCR Web (shown in Fig. 8.29) is a comprehensive and multidisciplinary journal analysis and evaluation report. It objectively lists the original data collected by *Web of Science*, such as number of articles published, number of references, and paper cited frequency. Then, it calculates the impact factor, immediacy index, and cited half-life of various journals based on the principles of bibliometrics to reflect the quality of journals and the effect of quantitative indicators.

For each collected journal, JCR Web provides journal statistical analysis indicators to users: the total number of each cited journal in the current year (total cites), the impact factors of each journal (the average cited number of papers in the current year of each journal that were published two years ago) (impact factors), the average cited number of papers published in the current year of each journal (immediacy index), the total number of papers published in the current year of each journal (articles), the duration of each paper's research subject (cited half-life), citing journal listing of each journal (citing journal), cited journal listing of each journal (cited journal), the changes in each journal's impact factor in recent years (trends), the situation of source data in each journal (source data), and so on.

The definitions of the specific structure and indicators of each Journal Citation Report in JCR Web are shown below (Fig. 8.30).

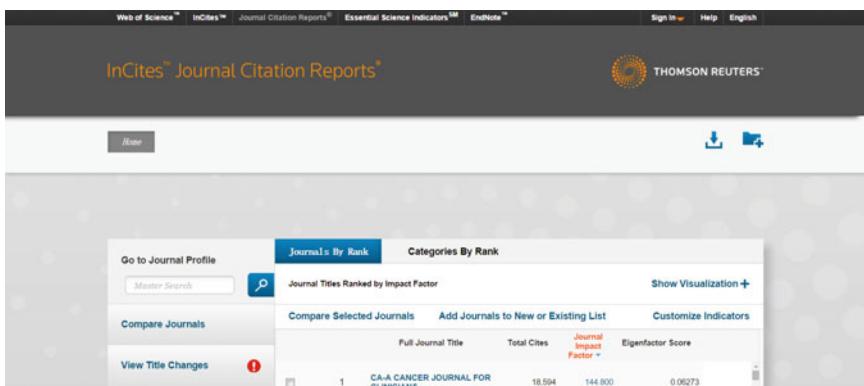


Fig. 8.29 Homepage of JCR Web

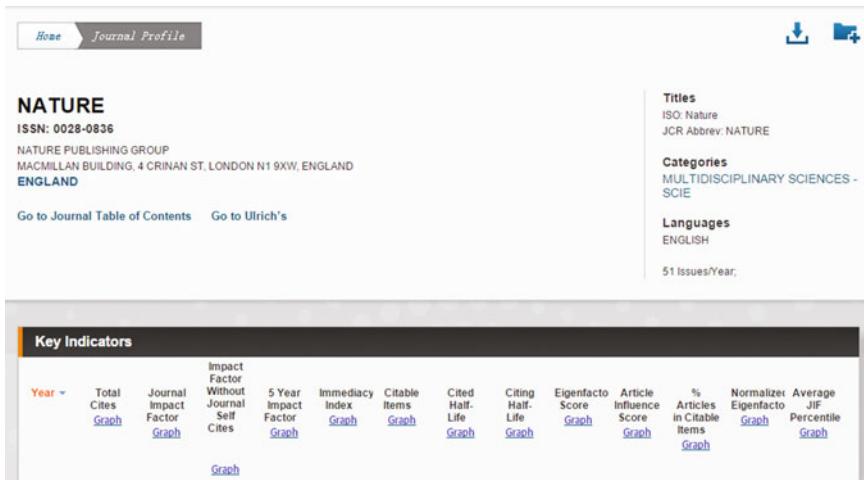


Fig. 8.30 Journal citation index overview interface in JCR Web: journal citation index and structure

By clicking on each index item on the page, one will see the detailed index page.

① Total cites

It refers to the total number of times that this journal is cited by other journals (collected by ISI)

② Impact factor

It is used to evaluate and compare journals in a discipline and finds the most influential and important journal. For example, based on the paper cited frequency, institutions will decide to order necessary journals.

③ Immediacy index

It is used to evaluate paper cited speed of journals and paper cited frequency of journals in the same year. It is useful to compare journals of new disciplines or cutting-edge disciplines.

④ Article counts

It only includes original research papers and review papers. Given that editorials, letters, news, and conference abstracts are not generally cited, the number of papers does not include their number.

⑤ Cited half-life

It is a persistent standard of cited papers. Starting from the current year onward, the number of years citing the half-life of the number of citation accounts for 50% of the total number of citations of cited journals at present.

⑥ Citing half-life

Starting from the current year onward, the number of years citing the half-life of the number of citations accounts for 50% of the total number of citation of citing journals at present. Compared with the cited half-life, it can be evaluated to edit the policy.

⑦ Source data

It provides the basis of reviewing the original research and the number of references in journals.

⑧ Cited journal listing

A user can arrange citing journals according to the frequency of citing specific journals. This citing link can explain the subject direction of journals, point to the closest similar or competitive journals, and form the specific professional journal network.

⑨ Citing journal listing

A user can arrange cited journals according to the frequency of cited specific journals. This citing link can explain the subject direction of journals, point to the closest similar or competitive journals, and form the specific professional journal network.

⑩ Impact factor trend graph

The impact factor trend graph (shown in Fig. 8.31) describes the past five years' impact factor of a specific journal. It measures the cited frequency of the average

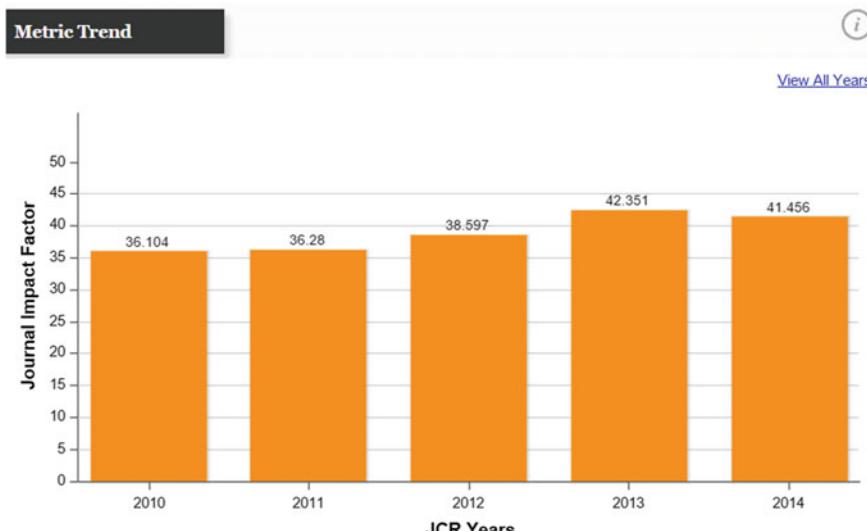


Fig. 8.31 Impact factor trend of JCR Web journal

article of this journal in a particular year and compares the influence of this journal in the current year with the influence in the past 4 years. At the same time, it also links the users of *Web of Science*® to JCR Web.

(1) Journal title

The title of journals in JCR is usually abbreviated, including the subtitle and title. The full name of a journal is in the full record.

(2) Marking a journal

Click on an option box under “Mark” and then click on the Update Marked List button to display the marked journals in the Marked Journal List. A user can mark journals in Summary list or Full Record.

(3) Go to Ulrich™

On the Full Record page, click on the Go to Ulrich button to directly link to Ulrich's Periodicals Directory™. If a JCR user does not subscribe to Ulrichsweb.com, the search interface does not have this link button.

(4) Application of JCR

Through JCR, a large amount of meaningful data can be obtained to comprehensively evaluate journals, such as disaggregated data of the source journal, number of articles published, citing rate of journals, mean citing rate, impact factor of journals, half-life of journals, and immediacy index. By using these indexes and data, one can carry out a wide range of quantitative analysis. These indexes and data provide convenient conditions and new applications for many quantitative analyses in philology, information science, and science of science. Some of the practical applications of JCR are briefly described as follows.

1) Research and selection journals—determination of core journals

The interpretation of mutual relation among journals and the degree and distribution of journals is the main function of JCR. Thus, data indexes (such as cited rate of journal) in JCR can be used to quantitatively and comprehensively evaluate the importance of scientific journals and determine the core journals of a single discipline or multi disciplines. It can be used to determine and select other professional related journals that should be collected by comparing the citing journal listing of a certain discipline with the cited journal listing of this discipline. It can also analyze the contact degree of using literature among different disciplines. Moreover, it judges the professional nature of certain special journals based on the size of the citing rate of a discipline. For example, if the impact factor of a journal is large, then this journal orients basic research and has a theoretical nature.

In discussing the application of Bradford's law, we have posited that using the distribution of the number of related published articles determines core journals. Finally, we will obtain the number of core journals and the specific core journal listing. However, the factors, such as the limitation of relevant papers, are rather

vague. The judgment criteria vary, and no related reference books provide the number of related published articles, thus making the determination of core journals through Bradford's law difficult, especially the determination of core journals in the field of natural science.

However, one can easily determine the discipline core journal listing and the total scientific journal listing based on the data used by users and the citation relation among scientific journals through the application of Garfield's centralization law of periodical literature and citation data provided by JCR.

For example, if a user wants to determine the core journal in a discipline or specialty, he/she can select 1–2 representative journals in this discipline or specialty, search journals that are cited by this journal by citing the journal listing of JCR, and order those journals according to the total cited number or impact factor. Finally, the user can obtain core journals in this discipline or specialty. If both citing journal listing and cited journal listing are used, the relevant professional journals that are required in this discipline can be obtained. This is because the citing journal listing not only cites its professional journals, but also cites many relevant journals; however, the cited journal listing mainly contains its professional journals. By comparing two parts of journals, one can access the relevant professional journals required in this discipline.

For determining the core journals of multi disciplines or even all scientific fields, the index data provided by JCR, such as total citations and impact index, can be used. Math can also be applied to process the index data and sequence them comprehensively to determine the core journals.

2) Research of journal aging law

The aging speed of several journals and the research journal aging law can be judged according to the age distribution and duration of cited literature provided by JCR. The application of this result can directly guide journal service mode and set the standard of journal saving time and journal ticking. Combined with the circulation data of a journal, a user can also carry out theoretical and applied research on literature aging.

3) Comprehensive evaluation of journals

JCR comprehensively provides various types of data, such as production output, feedback, age distribution, and evaluation index data. Thus, in addition to the two applications mentioned above, it also can comprehensively evaluate journals from other aspects.

① Determination of the contact degree among journals or relation among disciplines

By analyzing the overlapping portions of two mutually cited journals, a user can judge the contact degree between two journals and the relation among disciplines. The group and cluster of journals can then be obtained, used, and managed easily.

② Judgment disciplinary nature

Through the cited time provided by the cited journal listing, a user can judge the disciplinary nature and specialty areas of journals and the core journal group composed of highly cited journals.

③ Scientific management journals

4) Research on science of science

For research on science of science, JCR has an important role. By analyzing the interaction among journals, to some extent, it can learn the situation of crossing and penetration among disciplines and helps reveal the general law of the development of science.

JCR data can be used to objectively and accurately evaluate and assess journals from all aspects and provide a basis for the scientific management of journals. A user can correctly understand and judge the academic status, academic nature, and user trust degree of journals as well as the role and influence in scientific activities and literature exchanges. However, notably, owing to language and other aspects, SCI and JCR are mainly based on the literature of the United States, Britain, and other countries. Both of them have less Chinese literature. Therefore, we should vigorously advocate and compile *Chinese Science Citation Index* and *Chinese Journal Citation Reports* as soon as possible. Fortunately, the Institute of Scientific and Technical Information of China, Chinese Academy of Sciences, China Academic Journal Electronic Magazine of Tsinghua University, and other institutions have introduced the corresponding annual report in recent years, such as *Chinese S T Papers Statistics Report*, *CSCD ESI Annual*, and *Chinese Academic Journals Comprehensive Citation Report*.

8.5 Citation Network and Cluster Analysis

The citation and cited scientific literature make a large amount of literature grouped and clustered. Careful observation of the citation network map (shown in Fig. 8.1) introduced in 10.1 reveals that in the citation relation of literature, in addition to single mutual citation relations among literature, there are various complex networks and relations of clusters, such as two or more papers citing the same literature at the same time or two literatures are cited by other papers. This is the research content that this section focuses on, i.e., it regards the connection degree among literature as measurement units of citation analysis, which constitute the theoretical basis of literature cluster and discipline cluster analysis.

8.5.1 *Concept of Bibliographic Coupling and Co-citation*

(1) **Bibliographic coupling**

In the cited literature of scientific literature, people often see authors of different papers spontaneously cite an article or several similar articles. In response to this phenomenon, M. M. Kessler, a professor of Massachusetts Institute of Technology, first proposed the term “bibliographic coupling” in 1963. Kessler used the journal *Physical Review* to investigate citation analysis and found that if the disciplines and professional contents of papers are similar, the amount of the same literature in their references is large. Thus, he posited that two papers (or more papers) that cite the same paper at the same time are called coupling papers, and the relation between them is called bibliographic coupling.

In bibliographic coupling, citations establish a coupling relation through their reference (cited reference). Specifically, if two literature (A and B) have one or more co-cited references or have one or more similar references, paper A and paper B have a coupling relation in citation. Papers that have a coupling relation are considered to have some contact or correlation in the discipline content. The index of coupling strength can be used to measure papers’ coupling degree. The measurement unit of coupling strength is the number of common references between paper A and paper B. If two papers have the same reference, the coupling degree between these two papers is one citation coupling (or coupling unit). Through this analogy, if two papers have n similar references, the coupling degree between the two papers is n coupling units. Obviously, a high coupling degree means that two papers are close with regard to the discipline content and specialized nature. The contact between them is also close.

In general, citation coupling refers to the relation established by two citations, but it is not limited to two articles; it may be n articles, $n \geq 2$. Citation coupling is a relative term. With the different objects of coupling, coupling standards also differ, and a citation coupling group with different features can be formed. The index of coupling width can be used to measure the coupling range. Therefore, the phenomenon of citation coupling makes a large amount of scientific literature clustered. It not only provides the possibility of retrieving literature from the angle of literature utilization to improve the relevance and efficiency of the literature intelligence service, but also opens up a new means to investigate problems, such as citation structure and law of literature, the similarity of topics, and the structure of the discipline.

The starting point of bibliographic coupling theory is that two citations that have one or more common references must have a connected relation. In fact, the concept of coupling is not limited to the relation between two papers that cite the same papers at the same time. It also reveals a kind of mutual relation, i.e., the relation between two (or more) different subjects and the same object. Therefore, it can promote the concept of Kessler’s bibliographic coupling. The feature objects (discipline subject, journal, author, language, country, published time, and so on) of

literature can also present a coupling relation. In other words, the concept of coupling also reflects the coupling relation between two (or more) authors that have a common reference at the same time. For example, if we take the journal as the subject instead of the literature unit and if two journals cite other journal's paper at the same time, we can say that the two journals have a coupling relation. At the same time, if two journals cite a journal one time, it can be said that one citation coupling (or cited journal coupling) exists. A high citation coupling means that the relation between two journals is close.

In a broad sense, the coupling phenomena of scientific papers and their related media objectively combine subject objects together, which seem to have no relation on the surface, to reveal the internal and structural relation of the scientific literature system. Hence, in a broad sense, understanding and analyzing the concept of coupling has some theoretical and practical significance: it can comprehensively understand the research object and research range of informetrics and promote the application of informetrics in other fields, such as science of science.

In the beginning, Kessler regarded bibliographic coupling as a new type of search tool. If a user has a related paper P0, he can retrieve all paper groups GA(P0) that have a coupling relation with paper P0 through the retrieval system. Kessler called GA(P0) the logical reference of P0. As a search tool, bibliographic coupling has the following unique advantages.

- 1) Bibliographic coupling is not dependent on any artificial retrieval language and vocabulary. The entire process is completed by a computer with automatic match calculation. Thus, it avoids the difficulties caused by inconsistencies in language, grammar, and vocabulary habit and improves research efficiency and quality.
- 2) Like other types of citation index retrieval, bibliographic coupling does not need expert reading or judgment, which brings great convenience for the Department of Library and Information Science.
- 3) As a search tool, bibliographic coupling can break through the limitation of traditional static classification. When basic paper P0 is continually cited by others, logical reference group GA(P0) will continue to expand and the number of papers will continue to increase. This reflects the new change and direction of scientific research.

Although several scholars have theoretically disputed the concept of bibliographic coupling, the new CD-ROM version of SCI still provides papers that are close to their index papers to users.

(2) Co-citation

When analyzing the citation relation among literature, it can research the dynamic law of literature structure from two angles; the first angle is that papers have similar references, the second angle is that a paper is cited by later literature at the same time. In 1973, Henry Small, an American intelligence scientist, and Irina

Marshakova, a former Soviet intelligence scientist, proposed the concept of literature co-citation when they respectively researched the citation structure of literature and literature classification. They regarded it as another method to measure the contact degree among literature.

The so-called literature co-citation means that two papers (or more papers) are cited by one or more later papers at the same time. It claims that these two papers (cited papers) have a co-citation relation. In other words, if literature A and B (or more literature), regardless of their published time, are cited by one or more later papers at the same time, then literature A and B have a co-citation relation. The number of papers that cites them (citation amount) can be used to measure the degree of co-citation, i.e., the number of papers that cite these two papers is defined as co-citation strength or co-citation frequency. Description with the language of set theory may make co-citation strength easily understood. If set A is made up of papers that cite literature X and set B is made up of papers that cite literature Y, $A \cap B$ is a set that cites literature X and literature Y at the same time. The number of elements in $A \cap B$ is the co-citation strength of literature X and literature Y. If there are more papers that cite these two papers, their co-citation strength is much higher. It indicates that the relation among them is much closer. The index of co-citation amplitude can be used to measure the span of co-citation related group of literature. If there are many co-citation literature in the group, the co-citation amplitude is high.

In 1973, Henry Small proposed the concept of literature co-citation, regarded it as a method to describe the relation among the important concepts in the field of science, and simulated the real structure of scientific knowledge. He regarded the specialty of particle physics as a case and conducted research on co-citation analysis of journal papers. The result is shown in Fig. 8.32.

The citation network map shows that two pairs of important literature from 1968 have a close relation: Lovelace–Veneziano and Geu–Mann–Glashow. Their co-citation strength exceeds 49. They are linked by the earlier papers of Geu–Mann and Weinberg.

The co-citation relation can also be used for information retrieval. For example, it can take papers cited many times as the basis to establish a secondary index. Hence, people can use the co-citation retrieval point to retrieve new related literature.

In addition to existing between two literature, the co-citation relation still exists in the co-cited relation of tri-citation or multi-citation. In 1974, Henry Small proposed a geometrical model (the circle model) that vividly demonstrates the co-citation relations of double-citation, tri-citation, and multi-citation. Henry Small regarded six documents in particle physics (data in 1972) as the cited data object, determined the distance among literature, and produced an appropriate circular map (shown in Fig. 8.33).

This figure is an imagery of the co-citation relation among literature, and it provides a sense of the subject. Henry Small used this model to analyze the quantity of co-citation relation of tri-citation and found that it was consistent with the actual results.

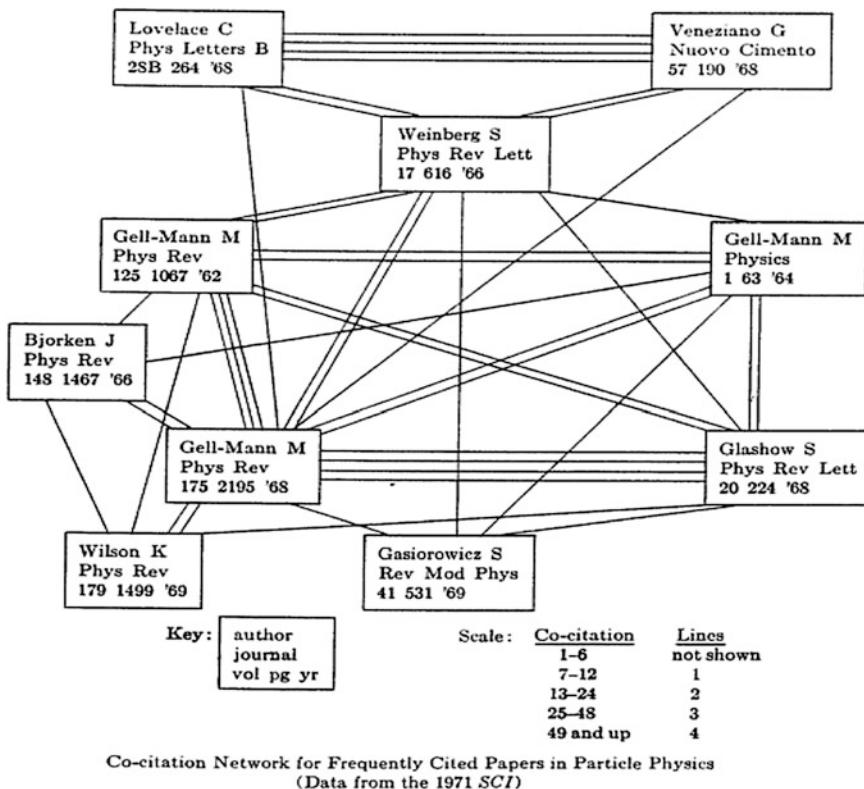


Fig. 8.32 Co-citation network for frequently cited papers in particle physics

Similar to the concept of bibliographic coupling, the concept of co-citation is also extended to various feature objects that are related to literature and forms various types of co-citation concepts, such as journal co-citation, author co-citation, and subject co-citation. In a broad sense, the phenomenon of co-citation is that the feature objects without external contact are objectively linked together through the authors who cite them at the same time. Therefore, from different angles, it reveals the complex structural relation among citations. It provides a new approach for comprehensive analysis of references and research on the citation structure.

(3) Similarity and difference between bibliographic coupling and co-citation

The concepts of bibliographic coupling and co-citation relation both have a similarity and a difference. They both mean that two papers establish the relation through another paper (or papers) to reflect the contact degree and structural relation among literature. In citation analysis, they belong to the same type: the analysis of network structure regards the contact degree among literature as the measurement unit. The angle of citation can reflect the similarity of subjects in papers and the role and contact among one another. Both methods can be used to study the relation

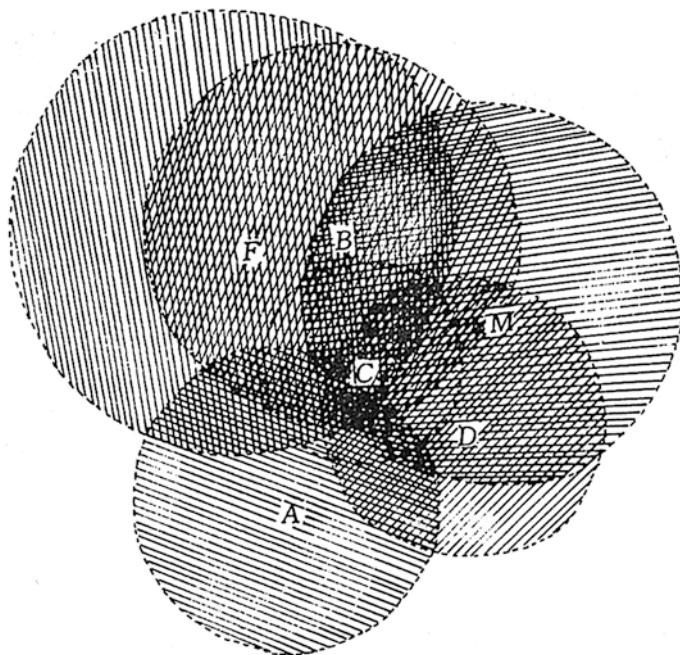


Fig. 8.33 Co-citation relation of six documents

among literature, retrieve literature, and reveal the discipline structure. However, the direction of dealing with problems and the subject/object of these two methods are different. Thus, there is a clear distinction between them, summed up as the following points.

- 1) Citation coupling reflects the relation between two citations, whereas co-citation reflects the relation among cited references. The former is established by the authors of two literature, and the latter is established by the authors that cite them.
- 2) Citation coupling strength is fixed, but co-citation strength may change at any time. For any published papers, the subsequent references are fixed. Therefore, the citation coupling relation does not change and becomes fixed in the long term. For the co-citation relation between two papers, the features of co-citation determine their position, which is always passive. Their relation depends on other literature to be established. Their strength is determined by the demand of other literature. Thus, the co-citation relation is still in a state of change.
- 3) The relation among literature reflected by citation coupling is a long-term fixed relation, but co-citation reflects a change or temporary relation. Therefore, the model formed by citation coupling is a static structural model, whereas the model formed by co-citation is a dynamic structural model.
- 4) Citation coupling is backtracking and belongs to back citation, whereas co-citation is forward-looking and belongs to forward citation.

Table 8.16 Comparison of bibliographic coupling and co-citation

Content item no.	Citation coupling	Co-citation
1	Reflects the relation between two citations	Reflects the relation among cited references
2	Must be established by authors of two or more citations	May be established by a single author of a citation
3	Relation medium is cited literature or references	Relation medium is citation
4	Its degree is measured by coupling strength index (common reference number)	Its degree is measured by co-citation strength index (common citation number)
5	Coupling strength is fixed	Co-citation strength changes at any time
6	Long-term fixed relation among citations and a static structural model	A change or temporary relation among cited literatures and a dynamic structure
7	In active citation status	In passive cited status
8	Backtracking, belongs to back citation	Forward-looking, belongs to forward citation
9	Cited literature Citation C ← A ↑ B First published Later published Bibliographic coupling relation between A and B	Citation Cited literature C → A ↓ B Later published First published Co-citation relation between A and B

- 5) For the research and interpretation of the internal relation and law among scientific literature and for depicting the dynamic structure of scientific development, co-citation has more advantages than citation coupling. Co-citation is more suitable for the continuous change and development of research objects in contemporary information science. For comparison purposes, the differences between citation coupling and co-citation are summarized in the following table (Table 8.16).

8.5.2 Coupling Analysis

The analysis and research on various types of coupling relation, which center on citation, not only enrich the research and method of informatics and open up new application areas, but also provide a new and effective means for research on science of science, prediction science, and technology management.

The coupling relation can serve as a new type of retrieval tool that can complete a special type of topic group retrieval. The other important application of the coupling relation is that it can examine the internal structure of disciplines, divide paper groups that have high coupling strength among similar disciplines and specialties, provide the different closeness among groups, and form a citation network that has an interaction effect among papers.

At the same time, bibliographic coupling reveals a common relation between citation and cited literature that exists between the subject and object. Thus, the concept of bibliographic coupling can be promoted. The concept of coupling can be used to reflect the similar coupling relation of various features of an object, such as discipline, journal, author, language, country, institution, and time.

Coupling analysis method is a type of metering analysis method. It is a very important and special method of informetrics and scientometrics. It is widely applied in philology, science of science, prediction science, and technology management. Besides, it continues to develop its own new research field. Below are some types of important coupling analyses.

(1) Bibliographic coupling analysis

Bibliographic coupling analysis is also called citation coupling analysis; it is the basic coupling relation analysis. It regards cited papers as the contact medium and links papers that seem unrelated to reflect the relation among citations and reveal the internal development law and organizational structure of scientific disciplines. It has some significance in research on philology, science of science, prediction science, and technology management.

Literature coupling can be analyzed from the following aspects.

- 1) It analyzes the relation among disciplines through the citation related paper group. In bibliographic coupling analysis, many so-called logical reference clusters $G_A(P_0)$ are formed, i.e., several papers that have a coupling relation form a paper group. Two types of citation-related paper group may be formed. The first one is the closed structure. In the citation group, if it has at least one coupling unit between each paper and any paper in the group, this paper group will form a closed structure in which papers in this group are interrelated. The second one is the open structure. In many papers, if it has at least one coupling unit between each paper and another benchmark paper (belongs to discipline A), those citations will form an open structure in which a paper is interrelated to any paper outside the group. Obviously, to some extent, these citation-related groups reveal the relationship among discipline literature and reflect some contact among discipline contents, i.e., they belong to the same discipline, branch of a discipline, or interdisciplinary and edge-discipline.
- 2) It can analyze the structure of the literate system and discipline through the citation network. The citation coupling phenomenon objectively links many papers that seem to be unrelated on the surface and forms a related paper network. In fact, a paper network that has a coupling relation must have several common attributes. The network attribute among papers decided by

citation relation causes many papers that do not have external contact to have an ordered structure. It can examine the structures and regularities of literature and information flow by analyzing these structures and then study the structure and development law of the entire discipline.

- 3) It provides a new approach for literature retrieval. The citation coupling phenomenon can put scientific papers into a related group with various attributes according to their citation relation. Thus, it provides the possibility to retrieve papers from the aspect of literature utilization. To some extent, this method of retrieving related papers from the coupling citation makes up for the lack of traditional literature retrieval and expands the retrieval range. Hence, it greatly improves the inspection precision and inspection rate of literature.
- 4) It helps to evaluate classic literature, track scientific development paths, determine the current hotspots of a discipline, and learn the exchange law among sciences.

(2) Journal coupling analysis

Journal coupling analysis is the promotion of bibliographic coupling analysis. The so-called journal coupling means that the journal is regarded as the basic unit to establish the coupling relation. Specifically, if 2 ($n \geq 2$) journals cite the literature of other journals at the same time, then these n journals have a coupling relation called journal coupling. The species number (or frequency) of cited journals can be used to measure the coupling strength. This measurement index is called journal coupling strength or journal coupling frequency.

The journal coupling phenomenon objectively combines many journals into ordered and related groups according to the citation relation. To some extent, it can reveal the interrelation among journals and provides an objective basis and condition for research on the structure and law of literature and information flow and the relation among disciplines. Journal coupling analysis is mainly conducted from the following three aspects.

- 1) Judgment of the relation and contact degree among journals. If two journals have a coupling relation, then some contact exists in several attributes between the two journals. For example, they belong to a professional field or they belong to journals of different disciplines, but their papers have some relation. It can reveal the intersect relation on the discipline content. Moreover, the increase or decrease in coupling strength can reflect the strength and changes in their relation.
- 2) Judgment of the professional nature of a journal. It can determine the professional nature of a journal through the coupling strength and coupling proportion of journal citations.
- 3) It can judge the interrelation and content degree among journals through content analysis of journal papers from the aspect of journal coupling relation.

(3) Author coupling analysis

The coupling relation among authors is the natural extension of the coupling relation among papers. The so-called author coupling means that it takes authors (contains author group) as the basic unit to establish the coupling relation. Specifically, if 2 ($n \geq 2$) authors cite the literature of other authors at the same time, then these n authors have a coupling relation. The coupling medium is the author whose paper is cited by other authors. The number of authors whose paper is cited at the same time can be used to measure coupling strength. This measurement index is called author coupling strength or author coupling frequency.

Author coupling analysis can reflect the objective relation among authors to some extent, and it reveals the organizational structure of a discipline's professional staff. This analysis method has a wide range of application, such as in the fields of library and information science, science of science, and talent science.

- 1) It can analyze the situation and trend of discipline research through the author group network. The author coupling phenomenon objectively combines many authors into ordered and related groups according to the citation relation. The ordered author groups are categorized according to the attributes of papers written by authors. The situation and trends of the author group (belonging to one discipline) in terms of quantity, quality, structure, and other aspects can reflect the team situation of this discipline research, the development process of the discipline, and its trends. Thus, it provides a new means to examine informetrics and the science of science.
- 2) According to the composition of the related author group, a necessary communication network can be established for scientific peers to promote information transmission and academic exchange.
- 3) In information retrieval, from the name of the author in the coupling author group, the author catalog and author retrieval can be used to search all related literature that were published by peer authors in a professional subject. Thus, it provides a highly targeted service of customizable subject retrieval for the research on this discipline subject.

This shows that through author coupling analysis, the network structure form of the paper author group in the field of a particular discipline can be determined together with the situation and development of the author team and the development process and trends of the discipline. Using the author network may help expand information transmission and academic exchange and promote the development of scientific research.

(4) Discipline coupling analysis

The so-called discipline (or specialty) coupling means that discipline is regarded as the basic unit to establish the coupling relation. Specifically, if literature of 2 ($n \geq 2$) disciplines (or specialties) cite the literatures of other disciplines (or specialty) at the same time, then these n disciplines have a coupling relation. The number of cited disciplines can be used to measure the coupling degree. This

measurement index is called discipline coupling strength or discipline coupling frequency.

Obviously, discipline coupling is a type of backtracking coupling. It forms a related group of relevant disciplines. With this coupling relation, the following analysis can be made.

- 1) The discipline coupling structure can be used to judge the relation among disciplines. Citation statistics can determine the coupling relation structure and coupling strength among disciplines. Obviously, some relation exists among disciplines in the coupling related group and cited disciplines. If the coupling strength of two disciplines is high, then the relation between two disciplines is close, and these two disciplines have a cross-penetration trend. If several disciplines are not coupled or the coupling strength is low, the relation among them is not close.
- 2) The change in the discipline coupling relation can reveal the situation and change law of the development of the discipline. If it accounts for and determines the discipline coupling relation and its coupling strength in different periods and continuously tracks and analyzes the changes in the structure of discipline coupling, then the development situation and the complex changing relation of a discipline can be understood. If coupling strength increases, the relation of relevant disciplines strengthens. On the contrary, it means that they are becoming independent and different.
- 3) It provides a quantitative basis for selecting and collecting special literature. Discipline (or specialty) coupling strength can be used to determine the correlation degree among disciplines. Thus, it provides a quantitative basis for the selection and collection of special literature.

The relation, contact degree, the relation of the branch hierarchy among disciplines, and their cross-penetration trends can be determined based on discipline coupling analysis. The situation and change in the development of a discipline can be learned based on the relation of discipline coupling to predict the trend of differentiation and combination of disciplines.

In addition, there are many couplings, such as country, regional, institutional, and language coupling. These coupling relations can be used to make an appropriate analysis to obtain many useful conclusions.

8.5.3 Co-citation Analysis

Similar to coupling analysis, various types of co-citation analysis can be examined and analyzed. Researching and analyzing the relation of various types of co-citations are important parts of citation analysis. They not only enrich the research content and method of informetrics, but also present a certain significance for research on science of science, talent science, prediction science, and so on.

(1) Literature co-citation analysis

Co-citation of literature is a basic co-citation relation. It mainly reflects the structure relations among co-citation references and further reveals some connection among disciplines. The analysis of co-citation relation among literature is significant. The main valuable points are as follows:

- 1) Theoretical research on philology can be performed through the analysis of co-citation related group of literature. For example, from the aspects of co-citation type, language, and other elements, the laws of the structural feature, distribution, and utilization of a scientific literature system can be studied.
- 2) Research on science of science can be performed through the analysis of the literature co-citation group network and its change. For example, it can be used to examine the interrelation, contact feature, development situation, and trends among disciplines.
- 3) It provides a basis for the establishment of a literature procurement strategy and model to open up a new approach of literature retrieval.

This shows that literature co-citation analysis can be used to learn the structural feature of the literature co-citation group, the distributed forms of disciplines, the types of literature and language, and the regularity of mutual citation in a scientific literature system. By analyzing the co-citation literature group network structure and trends, science of science and technology management can be studied, and the development situation and trends of interrelations among disciplines or the entire scientific system can be examined.

(2) Journal co-citation analysis

This is an extension of literature co-citation analysis. The so-called journal co-citation means that journals are regarded as the basic unit to establish the relation of co-citation. Specifically, the papers of n species ($n \geq 2$) journals are cited by other journals at the same time. It can be said that these n journals have a co-citation relation. The species (or frequency) of journals (citation journal) that cited these papers can be used to measure the co-citation degree. This measurement index is called journal co-citation strength or journal co-citation frequency.

Obviously, according to the cited relation journal, the co-citation relation combines a large number of journals. Then, from the perspective of utilization, it reveals the interrelation and structural features among journals of various disciplines. The analysis of journal co-citation relation is as follows:

- 1) According to the relation and strength of journal co-citation, the discipline (or specialty) nature of several journals can be determined.
- 2) It provides a basis and a new means to determine core journals. Journal co-citation reflects the connection of several disciplines or specialties. If the co-citation frequency is high, then such a professional relation is close.

(3) Author co-citation analysis

Author co-citation is an extension and development of literature co-citation. It regards authors as the co-citation measurement unit and examines the situation of literature published by n ($n \geq 2$) authors that are cited by other authors at the same time. It uses the number of citation authors to measure the co-citation strength.

According to the authors of co-citation literature, author co-citation establishes the co-citation relation. Many authors thus form a related author group based on the co-citation relation to reveal the organizational structure and contact degree of discipline professionals and reflect the contact and development situation of disciplines.

- 1) The situation of peer author can be learned through the constitution of the co-citation author group. If n authors are cited by the author of a special subject literature, then these n authors and citation authors are peers in this special subject research. If the co-citation frequency is high, the relation of disciplines is close. From the author co-citation group, one can learn the situation of total peer author, constitution, activity law, and other aspects. If the peer authors (researchers) of a special subject research combine together and form a collaborative network to strengthen academic exchanges and conduct collaborative research, the development of disciplinary research would improve.
- 2) The trends of disciplines can be speculated through the changes in the number of co-citation author groups and core author groups. To a certain extent, the changes in the number and structure of authors can reflect the trends of development, rise and fall, and differentiation and infiltration of disciplines and science systems. In a co-citation network, the changes in the number and structure of authors can be used as a basis for judging the discipline dynamics. The development direction and trends of a discipline or specialty can be tracked and speculated by regularly inspecting and analyzing the changes in these aspects. Therefore, author co-citation analysis has become one of the common methods in science of science, prediction science, and talent science.
- 3) It provides a new means of literature search. From the authors who have a co-citation relation, a given topic title search service can be carried out to improve the relevance and efficiency of the literature search.

(4) Discipline co-citation analysis

Discipline co-citation regards all disciplines as the research object. The so-called discipline co-citation means that disciplines are regarded as the basic unit to establish the relation of co-citation. Discipline co-citation analysis regards disciplines as a statistical unit of co-citation analysis and examines the situation of literature in n ($n \geq 2$) disciplines that are cited by literature in other disciplines at the same time. The number of citation disciplines is used to measure the co-citation strength. If the literature in n ($n \geq 2$) disciplines are cited by literature in other disciplines at the same time, then these n disciplines have a co-citation relation. The number of disciplines (citation discipline) that cited these disciplines can be used to

measure the co-citation degree. This measurement index is called discipline co-citation strength or discipline co-citation frequency. If the co-citation frequency is high, then the discipline relation is close, the nature is highly similar, and the degree of cross-penetration is high. The structural relationship and contact degree among various disciplines or various branches of a discipline are revealed. In this regard, discipline co-citation analysis proceeds as follows:

- 1) The relation among disciplines and the structure of science system can be analyzed based on the constitution of the co-citation discipline group. Many disciplines (or specialties) are gathered into a discipline group and form a network based on the co-citation relation. This relation and structure of the discipline group network not only reveal the relation of intersection and dependency among several disciplines at the microscopic level, but to some extent, they also reflect the discipline composition and structural features of a science system at the macro level. Thus, a new method for research on science of science is provided.
- 2) The trends of disciplines can be predicted based on the change in the amount of the co-citation discipline group. The changes in the amount, structure, and other aspects of discipline groups that have a co-citation relation reflect the differentiation, infiltration, and comprehensive trend of the discipline (or specialty) to some extent. The development process and trend of a discipline can be traced through a statistical analysis of the co-citation relation of the discipline (or specialty) at different times, i.e., co-citation analysis. This has some significance for research on science of science and prediction science.
- 3) The resource, composition, and exchange law of a discipline and information can be learned and examined based on the analysis of the discipline co-citation relation. This will provide a basis and material for research on information and science.

Therefore, through discipline co-citation analysis, the composition and structural features of disciplines can be learned at the macro level, the development trends of disciplines can be predicted, and the exchange laws among scientific knowledge and information can be learned.

8.5.4 Citation Cluster Analysis

Cluster analysis is one of the most commonly used multivariate statistical methods to reduce the dimension. It belongs to the category of reducing the dimension technology. The result of cluster analysis is usually a network diagram or dendrogram. From this figure, one can analyze and obtain the goal that needs to be predicted and judged.

Cluster analysis of literature is a specific application of cluster analysis in citation analysis. According to the different features of a citation, cluster analysis of

literature is used to research the group and cluster and analyze citations. Generally, citation cluster analysis mainly refers to professional cluster analysis. Given that a certain degree relation of disciplines and specialties exists among citations, according to specialized attributes, citations can be gathered into group clusters. Professional cluster analysis of citation is one of the important contents of citation analysis. Regardless of the research, both are significant.

Double citation cluster analysis, which was developed in recent years, is a type of intelligence analysis technology. It is also an important part of the citation analysis method. This analysis method is mainly from the perspective of bibliometrics and professional clustering to analyze the dynamic structure of science research and measure the achievements of scientists. Generally, cluster analysis of literature regards the *Science Citation Index* (SCI) and the *Social Science Citation Index* (SSCI) as the starting point of research involving a series of survey steps, statistics, classification, collation, analysis, and so on. Finally, a certain research purpose is achieved.

(1) Main principle

According to the concept of double citation, the principles of double citation cluster analysis can be understood. If an author in his paper cites at least two literature simultaneously (also known as literature pair), then these two literature exert a simultaneous influence on the author. If many articles refer to the same literature pair, they will be gathered into a citation cluster. In other words, a co-citation literature pair not only has a common professional literature content, but also acts as an intermediary among citation authors. It establishes a basis for contacting and knowing one another. From another perspective, when an author writes a paper, he needs to refer to related published papers. These references are used to identify the concept, method, and devices of early authors that were absorbed and cited in creating his own paper. Therefore, to some extent, a reference that an author provides formally expresses and explains the thinking method and theoretical substance during his scientific research. Thus, in the research of scientific literature citation, the research areas and directions of citation authors can be known by simply browsing cited related entries.

In general, each scientific paper has more than two references. According to the survey (in 1970) of D. Price, a professor in Yale University, in 1970, each scientific paper usually has 10–22 references. In 1980, through a survey on a large number of formal papers, *SCI* pointed out that each scientific paper has an average of 15.9 references. In addition, in an article, it is possible that author cites different professional cluster literature at the same time. In general, a double citation article is often from two clusters. The features of modern scientific papers not only provide necessary and sufficient conditions to research and interpret the internal structure and law of research activities, but also fundamentally ensure the source of the citation cluster.

(2) Method and step

The basic steps of double citation cluster analysis are as follows:

1) Examining the double citation paper pair and compiling a double citation list

A double citation cluster refers to a technical process in which papers are classified based on double citation intensity. To be specific, the so-called cluster uses the relation of co-citation among papers that have no external relation to gather and form a class. Co-citation strength is a judgment symbol that indicates whether papers have a co-citation relation and are eligible for a class. Therefore, the first step in the double citation cluster is to identify the paper pairs that have a double citation relation. First, a comprehensive investigation and the statistics of published literature in the discipline are required. Second, the related items in the literature and references whose cited frequency is high are selected. Finally, papers are classified and form an index content to identify the double citation relation among them. For any two papers, if they have a double citation relation, they can be selected and used as a literature pair. In this process, many literature pairs can be identified. Thus, cataloging work should be carried out step by step. Cataloging generally uses a two-way method. That is, each literature pair (two papers) is catalogued based on AB or BA sequential order to form two catalog tables. Then, they should be combined into a total double citation catalog table.

2) Professional cluster

The total double citation catalog can be used for professional clustering. It produces a special literature cluster in various disciplines to complete the process of clustering. During the process of double clustering, how to select a moderate and reasonable threshold is a bottleneck that restrains the smooth progress of the clustering. At the same time, it also decides the size of the cluster and its success. In many paper pairs, a standard is needed to determine the people who are eligible to participate in clustering. This is a threshold issue of selecting double citation strength. In general, if the co-citation strength is greater than the selected threshold value, a paper will be gathered a cluster and form a special subject. The papers in this cluster is the core literature of this special subject. Paper pairs whose co-citation strength is smaller than the selected threshold value cannot enter a class; they are used to present the relation among clusters. A selected threshold should be not too high nor too low. If it is too low, it would affect the cluster level and analysis accuracy. If it is too high, it would make the entire cluster system unbalanced. In general, a selected threshold should be based on the specific situation and should regard comparative simple literature in each special subject cluster and the clear mutual relation as the principle. In fact, the more cluster projects there are, the greater the size is and the more complex the threshold selection is. Currently, the amount of data stored in SCI has exceeds 900 million each year. When SCI is used for double citation clustering of various special subjects, the selected threshold often reaches 15–17.

3) Various special research and analysis

Double citation cluster analysis is used to analyze and study the relations and dynamic structure among disciplines, to investigate the laws and structure of scientific activity, to evaluate the capacity and performance of scientists and measure their impact on actual economic results to provide an important reference for awarding and selection, and to establish an important tool to examine science and science history.

Given that the scientific citation source is wide, the amount of data is vast, and the relation of double citation of papers is complex, the process of citation clustering usually uses computer systems to implement. First, a computer program is used for processing, arrangement, classification, and cataloging. Second, according to their respective disciplines, a large number of citations are classified and form the professional subject of literature clusters. Finally, the citations of the individual cluster of papers with high citation rates generated by a survey are organically combined and come in contact together. A computer is used to display the cluster image on the screen. The distribution patterns and contact of various professional disciplines are identified to achieve the study objectives of investigating science and technology activities and structure.

Thus, cluster analysis of literature mainly refers to the coupling strength or co-citation strength as the basic unit of measure and is a quantitative processing technology to classify and cluster the given set of citations and cited literature and closely linked literature of a discipline or professional content. This technology can aggregate papers whose content is closely linked and whose discipline nature approaches literature clusters. It determines the quantitative extent of contact among clusters. Based on these quantitative data indicators, a clustering network diagram or dendrogram of papers in a certain discipline can be created.

Exporting this objective existence of a potential network diagram exerts a huge effect: analysis of contact among literature, analysis of the organizational structure of science, analysis of discipline or professional trends, and analysis of the size and condition of national, regional, or institutional research. It provides reference data and decision-making basis to carry out scientific metrological study, literature service intelligence analysis, and technology management. With the development of modern information technology and the emergence of large-scale citation databases, cluster analysis of literature has become one of the commonly used methods to manage applied research on science of science, information science, and modern technology.

Clustering analysis involves much mathematical knowledge, such as matrix, multivariate statistics, and vector analysis. Thus, it does not do too much to explore theories and regards just one practical example for readers to have a basic understanding in this aspect and generally grasp the method of data processing.

(3) Garfield's cluster analysis of natural sciences

In 1972, Garfield used the citation database of SCI to research all natural sciences by discipline cluster analysis. His raw processing data source includes 867,600

citations, 93,800 literature, and 2,400 journals. It covered all areas of natural science at that time. Cluster processing can be divided into the following steps.

- 1) The threshold is determined and original literature are selected

The threshold of time cited must be determined, i.e., selecting paper quality standards, to ensure that a high time cited literature is selected for cluster analysis. Then, the work of the cluster will have a practical meaning. The size of the selected threshold is related to the size of data processing. According to the size of data processing, Garfield selected threshold = 10. Finally, there were 1,832 cited references from 16,927 journals.

- 2) The work of cited literature are paired and matched

Garfield performed co-citation cluster analysis. First, he paired the 1,832 selected cited references. According to the formula of permutations and combinations, he obtained about 1.7 million paper pairs. Then, each was matched with the source literature and screened (computer processing). Finally, 20,414 paper pairs were obtained. L • Ai Gexi and R • Russo, the authors of *Informetrics Introduction*, reminded readers that most of the papers will not be the same citation. This was verified. Meanwhile, they explained that the entire natural science discipline system is quite open, and the knowledge structure is relatively loose.

- 3) Cluster analysis

Garfield used the co-citation strength of 3, 6, and 10 as three types of threshold and clustered the above 20,414 pairs of literature. The results are shown in Table 8.17.

The table shows that the cluster result threshold value is 3, and 44 literature clusters are formed. The detailed classification table of these 44 literature clusters is shown in Table 8.18. These literature clusters represent the basic structural units and research fronts of science at that time.

- (4) Drawing the cluster map of a discipline

According to the scientific basic unit of cluster analysis, a cluster map of a discipline can be drawn. Dinsmore, a scientist, adopted the data of Garfield and drew a

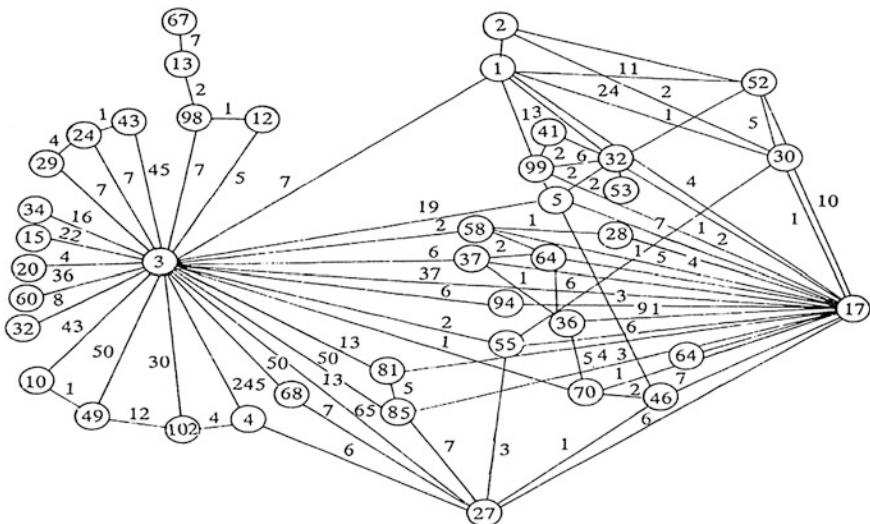
Table 8.17 Number of clustered literature of different thresholds

Threshold	Number of literature pairs	Number of matched literature	Number of unmatched literature	Number of literature cluster
3	3067	1310	522	44
6	791	594	1238	47
10	213	193	1639	18

Note The number of literature clusters in the table does not include literature clusters that consist of two literature

Table 8.18 Cluster group classification with a threshold value of 3

Discipline/specialty	Number of participated cluster literature	Number of matched pairs	No. of groups
Biomedicine	801	2205	3
Chemistry	92	291	17
Nuclear physics	41	59	1
Particle physics	32	99	2
Australia antigen	15	70	10
Crystal structure of enzyme	12	30	27
Plate tectonics	10	35	13
Virus transfected cells	9	25	4
Nuclear magnetic resonance	9	8	5
Visual neurophysiology	7	14	29

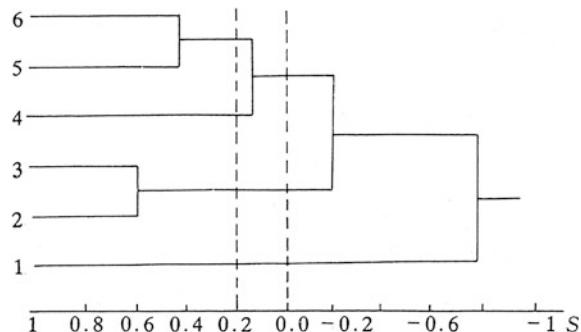
**Fig. 8.34** Cluster map of the same citation

cluster map of the same citation (shown in Fig. 8.34). It clearly shows the complex relation among disciplines and specialties.

The figure shows the following:

- 1) The relations among physics (clusters 1 and 2 with 73 literature), chemistry (cluster 17 with 92 literature), and medicine (cluster 3 with 801 literatures) are close. They constitute the main part of the discipline dendrogram and were the emphasis of natural science research work at that time.

Fig. 8.35 Dendrogram of a cluster



- 2) Among these disciplines, chemistry has the most prominent position. It has a close relation with almost all disciplines. It can be said that chemistry is a binding site of each discipline in natural science.
- 3) Biomedicine is a relatively closed discipline system and has less contact with the other disciplines and specialties in natural science.

In addition to the network graphics, after threshold determination, a dendrogram can be obtained (shown in Fig. 8.35). When drawing a dendrogram, the intensity of the same citation must be converted into a similarity coefficient (various conversion options can be adopted by referring to the related monograph). The horizontal axis represents the similarity coefficient that regards a literature similarity scale. The ordinate axis represents the number of clustered literature. The right of figure is the root end of the dendrogram that represents a set of all the above literature. It expands gradually from right to left. The left of the figure is the part of a treetop that represents every clustered literature. Each node in the dendrogram represents a combination of two low-level literature with a higher level group. The horizontal axis of the node represents the related similarity coefficient and the class level (-1 to 1).

Figure 8.35 shows the different grade standards, S , which may obtain different cluster results. Assuming that $S = 0$ will result in three literature clusters: A (1), B (2, 3), and C (4, 5, 6). Assuming that $S = 0.2$ will result in four literature clusters: A (1), B (2, 3), D (4), and E (5, 6).

8.6 Application of the Citation Analysis Method

Since the citation analysis method was introduced, it has gained widespread attention and application. When SCI and JCR came out, they provided a very useful condition and tools for the application of the citation analysis method. Citation analysis technologies are maturing now, and their application continues to expand.

They have developed into important methods of bibliometrics. The citation analysis method, a widely applied practical technology, has important applications in both library and information science and the bibliographic field. For example, it is used to determine core journals and establish a new retrieval system. It is also widely applied in technology management. For example, it is used to investigate the scientific level, select scientific and technical talents, evaluate scientific and technical journals, and predict the development of technology. Citation analysis is the most outstanding contemporary contribution of scientometrics. This section focuses on the application and evaluation of the citation analysis method and illustrates the method and steps of citation analysis with examples.

8.6.1 Application of the Citation Analysis Method

(1) Determination of the effect and importance of a discipline

The influence of a discipline and the importance of a certain discipline in a country can be measured by analyzing the frequency of literature citations. Figure 8.36 shows the situation of the cited scientific and technical literature of the Federal Republic of Germany. The figure shows if a discipline is below one, its literature citation frequency is lower than that of other countries' average citation frequency. If it is higher than one, the literature citation frequency is higher than that of average citation frequency. Thus, it can be concluded that chemical research in the Federal Republic of Germany has a large effect on other countries, whereas medical research has minimal effect on other countries.

(2) Research on the discipline structure

Relations in discipline content have always existed between science citation and cited literature. Through citation cluster analysis, particularly research on network relations among citations, one can explore the genetic relation and structure among

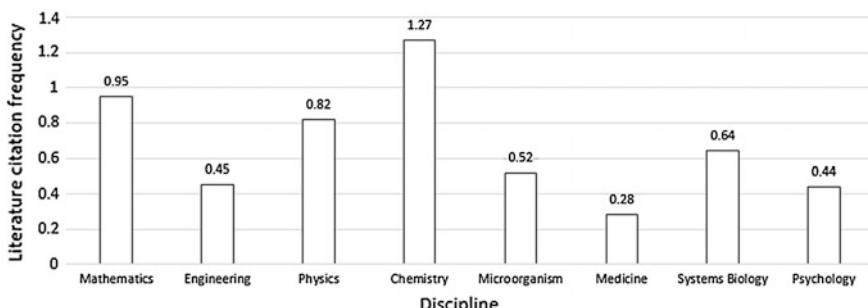


Fig. 8.36 Frequencies of citations in the Federal Republic of Germany

related disciplines and delineate author collective in the discipline. Furthermore, one can analyze and speculate the crossing, penetration, and derivative trend among disciplines. The background, development profile, groundbreaking achievement, mutual penetration, and future development direction of a discipline can also be analyzed. Thus, the dynamic structure and development law of science can be revealed. As early as 1974, Garfield used the computer system and citation analysis method to depict all major subjects, relations among subjects, and diagram of the new topic in the field of biomedicine from 1972 to 1973 to reflect the internal structure of biomedical research.

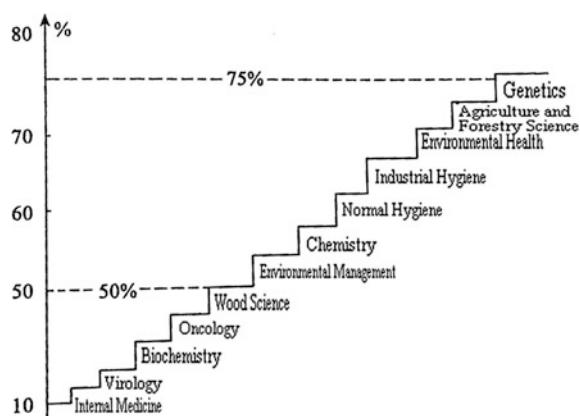
(3) Distribution of the discipline information source

Through the mutual citation relation among literature, one can analyze the source and disciplinary characteristics of literature references in a discipline (specialty) not only to learn the contact between this discipline and other disciplines, but also to explore the source and distribution features of information. Thus, a basis for information management solution and development plan of this discipline will be provided. For example, using the citation analysis method shows that the distribution of information source in environmental science is very wide. The papers from agriculture and forestry sciences, health science, biochemistry, and 12 other disciplines account for 75% of the total amount of citation (shown in Fig. 8.37). According to this feature, one can formulate an appropriate information strategy.

(4) Identification of core journals

The citation analysis method is one of the commonly used methods to identify core journals. The main feature of this method is to evaluate and select a literature from the perspective of the literature utilized. It is highly objective. Garfield used the citation analysis method to examine the cluster law of literature. He ordered journals according to the journals' citation rate and found that literature of a discipline contain the core literature of other disciplines. In this way, literature in all disciplines constitute a scientific, holistic, and multi-disciplinary core literature.

Fig. 8.37 Distribution of information source in environmental science



However, the journals that published these core literature are about 1,000 kinds. According to the analysis of JCR in 1974, in this 1,000 kinds of journals, literature published by 206 kinds of journals accounted for 50% of the entire core literature. The centralization law of literature can be used to identify core journals. Analysis of Chinese scientific journal citation indicates that literature whose citation amount accounted for 75% were from 72 kinds of cited journals. These journals were not only frequently quoted by the journals in this discipline, but were also the citation objects of other journals. Therefore, we can say that they constitute the core of Chinese natural science journals.

(5) Law of scientific communication and information transmission

The citation chain and citation network can be used to research the direction, processes, features, and laws of information to analyze the history and law of a discipline.

(6) Law of literature obsolescence and intelligence utilization

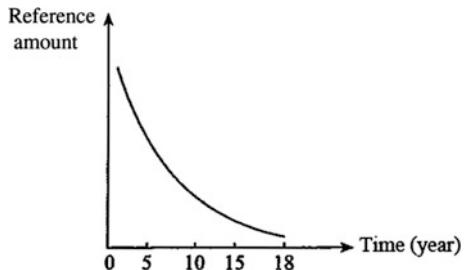
Currently, related research on literature obsolescence is generally from the perspective of the literature utilized. D. Price used the citation analysis method to explore the laws of literature obsolescence. Through an immediacy cited index and journal average citation rate analysis, D. Price believed that papers are composed of two classes of literature whose half-life is absolutely different, i.e., archival literature and the currently effective literature. The basic form of citation relation among scientific literature is citation time series. The citation distribution curve can be analyzed to measure the half-life and maximum citation life of journals in various disciplines to provide a basis for creating the best literature collection management program and for the quantitative analysis of literature utilized. According to the citation analysis method, scholars measured the half-life of various Chinese disciplines, as shown in Table 8.19.

Research shows that the citation distribution curve of a discipline is very similar to its aging curve. This strongly suggests that the citation distribution reflects the laws of literature obsolescence. Therefore, from the perspective of citations, it is a useful means to research the laws of literature obsolescence and intelligence

Table 8.19 Half-life of Chinese scientific journals (years)

Discipline	Half-life
Social science	3.8
Natural science	3.3
Mathematics	5.2
Physics, mechanics	2.5
Chemistry	4.7
Earth science, astronomy	4.7
Biology	4.8
Medicine, health	4.2
Agriculture, forestry	5.9
Technological sciences	3.2

Fig. 8.38 Biochemistry citation's distribution curve



utilization. For example, the citation distribution curve of biochemistry presented a negative exponential curve in 1980 (shown in Fig. 8.38).

(7) Demand characteristics of information users

It is an important approach to use the citation analysis method to research user intelligence. According to the citations of scientific literature, one can learn the demand characteristics of information users. In general, citations attached to the end of papers are the most representative literature that users (authors) need. Therefore, the characteristics of citations can reflect the main features through which users obtain intelligence via formal channels. A particular intelligence center analyzes the citations of its served users' published papers. It provides direct guidance. Through the statistics of citations from published papers of the same specialty, many indicators can be obtained, which are related to information demands. These indicators include citation amount, citation literature type, citation language distribution, and citation time distribution. To some extent, these indicators can illustrate the main features of users' information demands.

(8) Evaluation of scientific level and talent

The scientific capacity and academic level of the state and academic institutions can be compared and evaluated by using the indicators of scientific literature rate, duration, and so on. The citation analysis method is often used to evaluate talent because the extent of a paper cited by others can be regarded as a measurement that may be used to measure the academic value and influence of this paper. From the perspective of cited research results, it also can reflect the influence and status of an author in his discipline. Therefore, citation data provide a quantitative basis for talent evaluation. From the statistics of papers that cited the situation of Nobel Prize winners over the years, Nobel Prize winners in the field of physics, chemistry, and medicine have a high number of citations. The highest number of citations (L. D. Landau) is 1,888 times. The lowest number is 79 times (J.H.D. Jensen). There are only six Nobel Prize winners whose citation amount is less than 200 times. These elite papers have a very high citation amount in the scientific community.

8.6.2 Application Example of the Citation Analysis Method

To illustrate the application and steps of the citation analysis method, we provide an example that uses the citation analysis method to examine the laws of literature utilization.

According to the principle of bibliometrics, we use the dynamic citation analysis method to analyze the scientific papers on agriculture. The steps are as follows:

(1) Selection of the analysis object

We select research papers from four Chinese journals (*Chinese Agriculture Science*, *Genetics*, *Crops*, and *Soil Technology*) in 1980. We also select research papers from four foreign journals (*Agronomy Journal*, *Heredity*, *Crop Science*, and *Soil Science*) in 1980. Random sampling makes samples have an extensive representation, i.e., it can represent the general level of agricultural scientists rather than concentrate on a few prestigious scientists. Therefore, we use the random sampling method to obtain 100 papers from four Chinese journals and 100 papers from four foreign journals as the analysis object.

(2) Citation statistics

In the selected 200 papers, we add up the citation amount, citation published year, self-citation amount, citation languages, and citation literature type of each paper. In the randomly selected 100 Chinese papers, there are 1,055 citations. Among them, the largest citation amount of the paper is 39, and the smallest citation amount of the paper is 2. In the 100 foreign papers, there are 1,604 citations. Among them, the largest citation amount of the paper is 42, and the least citation amount of the paper is 3. These statistics roughly show that in the same 100 papers, the amount of citations in foreign papers is more than that in Chinese papers, and the average citation amount of each foreign paper is more than that of each Chinese paper. This reflects the gaps among Chinese and foreign authors in utilizing literature.

Further analysis requires packet processing of raw statistics. The typical packet group number (K) is obtained according to $K = 1 + 3.3 \log N$, where N is the sample size (i.e., N is 100). Therefore, K is 7.6, which means we can roughly divide the sample into eight groups. We can use the difference between the maximum amount and the minimum amount of each paper's citations divided by 8 to obtain the group interval (i). Then, we can set the group limit, group midpoint, corresponding frequency (f), level difference (d), and so on (Tables 8.20 and 8.21).

(3) Citation analysis

The two tables above show that most Chinese papers have 7–11 citations (frequency is 37), followed by those with 2–6 citations (frequency is 31). Meanwhile, most foreign papers have 13–17 citations (frequency is 35), followed by those with 18–22 citations (frequency is 22). The distribution of the resulting frequency impression is more profound than the meaning of the average. According to the data

Table 8.20 Distribution of citation frequency of 100 Chinese papers and simplified calculation table

Group	Group limit	Group midpoint	Frequency(f)	Level difference(d)	fd	fd^2	$f(d+1)^2$
1	2–6	4	31	-3	-93	279	124
2	7–11	9	37	-2	-74	148	37
3	12–16	14	16	-1	-16	16	6
4	17–21	19	9	0	0	0	9
5	22–26	24	5	1	5	5	20
6	27–31	29	0	2	0	0	0
7	32–36	34	1	3	3	9	16
8	37–41	39	1	4	4	16	25
\sum			100		-171	473	231

Table 8.21 Distribution of citation frequency of 100 foreign papers and simplified calculation table

Group	Group limit	Group midpoint	Frequency (f)	Level difference (d)	fd	fd^2	$f(d+1)^2$
1	3–7	5	12	-3	-36	108	48
2	8–12	10	14	-2	-28	56	14
3	13–17	15	35	-1	-35	35	0
4	18–22	20	22	0	0	0	22
5	23–27	25	14	1	14	14	56
6	28–32	30	1	2	2	4	9
7	33–37	35	1	3	3	9	16
8	38–42	40	1	4	4	16	25
\sum			100		-76	242	190

in the table, a citation analysis of each project can be carried out. Several aspects of this example are analyzed as follows:

- ① Concentration and dispersion tendency of the citation amount
- ② Theoretical distribution of the citation amount
- ③ Analysis of citation main indicators, including the analysis of self-citation amount, the analysis of citation language, the analysis of citation type, and the analysis of the growth of citation amount
- (4) Conclusion

Based on citation data and analysis, appropriate conclusions can be drawn.

- ① Utilization of agricultural science books and intelligence information can be explored by citation quantity analysis of agricultural papers. The results show that Chinese agricultural researchers should improve citation

quantity, and their intelligence work should be enhanced to provide a high level of literature utilization.

- ② To increase workers' service capability and research level on agricultural science and technology and library and information science, if a researcher masters English and Japanese, he will get 90% of information. Thus, language learning should be promoted.
- ③ In 2,598 citations, more than half are journals. Therefore, each library and information department should strengthen journal ordering and circulation work and appropriately increase classic journals. Besides, each department should examine common and edge periodicals to provide a basis for SDI service and selected journals.
- ④ The analysis result of 50 years of citation shows that the citation amount in nearly 20 years is predominant, which accounts for 80%. Besides, the citation amount in nearly 10 years accounts for 60%. Therefore, it is important to do well the work that provides books, journals, and new information in the nearly 20 years. At the same time, it is also important to research the aging and weeding of books and information.
- ⑤ With the progress of time progress, the citation amount increases rapidly. It fits the exponential curve, and its mathematical model is $y = ae^{bx}$. In this example, according to foreign papers in nearly 50 years, we can obtain a mathematical model. The trend curve equation is $\hat{y} = 4.39e^{0.098x}$. Researchers can use this model to estimate any year's citation amount in 50 years. Thus, the trends of literature utilized can be analyzed.

8.6.3 Limitation of the Citation Analysis Method

Although the citation analysis method has a wide application and meaning, it presents several limitations. Author citation is a human-controlled thinking and judgment process. As a manifestation of citation, it is just a macro and surface measure. Therefore, research on citation law requires enhancement. Many limiting factors affect citation analysis.

(1) Citations are not entirely equal in importance

For example, several papers may have wrong ideas or conclusions, and researchers may criticize them for further discussion. Thus, these papers' citation may be high. In the other aspects, papers whose citation is seldom are not entirely of equal unimportance. Citation is limited by many factors, such as publication time, language, and discipline. Minor differences in citations cannot completely describe the pros and cons of quality. It involves much randomness. Only when the difference is very large can the problem be explained.

(2) Accessibility influences the citations an author selected

M.E. Soper pointed out that most of author-cited literature are personally collected literature, and a small part of author-cited literature is in the author's sector and nearby library. The proportion of literature cited from other cities and countries is very small. It shows that if an author cites literature, he will consider accessibility and possession. The author's language ability, literature age, circulation of the literature itself, and the secondary publication reported also influence author-cited literature.

(3) Influence of false connection on the citation relation

Literature is cited for many reasons. Two papers cite an early literature for entirely different reasons or from different perspectives. For example, a paper may cite its methods, and another may cite its results. Thus, the link in the content of these two papers may be false. Citations may appear in the preface, title, text, conclusion, and discussion. In these cases, the content and extent of the original literature that the author cited are not the same. The relationship and importance of the citation for the original literature are dissimilar, but in the current citation analysis, they are treated equally and without distinction. A false relationship is easily caused. New papers do not have a large number of citations. The cited rate of small journals is lower than that of large journals, and the phenomena of cited but not used and used but not cited also occur. These citation phenomena affect the application and effectiveness of the citation analysis method.

(4) Influence of the Matthew effect

Several researchers believe that the Matthew effect exists in the aspects of citation. People always regard classic and authority as the selection criteria for citation. They do this out of necessity, but some do this to raise the social status of their papers. A journal that publishes celebrities' articles has a large number of citations. Therefore, this journal has a high citation rate. This psychological Matthew effect covers and affects the authenticity of citation.

Chapter 9

Methods of Computer-Aided Informetrics Analysis

Computer-aided informetrics is performed with the aid of computer systems during statistical analysis, simulation, and reasoning to yield corresponding statistical analysis and research results. With the improved performance of computer systems, current information measurement research, whether it is theoretical or practical, relies heavily on computers. From data collection to statistical analysis, all work is completed with the assistance of a computer. With the continued surge of information from documents and literature, computer processing of information is becoming increasingly popular, information retrieval systems are being widely used, and information measurement research is deepening. The computer plays a vital role in information measurement research, and computer-aided informetrics research has become a very important research content and a new development trend of informetrics research.

9.1 Significance of Computer-Aided Informetrics

9.1.1 Improve the Efficiency of Informetrics Analysis

Current research issues as well as research environment are largely different from those in previous times. The traditional manual method cannot meet current needs any more. First, bibliometric research progresses and deepens. Now, the full-text database text in any knowledge unit includes a word that can be used as a search portal and metering unit. The measuring unit of bibliometric research has entered the word level, so bibliometric analysis goes deep into the internal contents of documents. That is, the measuring units, such as a chapter and book, are converted into a knowledge unit and documentation information, such as title, keywords, word frequency, knowledge items, citation information, author, publisher, date, language, and format. Second, the quantity of document information has surged.

Owing to advances in technology, documents and literature, which are records of increased new knowledge, grow rapidly. Bibliometrics is a specialized discipline that focuses on quantitative and practical analyses. Both theoretical and practical research require systematic large-scale information as data support. These systematic and standardized data are a prerequisite for bibliometric analysis. Undoubtedly, the manual mode cannot obtain the required data for current bibliometric studies. However, computers can. To meet the challenges of statistics and data processing, computers work in a very efficient and effective manner. For example, they can quickly show data from the database for a specific bibliometric research and then process the data for analysis. If the work done by hand, it may take months or even longer, but the computer is likely to complete it within several hours or even minutes. Thus, computer-aided bibliometric and informetric research can obtain massive data in an efficient manner. Statistical analysis of past bibliometric research had two problems. First, past statistical analyses were carried out manually, so the workload was very large and work was inefficient. Second, data for past bibliometric analyses were not completely covered and obtained, with a very narrow coverage of subjects, too short time span, and single document type. Therefore, making full use of current computer-aided database retrieval systems is important for the realization of automated bibliometric analysis to cover more subjects and more document types and to improve research efficiency. Moreover, computers not only obtain the data accurately and efficiently; several computer systems can also analyze units of knowledge and information in documents and literature in a spectacular, efficient manner. For example, the American Statistical Package for Social Sciences (SPSS) can carry out correlation and regression analyses. The European Space Agency Information Retrieval System not only has an information retrieval function, but can also meet user's requirements. For example, analyzing data, such as keyword, author, organization name, code, and their frequency, can help understand the development of a discipline, predict its development prospect, possibly reveal what new materials, new technologies, new processes, etc. have penetrated into what areas, or reveal the relation of one technical field detected in another field.

9.1.2 Improve the Reliability of Informetric Analysis

The data collection of past bibliometric research was quite incomplete and had problems, such as covering narrow subjects, short time span, and single document type. With limited data coverage, past analyses generated comparatively lower credibility of research results and could not completely show a holographic picture of research progress in a certain field. The accuracy of data analysis results depends primarily on the typicality and completeness of data; otherwise, it will lead to one-sided conclusions. For massive data analysis, the traditional manual manner of analysis needs to be changed. It is necessary to make full use of the current computerized database system to realize the automation of informetric analysis.

Owing to the rapid development of methods for informetric analysis, data collection, data modeling, data processing, and data analysis have become more complicated than before to meet raised requirements. Several methods involve many parameters, and the modeling is far more difficult than that in manual capability. This requires computer aid, and computer technology development has opened up broad prospects for bibliometric analysis. The popularity of computers and the establishment of large document and literature databases provide strong technical support to data acquisition and statistical analysis for citation analysis in computer-aided bibliometrics research, which is a breakthrough in bibliometric analysis methods. Full-text databases, full-text retrieval, improved social information networks, and expanded electronic communication networks contribute to a more accurate bibliometric analysis with more accurate and reliable analysis results.

9.1.3 Promote the Accuracy of Informetrics Research

Since the last century, theoretical study of bibliometrics has achieved a series of breakthroughs, such as literature exponential growth law, literature dispersion law, and literature aging and citation law. The entire theoretical system has evolved, which now provides the theoretical foundation for computer-aided bibliometrics. Given that these rules are generally in mathematical modeling and very suitable for computer processing requirements, they lay a solid foundation for bibliometric research models. How to build mathematical models for bibliometric laws is the research focus of bibliometrics. The Ninth International Conference on Scientometrics and Informetrics in August 2003 held in Beijing included the “mathematical modeling of scientific development and informetric laws” as one of the topic themes. Bibliometric research has reached the scale where it needs to establish a large database with featured functions. China and foreign countries have established many databases. For example, the world’s most famous citation index, the American Science Citation Index (Science Citation Index) referred to as SCI, was founded as the American Institute for Scientific Information in 1963 by Eugene Garfield. SCI provides an international, multi-disciplinary, comprehensive index and online services supported by corresponding databases.

9.1.4 Expand the Research Area of Informetrics

Informetrics is featured with its quantitative analysis being based on big and systematic data. With informetrics being investigated further and data becoming large in quantity, the manual method for data collection can no longer meet the needs of researchers because sufficient and comprehensive raw data need to be covered as a basis of data analysis. Accurate data source is the key to guarantee the quality of citation analysis results. However, with the information explosion, data from

literature increase in size and widen in terms of time span. As a result, the traditional manual method lags far behind. Besides, along with the wide application of bibliometrics analysis, new concepts, analysis methods, index systems, and mathematical methods are emerging, such as matrix, model, and regression methods. These emerging new methods increase the complexity and unreliability of traditional manual data calculation. Thus, the trend of research methods for biliometric analysis needs to be modernized. With the popularity of computers and the development of information technology, the research methods of biliometric analysis are becoming increasingly diversified. For example, computer-aided citation analysis can do a much better, efficient, and reliable job of exploration. Another example is citation cluster analysis combined with keyword analysis, which is a new hybrid method for citation analysis. Co-citation cluster analysis presents the citation network that categorizes reference literature. The computation of the similarity and theme differentiation among similar literature as well as the test of clustering results are beneficial to expanding the miro-research on literature topics with scientific explanations. Quantitative and modeling analysis is a significant feature of citation analysis. The gray linear differential model can be fitted to diachronic citation data, which is a method for diachronic measurement of the aging of literature.

9.2 Feasibility of Computer-Aided Informetrics Analysis

9.2.1 *Computer and Network Technology as Fundamentals*

The earlier application of computers has shifted to the computing of non-numeric characters, which caused a revolutionary change in library and information science. Since the advent of the third generation of computers, hardware and software have greatly improved, thereby greatly enhancing the adaptability of library automation, such that computers have realized the processing of documents. At the same time, the computer information retrieval system has developed rapidly. The computer shows high capacity and efficiency in processing data. When data are inputted, the computer can perform various data analyses and output various results for different requests. The computer can also easily obtain dynamic data over document changes, add data, update data, and perform repetitious jobs in statistical calculation. Through the use of computers, complex calculation and model processing can be performed. Furthermore, data can be further developed by using the simulation computing technique, which deepens theoretical and practical research on informetrics.

Currently, computers can retrieve and filter databases, perform statistical calculations, and make charts. It is not only feasible but also extremely necessary to use computers and databases. It is also important for the development of multi-functional and integrated library and information systems. The current

computer systems in library and information science are mainly systems for document retrieval, which have the common problem of low efficiency. Among all the reasons for this problem, its single and undeveloped function is the major one. If the function of computer systems can be improved to meet information needs from various levels of customers and researchers, computer systems will undoubtedly not only attract customers in all aspects but will also expand information services at a large scale, which in turn increases the utilization frequency of the entire system. In recent years, much effort and many developments have been witnessed in domestic and foreign database systems, full-text retrieval systems, and multimedia systems. These systems are attractive to users because they provide statistical analysis results. Meanwhile, bibliographic retrieval systems have also been greatly enhanced to provide systematic bibliographic information and databases for statistical analysis. Artificial intelligence technology offers another possibility for the research development of computer-aided bibliometrics. Network technology and information resource network provide a basis and prerequisite for the statistical analysis of networked information of literature.

9.2.2 Development of Literature Digitalization as a Prerequisite

To use a computer to carry out research on bibliometrics, we must properly establish corresponding databases for informetrics research. The development of databases has expanded the measurement function, and the ability and adaptability of quantitative analysis are enhanced. As databases increase in size, quantitative analysis becomes more functionally enhanced and adaptive. In particular, the development of the US “Science Citation Index” not only provides an effective tool for informetrics study, but can also be used as a reference for the construction of databases in informetrics studies. With a citation database, statistical analysis can be automatically performed, and new documents of data can be derived for various studies. For example, the Hungarian Academy of Sciences Library used the SCI CD to set up the Science Citation Index database and has performed meaningful work in informetrics. It implemented scientific informetric indicators and set up a comprehensive quantitative index system of worldwide publication activity and citation impact. A computer can be used at any time for data analysis with various perspectives and purposes. The institution also studied the scientific communication process and tried a random model to analyze the process and determine the law for scientific publications and citations. Electronic publications can be adopted for informetrics and knowledge clustering and have opened up a new field for the application and development of informetrics. Electronic publications, as a full-text database, can be retrieved and statistically analyzed down to the unit of every word or every piece of knowledge. This makes it possible for bibliometrics to narrow down the analyzing unit from one literature to the unit of knowledge within the

literature, thus allowing informetrics to work at a deeper level. This is an important progress because it indicates that bibliometrics has entered a new stage of informetrics and will continue to develop. Many publically accessed information retrieval systems, such as the US DIALOG and China's "Chinese Science and Technology Journal Title Database," mainly serve as tools for information retrieval but are seldom used as data sources for bibliometrics. Actually, several of the functions of these retrieval systems can be used for bibliometric analysis, which should be paid more attention and taken advantage of.

9.2.3 Theoretical Foundation Laid by the Development of Informetrics

Whether it is manual or computation aided, bibliometric analysis should be guided by scientific theories to obtain sufficient and correct conclusions. At present, the theoretical system and analysis methods are approaching maturity. This is because on one hand, the research targets of informetrics and the laws revealed by the research can be mathematically modelled, which offers a possibility for computer processing. It provides an objective basis and possibility for the testing, calculation, and development of computers; on the other hand, statistical analyses for both manual and computer-aided bibliometrics have common features in sharing basic theories, but computer-aided bibliometric analyses are performed more efficiently on a larger data coverage. Thus, the basic theories and analysis methods of informetrics are not only applicable but also provide guidance to computer-aided bibliometric information and data analysis.

9.2.4 Research Development from Abroad Provides Experience

Although research on computer-aided informetrics is not as popular as that on computer-aided design, computer-aided manufacturing, or literature retrieval, in fact, the study of this field has been carried out abroad in earlier times. In particular, in several developed countries, research work is actually obtaining increased attention from the field of library and information management and has achieved many promising research results and progress. Some work used the literature retrieval system to conduct a bibliometric analysis, and some established databases for bibliometric analysis. Statistical software has also been commercialized and widely used. Research work is not only related to the analysis of theories, construction of databases, expansion design of functions, and program design, but also to broad and active practical applications. For example, the United States, Russia, Britain, Japan, and other countries used a patent literature database to carry out

technical evaluation and forecasting work. The British Derwent patent companies used a computer and analyzed the WPI library to obtain the corresponding topic words, synonyms, and eventually the patent. They could also obtain the yearly distribution of a patent of any company owned by WPI. The United States Patent and Trademark Office now has a huge database of patent documents. They used a computer to analyze various technological sectors, output research reports regularly, and make customized research reports for specific users. The American Institute for Scientific Information used the SCI database to analyze co-citation clusters, research situations in different countries, laws of scientific development, and talent evaluation. Russia uses the Chemical Abstracts (CA) database and the Russian chemical database to analyze the growth and changes in the number of the world's patents for polymer. Japan used INSPEC and CA databases to conduct retrieval and statistical analysis. Japanese researchers identified activities in physical, electrical, and chemical fields by applying the quantitative analysis method. All these facts illustrate that computer-aided bibliometric research is not only necessary but also feasible. They provide examples of successful experiences that can be learned from.

Since the 1990s, the study of computer-aided bibliometric analysis has increased in number, with plenty of achievements, and has been applied widely. Through theoretical analysis, computer-aided informetric research focuses on the design and development of computer-aided informetric software tools and achieves database reform and construction with the statistical analysis function for various kinds of resources. The establishment and maturation of computer-aided informetric analysis indicate that the research methods of informetrics have been basically formed and will mature. In recent years, the successful establishment and publication of the citation index database in China offer tools and conditions necessary for the automated statistical analysis of bibliometric information.

9.3 Theoretical Basis of Computer-Aided Informetrics

9.3.1 Major Types of Computer-Aided Informetrics Analysis

There are three types of computer-aided informetrics analysis. The first is through the use of a computer to construct a database for bibliometric research. Upon the construction of such a system, obeying the laws of informetrics, the design of different types of informetric information system is carried out according to different research purposes and requirements. For example, to establish an evaluation system toward core journals, the contents and characteristics of Bradford's law need to be observed to select criteria for the evaluation. Then, using certain algorithms, a computer is used to model the law and make block diagrams.

The second is to utilize and improve existing information retrieval systems. Based on the original information retrieval system, this method utilizes a computer

to add data and expand the range of statistics and analysis to fit documentary information metering analysis in accordance with the requirements of documentary information metering research. This method comprehensively improves those systems in accordance with the features and demands of documentary information metering research by sufficiently utilizing existing functions, such as increase related data record, expand statistic range, and enlarge quantitative analysis, which will make documentary metering analysis an assisted method and function in favor of comprehensive development and utilization. This method can address the dilemma faced by the low utilization ratio during the development of a bibliographic database and can offer efficient modernized methods and conditions for the computerization of bibliometric research. Whether in consideration of the economic cost and social benefit or design and technology, utilizing and improving existing computer information systems are important and effective methods to develop automated bibliometric research. Recently, a few large-scale information retrieval systems have improved their abilities of supplying data, fact information, and statistical analysis. The third is to utilize download and linking technology to establish a database specially utilized for documentary information metering. Adopting download is an important technological method to establish, transform, and develop related computer database systems that are fit for bibliometric research. It mainly utilizes a network database and adopts a means of communication to download data and store them in storage media for temporary or permanent utilization after transmitting. This method is applied to extract necessary data from various databases in a large computer system, store data on storage media for offline analysis research or further treatment, and establish personal or departmental databases for specialized application.

There are two methods of computer-assisted documentary information metering analysis in the data analysis view. The first is to utilize a computer software to perform correlation and regression analyses and to count the word frequency. The second is to utilize fundamental mathematical expressions of bibliometrics and computer software to establish a mathematical model and perform documentary information metering, which is more comprehensive, correct, and efficient than traditional information metering analysis. However, the fundamental theory of the two methods is the same.

9.3.2 Structure and Function of Computer-Assisted Information Metering Analysis

The structure of information metering analysis includes many systems. The first is the data treatment system, the main task of which is to complete the work of inputting, amending, and handling. The second is the retrieval system, the function of which is to supply the cited references and citing references, retrieve statistical results, and display, download, and print the retrieval results. The system includes a

user interface, an input analysis subsystem, a retrieval database, and an output analysis subsystem. The third is the statistical analysis system. With source literature and citation indexes as a base, it obtains statistical results on source literature, institutions, references, citation condition of journals, and outgoing message of authors. It also explores problems related to the theory and laws of documentary metering to form a new awareness.

The main functions of informetrics analysis are as follows. The first is data processing. Quantitative analysis methods are adopted in bibliometric analysis, and special data are needed to support the method. Thus, the function of processing data is a basic function of bibliometric metering analysis. Any bibliometric analysis system needs functions of assessing and sorting material. The second function is to provide statistics. A bibliometric analysis system can provide specific statistics in accordance with each kind of need, such as statistics on issued articles sorted by authors, institutions, and disciplines. The third function is evaluation. A bibliometric analysis system can make statistics and sort in accordance with issued articles and citation of papers and works to evaluate scientific achievements, talents, institutions, and print publication. The fourth function is query. Users can refer to cited references, source references, and count results by utilizing the bibliometric analysis system. Existing bibliometric analysis systems not only supply various retrieval methods, such as full text, author, key words, abstract, references, journal title, and institution, but also offer retrieval means, i.e., Boolean logic retrieval, truncation search, and quadratic search. As to the retrieval results, several retrieval system not only retrieve the full text but also similar articles, citing articles, and cited articles. We should confirm the data model first to make the computer automatically handle and produce document metering. We must handle, analyze, and make a conclusion for statistical data after calculating, sorting, and expressing raw data by utilizing the database retrieval system. For the data statistical software, we can select common software, such as SPSS, and common methods, such as multivariate analysis of variance, factor analysis, and clustering analysis.

9.3.3 Steps of Computer-Assisted Informetrics Analysis

Similar to traditional informetrics, computer-aided informetrics also uses quantitative analysis methods. However, it has advantages, such as collecting more comprehensive and accurate data, analyzing the data with higher speed, and obtaining more reliable results. The three stages of computer-aided informetrics are data collection, data processing, and data analysis.

(1) Data Collection

Scientific databases and information networks are the main data sources for computer-aided informetrics. With the development of computer science, an increasing number of databases of different types are constructed with improved

functions and data quality. To collect data, first, one needs to make a proper choice of database based on analysis goals. Second, one needs to ensure data accuracy and representativeness. Although collecting data from databases is more comprehensive and accurate than manual work, it does not guarantee that the data are not completely without mistakes. Thus, one needs to expand the coverage of data and check the data representativeness to avoid mistakes.

(2) Data Processing

Data collected by a computer are usually disordered. However, with data processing and categorization, the data can become more organized for finding their characteristics. Upon requests, calculations such as average number, numerical calculation, percentage, and logarithm are performed on raw data to obtain valuable results. Then, the data need to be arranged. Finally, the data are presented via statistical tables or charts for data distribution analysis.

(3) Data Analysis

Given that the purpose of statistical analysis is different, the methods to analyze data vary and are of two kinds. The first kind is verification law. It involves making a statistical analysis on data to check if the data meet the descriptive law and to verify the accuracy of the law under some laws of bibliometrics. The secnd kind is discovery law. It involves determining which distribution the data comply with in accordance with an approximate distribution condition in the statistical table or on statistical graphs. Then, it determines the possible obedient model distribution of the data and utilizes a computer statistical software to perform regression analysis with special data to determine the values of parameters. Given that we cannot affirm if this mathematic model is fit or not, we must verify the model. Analysis and research can be performed after verification. For example, one can utilize a certain algorithm to confirm the development condition of each researched object in each phase or predict the future developmental trend through comparative analysis.

9.4 Construction of a Citation Database and Data Mining Analysis

9.4.1 *Citation Analysis Database*

A digital citation analysis system must be based on computerized citation networks. First, there must be a documents database, records that reflect the objective reality in detail, and various literature and their characteristics. Meanwhile, the documents should not be isolated in the system but interrelated. A citation system should be able to reflect the real connection among literature, so the design of the system requires consideration of the file record structure. Second, the main task of a citation system is to promote the means of computerized citation analysis, so the system

must be able to make full use of computer resources and must be powerful. For example, it can achieve the functions of manual citation analysis system more efficiently, such as author, subject, source journals, and other matters of statistics; discover the distribution of citations; and analyze bibliographic coupling and the co-citation structure. Owing to the application of computers, a new citation system can achieve not only the full functionality of the manual system, but also some of the features that are difficult or impossible in manual conditions, such as publication of Chinese science citation index, journal study, topic detection, and forecasting based on large data sets.

A citation analysis system consists of three parts: data entry system, citation analysis system, and citation retrieval system. The data entry system is a typical type of processing system, and the citation analysis system is a typical analytical system; they are different. The citation analysis system must analyze the data extracted from the input system and reorganize the data according to analysis process needs. The data warehouse can be used to do this. To return the query result from the citation retrieval system rapidly, the data must be organized effetely. A variety of index databases, such as sources index, citation index, author index, group author index, and keyword thesaurus should be constructed.

9.4.2 Citation Analysis System Design

A citation analysis system is an integrated system with multiple functions, which can be divided into the following subsystems.

- (1) Documents Subsystem: It contains document inputs and maintenance tasks and is the key to the entire system and the bases of other subsystems. This system enables the functions of input, modify, and delete documents, data conversion, and establishment of citation relationships among literature. It is a typical transaction processing system that requires frequent additions and deletions and other data access operations. The amount of data in each operation is small, the processing time is short, and data integrity and referential integrity constraints are required strictly. It can be constructed according to the theory and method of the traditional database system.
- (2) Index Database: The index database organizes the data in different ways to extract data quickly and accurately. It includes sources index (for records of the collections storage), citation index (for the storage of the citations of collections), personal author index (for a personal author's information storage), group author index (for group authors' information storage), key word thesaurus (for source documents' keywords storage), and so on. When users submit a query, the system will extract data from the keyword thesaurus or from the personal author index directly. In a word, the index database extracts data according to queries from the input analysis subsystem and transmits them to the output analysis subsystem.

- (3) Retrieval Subsystem: The retrieval system is designed for two goals. First, it retrieves literature from the database in different ways, such as item retrieval and group matching retrieval. Second, it retrieves citations, which mainly include archival retrieval and quotation retrieval. The latter needs to supply a fast, effective, flexible, multi-approach and all-around retrieval method to help users find their necessary information quickly; thus, it has a very high requirement on response time (i.e., it should extract data from the database within several minutes and even seconds after receiving users' retrieval requirements).
- (4) Statistics Subsystem: It utilizes the calculation function of computers. This subsystem makes quotation relationship statistics in accordance with the features of each kind of literature rapidly and precisely and outputs such statistics in tables or graphics.
- (5) Intelligent Analysis Subsystem: It utilizes each kind of statistics data. This subsystem performs comprehensive research under citation analysis theory, orderly outputs core journals or papers, or performs law verification.
- (6) Compilation Subsystem: This subsystem realizes editing and publishing of Chinese Science Citation Index (CSCI) and editing of source citation, subject citation, institute citation, and journal citation reports.
- (7) Output Analysis Subsystem: This subsystem evaluates the type of extracted data and converts it into the type selected by users. For example, the output type only contains author, title, journal title, and publish time, whereas the users require the abstract to be added. Given that it is different from the data input system, its data amount is small or would not be updated. Access to large data is involved during each retrieval, and it has low requirements on response time. Moreover, related data cannot be obtained in the data input system under its data mode for further analysis with a high degree of synthesis, but special data extraction must be performed, and a mass of intermediate data must be calculated. Lastly, analytical requirements cannot be determined beforehand.
- (8) User Interface. The user interface of a retrieval system often offers retrieval approaches, such as title, keywords, author, journal title, and full text and retrieval methods, such as logic retrieval, truncation retrieval, and quadratic retrieval, to allow users to clearly express their own retrieval needs. The input analysis subsystem judges the users' input type, content, and actions that the system should adopt after receiving users' retrieval requirements from the server. This function includes analyzing which retrieval approach has been adopted by users, extracting the corresponding input value, determining if there are matched items, and extracting special retrieval and matched words to construct a retrieval strategy and transfer the strategy to the index database.

The main program design of a citation analysis system should select proper programming language and utilize the strong function supplied by such language to accelerate the development speed and realize the goal of the system efficiently. The key point of citation system research is the quantitative relationship. With philology knowledge, we know that a congruent relationship exists among quotation

literature. Several such relationships form a table; therefore, selecting a relationship database to support the system and embody such a congruent relationship is the best choice. The design of the data structure directly affects the realization of a system's functions. Therefore, we should design the data structure seriously in the initial stage of system design and construct the best database mode to allow the system to effectively save data and meet the information and treatment requirements of each user. If problems arise after a database is operated and after performing modification and maintenance, the cost will be very high. Entity-relational graph is a common method for the design of a database's logic data structure. It extracts an entity, determines the entity relationship, and sets up a database conceptual model by knowing and analyzing the real world. With research on the literature stream and the citation relationship among literature systems in the real world, we would know that the involved entities in a citation database are mainly document collection, author group, topic dictionary, and periodical collection; the involved correlations are citation correlation, paper–author correlation, paper–journal correlation, and paper–keywords correlation.

Several issues should be considered in the integrated design of a citation analysis system. The first is the establishment of a literature database. To reflect the relationship among references completely, the references of literature often have a wide coverage, long time span, various discipline scope, multi-source journals, and diverse literature types, which increase the size of the citation database and impose a heavy work load for the establishment of the database. Therefore, it is necessary to confirm a proper reference collection scope before the execution of the system to guarantee a certain coverage and a system that is not excessively large. Given that basic databases in reference systems are ordinary literature databases, the data in existing literature databases can be considered to transcribe and set up correlations among literature through proper methods and to establish reference databases to reduce repetitive labor and relieve onerous data input work. The second issue is the efficiency of statistical analysis. A citation database is often enormous, so we must consider how to organize data and design the best algorithm to improve the efficiency of program operation, reduce response time, and increase the usability of the system when we perform retrieval and statistical analysis in the database. The third issue is achieving a friendly human–computer interface. Designing a friendly interface to allow users to use the system conveniently has become an important problem that system designers and developers must face. A successful system would not force users to consider data in a non-natural manner, so designers should make the system consistent with several general systems. The data input should meet operation habits in the manual system as much as possible, practical orders should be easily understood, and the data output must be clear and perceptual as much as possible to offer related online help or hints in accordance with different users.

9.4.3 Case Study on the Design of the Online Edition of Chinese Social Science Citation Index (CSSCI)

Chinese Social Science Research Center of Nanjing University began to develop the online edition of CSSCI in 2000 to allow users to retrieve information on CSSCI through the Internet. At present, the center offers its service on the Internet (its website is <http://www.cssci.com.cn> or <http://cssci.nju.edu.cn>).

(1) System Goals

The online edition of CSSCI provides open and convenient information service to users. Individual users want to look for the influence of their published papers, and research institutions hope to know their research abilities and comprehensive strengths through the information of CSSCI. Therefore, the CSSCI system set up the following goals in accordance with the requirements of users. First, against users' need, i.e., looking for the source of some achievement, the system supplies source document browse and retrieval, which mainly supply recent published literature and related information for searchers. At the same time, users can select a necessary literature on the basis of the source literature retrieval result and obtain its references and related literature. Second, users need to look for the latest literature that is related to their published papers in early time. The system supplies a function of cited reference retrieval, and users can know the latest development of a research through its cited condition. When users cannot log on the Internet but have retrieval needs, the online edition of CSSCI supplies a delegation inquiry that can supply a convenient information service for users.

(2) Structure Design of the System

A completed B/S mode is adopted by the online edition of CSSCI in accordance with the above goals. For managers of the system, its function is to manage users. Managers only need to input correct commands. Then they can browse and alter users' information and fees, so managers can know the requirements of users on the Internet in time. For retrieval users, the system mainly supplies information service, so openness and interaction are required for users to retrieve through online pages on the Internet. Program codes, data, and supportive software of the application system all concentrate on the server under the B/S mode, and any supportive software is unnecessary in the client; only a browser is needed. Such a B/S structure reduces the load of the client, benefits the convenience and operation of the system, and makes the most of the open and interactive function of WWW browsers. Considering the convenience and mobility of user retrieval, the system design has two different retrieval interfaces, i.e., source literature retrieval and cited literature retrieval.

The CSSCI database has three function modules, i.e., data processing, data inquiry, and statistical analysis. The primary aim of the data processing module is to complete the input, modification, treatment, and maintenance of the digital dictionary of systematic data. The data inquiry module supplies a fast, effective,

flexible, multi-approach, all-around retrieval method. The statistical analysis module regards the bibliographical and citation indexes as a base to produce statistical analysis results, such as source literature, institute paper, quotation situation, cited journal, and quantity of authors' published articles. These statistics are powerful analytical tools for promoting a discipline's development, motivating each institution's academic activity, rearranging core journals' rankings, and knowing the core author groups of each discipline.

(3) Function of the System

Browsing and retrieval of source literature. To provide a comprehensive retrieval, source literature retrieval supplies the main retrieval item and the subsidiary retrieval item. The main retrieval item consists of several access points, such as author's name, indexing subject term, article title, journal title, and institution name. The subsidiary retrieval item includes access points, such as authors' regions, classification of disciplines, publication date, and article type. Users can match the above access points freely to select their own retrieval results.

Cited references retrieval. As for cited references retrieval, the system supplies the following retrieval points: author names, titles, and type and publication date of a cited reference. Users can perform a retrieval by combining the retrieval points.

User management. Users can make a retrieval after registering online, filling in tables, and paying for some fees. The website can control the quantity of concurrent users in accordance with users' retrieval types.

(4) Design of the System Interface

The interface is a bridge between the system and users. The knowledge background of users on the Internet varies considerably, so the design of the user interface should be concise, convenient, and simple to make each user view the functions of the system with half an eye and complete necessary information without any special training. Therefore, the system should put all operations on the same screen to reduce the operation time. The user interface of the system consists two parts. Users can perform source literature retrieval, cited literature retrieval, and user information management conveniently through the menu buttons on the above. Retrieval access consists of the main retrieval items, subsidiary retrieval item, and submit button. Retrieval results can be displayed on split screens or scrolling display; users can opt to check detailed information of retrieval results, including references and related literature. The retrieval interface of source literature consists of menu buttons and access points. Considering the requirements of various users, users can utilize three matching methods, i.e., precise, similar in the front, and vague matching for several access points, such as author name, journal name, institution name, and grant. As for article title and indexing word, the system offers three input boxes to input a phrase, and the logic relationship among the three input boxes is "AND." For example, if an individual user needs to search an author's source articles, he/she can input the author's name in the author input box. If a user wants to search "information" in the journal name box, he/she will get 0 article through

precise retrieval, 691 articles through similar in the front, and 1225 articles through vague retrieval. Such a retrieval strategy can meet the requirements of various users, and the combination on one interface is convenient for the retrieval of various users. The cited literature retrieval interface also consists of menu buttons and retrieval points and has the same function as the source literature retrieval interface, except that the access points are different.

(5) Charge Management of the System

Given that the CSSCI project caused large fees on its research and production, a user charging design is adopted. In consideration of the retrieval requirements of users, there are three charging modes, i.e., package system, rating system, and entrust inquiry system. As for group users, such as schools and research institutions, who hope to make a comprehensive data retrieval, the package system can be selected (i.e., a special institution can use the database on any computer at any time within its limited IP address scope). Such users should fill in the quantity of concurrent users and the IP address field when making an application to pay for different fees. As for individual users who just need CSSCI data for a research or a personal situation, they can select the rating system and can make a retrieval after paying for a certain fee. The website will deduct users' fees in accordance with the retrieval results of each time. A delegation inquiry is available for users who need data but have difficulties logging on to the Internet. The website will charge for certain fees in accordance with the delegation inquiry and its results.

(6) Construction of the Database

As an important component of the retrieval system, the CSSCI database sufficiently utilizes the strength of Nanjing University Library on information collection, classification indexing, and processing and treatment. A total of 496 Chinese humanities and social journals published in 1998 and more than 60,000 source articles and 28,000 cited literature are available. In the near future, the database plans to collect more than 200 Chinese humanities and social science academic journals in Hong Kong and other regions in the world, and journals published before 1998 will be collected successively. The database consists of source article, related literature, cited literature, and related code databases and will be updated in real time to stay in accord with the electronic edition database.

9.4.4 Mining Analysis Methods of Citation Data

(1) Statistical Analysis Methods

Informetrics is based on statistical literature information, and statistical methods are the fundamental methods of informetrics. The main application methods of statistics consist of descriptive statistics, such as table, measuring scale marker, graphical

representation, central tendency measure, and discrete measure; probability theory, such as probability distribution function; inferential statistics, such as hypothesis testing and significance testing; and sampling theory and multivariate statistics, such as multiple regression and correlation, principal component analysis, and multi-dimensional scaling. Statistics is applied to explore the internal quantitative relationship among research literature information, such as quantitative balance among the primary, secondary, and tertiary literature; quantitative relationship among literature's contained information; quantitative relationship between literature and authors; quantitative relationship between literature and science technologies; quantitative relationship between literature and time, country, and language; and quantitative relationship among literature, citation, and users. Citation analysis based on literature statistics describe literature's systematic behavior process; analyzes, compares, and arranges literature in accordance with citation indicators, such as journal, author, article, language, literature type, time distribution, and citation rate; examines citation laws; and evaluates quality. It also explores the flow direction law of literature information, literature obsolescence, and the concentration and dissociation laws of papers.

(2) Clustering Methodology

On the basis of quantitative analysis of citation, performing an analysis on the quotation network relationship and an analysis on co-citation or multi-co-citation clustering for several important and representative analytical objects will reveal the structural features of a discipline, research fronts, development source, professional degree of correlation, and scientific communication methods (Qiu 2000). Co-citation analysis and co-citation clustering analysis are effective methods to investigate the citation relationship and microstructure of literature.

1) Co-citation Clustering Analysis

Co-citation clustering analysis can be utilized to explore the research structure and situation of a discipline or subject. On this basis, a continuous co-citation clustering analysis on highly cited papers of a discipline and subject will dynamically reveal the changing circumstances of the discipline and subject. The process of such an analysis involves collecting references of related literature in a subject and selecting papers whose cited times exceed a threshold value; the threshold value can be determined by sample size and research goals and by referring to other factors properly, matching up with highly cited papers, making statistics of co-citation times, and attaining the co-citation matrix of highly cited papers. The co-citation matrix is changed into a related matrix through a coefficient, and a proper testing index and clustering methods (include the selection of familiar coefficient and clustering process) are selected to obtain co-citation clustering mapping. For example, certain distance methods (maximum, minimum, or average distance) can be adopted to perform clustering analysis and obtain a clustering analysis tree diagram. In quotation co-citation clustering analysis, the more similar the subjects are in the same cluster, the larger the differences are among clusters. The level of

similarity of subjects within the same cluster can be revealed from the size of the calculated value. The larger the subjects are among clusters, the better the co-citation clustering effect is. Co-citation clustering analysis of continuous and dynamic highly cited papers can reflect the structure and development process of a subject.

2) Co-word Clustering Analysis

The main principle of co-word clustering analysis is similar to that of co-citation clustering analysis. It makes statistics on co-occurrence times in the same papers for a group word and then performs clustering analysis on the co-occurrence base to reveal a close or distant relationship among those words by analyzing the structural change in their represented discipline and subject. The indicator of co-word clustering analysis is word (descriptor or free word), and the relationship among words represents a conceptual relationship. Therefore, the established clusters after clustering treatment should simply and clearly reveal the structure and change in a discipline or subject. Studies have proven that co-word clustering analysis on high-frequency subjects is feasible with a clear expressive pattern. Compared with traditional bibliometrics methods, i.e., simply making statistics on subject words, determining rankings, and analyzing research fronts, co-word clustering not only looks for high-frequency words but also emphasizes the correlation among words to reveal the conceptual relationship better.

Co-citation clustering analysis explores the main concern of people at present through the citation situation in the past; co-word clustering analysis is a direct statistic on published literature at present to reflect the concerns of existing papers. Therefore, the function and results of the two analyses are similar, but the approaches are different. Co-word analysis is simpler and easier to be understood by people. Complementarity exists when people examine a science structure by combining co-word clustering analysis with co-citation clustering analysis. Co-citation clustering analysis can classify important references, form a reference network graphic, research and reveal the internal relationship among literature, and describe science development and dynamic structure. Given that co-citation clustering analysis regards the full text of a literature as a unit, its pertinence, research depth, and precision are obviously insufficient in revealing the similarity of a literature's subject compared with content analysis. Other abstract word, topic word, classification code indexing word, and content word analyses aim to improve and optimize the effects of co-citation clustering analysis to make up for its deficiency. Co-citation clustering analysis regards literature as a unit and focuses on the classification of literature, whereas co-word clustering analysis focuses on the topic content of literature. Co-word clustering analysis can perform a similarity inspection for literature groups produced in clusters. It is beneficial for in-depth, specific research on literature subjects and for making scientific explanations by calculating the similarity of literature subjects in the same cluster and the differences among clusters and testing the effect of clustering. The combination of the two ways is of unique strength.

3) Subject Term Chain Clustering Analysis

By utilizing the strength of literature keywords to perform clustering analysis, subject term chain clustering analysis is more convenient in terms of technology and more extensive than co-citation clustering analysis. Its theory is as follows: if there is one or more similar keywords in two or several scientific literature, a potential relationship must exist between the two (or several) corresponding literatures. Such a relationship is called the scientific literature keyword chain. This method can predict the relationship between research fronts of a discipline and related literature.

(3) Link Mining Analysis Method of the Citation Index Database

Most index databases offer related links at present. For example, the Web of Science design has many interlinks in hypertext format. These interlinks can help users find their required information and offers related evidence for knowledge discovery at the same time. The links of Web of Science are of two kinds, i.e., internal data link and external data link. The internal data link includes the following:

- ① Cited times link. Users can see the cited times of a paper in a retrieved source record, i.e., the full record of a paper. Users can know where the paper is from and what other papers cited it. Besides, users can see its complete bibliographic record and cited times. Users can know which papers cited the article through the “cited times” hotlink and can control the situation and development of a research field at present through layer upon layer links.
- ② Bibliography link. This link shows the bibliography list of the current source record. Bibliographies cited by authors, such as books, journal articles, patents, and other literature, can be retrieved. Users can see the full record of the source record to understand the development history of a research subject and reveal the researchers’ disposition when they absorb the research achievements of their predecessors.
- ③ Related record link. Users can see a group of citations that cites one or several similar articles published in different years with the current retrieved record, i.e., related record, and sort these co-quotations by their relevancy. The more similar a citation is to the current record, the closer the subject is to the current record. Therefore, the article’s place in the bibliography will be in front. Related record is an extended retrieval approach, and there is no need to adjust or replace the previous search strategy. This approach reveals the relevancy among research projects and offers continuous and correlative materials for a research project. External data links connect with information products from each kind of sources, and users can access necessary information through a uniform retrieval interface. Users can see original literature of the current record directly through the “Full Text” link, which includes collected original journal literature as a full text link. The library public directory link tells users if

there is library holding record; a secondary and tertiary document information content database link, such as Derwent innovation index database and social sciences and humanities conference proceedings database links, is also available (Qiu 2000). The relationship and level of citation can be analyzed through link mining analysis and clustering, and correlation analysis can be performed through the similarity of subjects reflected by citation chains.

(4) Correlational Analysis Method

The mining of citation data through correlational analysis of knowledge discovery will reveal the related laws of scientific development. For example, the inheritance of scientific development can be revealed in the citing and cited relationships of literature. The mining of citation relations will reveal the vein of scientific research and determine the historical evolution process of scientific theory and method. Bibliographic coupling and co-citation analysis will reveal the overall development trend of science and probe scientific crossover permeability and the law of scientific development. Systematic synchronic and diachronic analyses on science through the citation analysis method can examine the hierarchical structure and dynamics of scientific development to explore its development trend. Scientific and ultra-micro-structure relationships among professional substructures can be explored by combining co-citation clustering analysis and multi-dimensional scale analysis. Histogram and net-like relationship produced by the year distribution of citation can reveal the science production background, general situation of development, breaking achievement, and direction of future development. The results of citation analysis can reveal the status quo of a country's national policies and talent strategy on one side and explore the relationship among scientific research, policies of science and technology, and cultural background on the other side. When we explore transnational scientific communication and the relationship between the information output and input among various countries, the data volume of citation can be viewed as an indicator for measuring the number of times a country cites the research achievements of another country, and the quantitative analysis of each countries' citation will reflect the comparative level of each country's scientific development. Quantitative research on transnational scientific communication is realized through a quantitative study on information exchange, but the crux of quantitative research information exchange is the construction of the citation crossing matrix. First, the research field, countries, and representative journals should be selected. Second, a citation statistical analysis is performed. As a measurement index, the quantity of information exchange can be measured with the quantity of literature cited by one country from another.

9.5 Application of Computer-Assisted Information Metering Analytical Method

The application range of computer-assisted information metering analysis is wide, and the analysis is not only applied to the library and information field but also to related fields, such as management science, scientology, sociology, and science of personnel and history. For example, the CSCI database is equivalent to the SCI database; it not only supplements the deficiency in quantity of Chinese journals but also supplies data as judgement evidence of science. At present, the CSCI database is the assigned inquiry database for selecting the academician candidates of the Chinese Academy of Sciences and National Natural Science Fund for Distinguished Young Scholars as well as for the performance evaluation of projects funded by the National Natural Science Foundation of China (NNSFC) during latter periods, professional title evaluation of colleges, universities, and scientific research institutions, achievements declaration, and promotion check and evaluation. CSSCI, similar to citation retrieval tools (e.g., SCI), possesses unique retrieval functions and points that other retrieval tools do not have and a quantitative tool to evaluate authors, institutions, academic production, and the influence of regions for scientific institutions and management departments. Users can learn the deep influence produced by academic papers, journals, researchers, and institutions.

9.5.1 Application in Scientific Research

All scientific research and technological creation is reflected in words, so the quantity and quality of scientific literature can reflect the technological development situation. Knowing the content structure and quantity change in scientific literature through computer-assisted bibliographic information metering statistical analysis can evaluate or analyze the history and situation of a technological development and predict its development trend. The dynamic situation of a discipline, such as penetration and intersection, shift of developing emphasis, and locating position of a discipline, can be evaluated by utilizing the citation relationship of scientific literature and making statistics on the literature quantity of each discipline. Scientific research is meant to innovate, which is to discover things that have not been found by predecessors, explore unexplored regions, solve problems others cannot solve, and introduce unbeknown knowledge. Therefore, any scientific paper is not isolated, and the production of any paper should refer to related papers as theoretical foundation, compare objects and the data source, supplement those papers' deficiency, explain authors' innovation, or make a historical review of a research field. We can trace the opinion or process of a discovery and know whether these views and creations are applied, amended, executed, developed, or innovated through mutual citing and citation among papers. Therefore, computer-assisted bibliographic information metering analysis can offer a theoretical foundation and

evaluation and guiding information for scientific research innovation. Selecting source journals through a rigorous assessment and high-standard selection system and utilizing the mutual citation relationship can reveal the internal relation among academic research and thus allows researchers to conveniently control the intersection and interaction among disciplines and fields and supply high quality, all-around, reliable information material for project approval, planning development, and in-depth scientific research.

9.5.2 Application in the Information Source Field

The application in information source field includes the following:

- ① Determination of core journals. Core journals refer to a few important journals that publish a large number of professional papers with a high utilization factor in a discipline or professional field. Core journals change with the development of a discipline. Core journals should be re-determined every few years to reflect new changes in journals. The strengths in evaluating the quality of journals and determining core journals are very obvious through computer-assisted bibliographic information metering statistical analysis, which not only accesses data with a wide scope and large quantity with comparatively objective evaluation results but is also accurate and fast. Hence, determining time is reduced obviously.
- ② Theoretical research that deepens informatics. A large amount of completed data can be obtained, and classical laws of informetrics can be tested, calculated, and deducted through a computer system to promote further research on the theories of informetrics.
- ③ User study. Users' behavior, interests, and hobbies can be quantified through a computer system, and the quality of a library and information can be continuously improved based on these data.
- ④ Document classification. Citation is an external feature of literature. The co-citation and quotation coupling phenomenon among literature reflects the relevance of documents. Utilizing such a relationship can provide valuable information for document classification. Such an application can be subsidiary and supplementary of traditional classification methods. An independent classification method is unreliable and infeasible because its precondition is that existing articles in the database have been classified.
- ⑤ Journal citation reports dynamically evaluate and sort journals in each research field and list the impact factors of those journals. Doing so not only traces the work of processors and reveals the cited situation of researchers' work, but also identifies the research groups, institutions, and scholars with the most achievements. It supplies a large amount of information for identifying the focal points of an interdisciplinary subject or

developing a new research field. Moreover, it can reveal and measure the influence of literature results on scientific research and offer information for expanding new research fields. It supplies information for libraries and institutions to select and subscribe to journals in accordance with the evaluation results and rankings.

9.5.3 Application in Competitiveness Analysis

We must control the developmental situation of patent literature at home and abroad and what technology has been patented by other companies in a field to offer references and evidence while avoiding repeated work when we give project approval, research and development, patent application, and patent developmental strategies for a special technological field. When we perform quantitative analysis on hundreds of patent literature in each field, we must search, select, count, tabulate, and draw these literature. By utilizing computers, we can retrieve, select, count, and tabulate patent databases. Each type of statistical analysis graphics can reflect each country or company's quantity and content of patents. People can control the developmental history, status quo, and future development trend of each technological field and the competitive situation of each company in each country.

Computer-assisted bibliographic information metering statistical analysis is strict with regard to the selection of the database retrieval system. It requires researchers to understand the structure and software of the selected database retrieval system and predefine if it meets the established research target or not. As traditional manual or self-built database design structure and software in accordance with research goals, existing databases and their retrieval systems at home and abroad are designed and produced for information retrieval service, which does not fully conform to the needs of bibliometrics research in several respects. Therefore, the selection of databases, especially the professional database, must meet the requirements of established research objectives first. Second, we must pay attention to the standardization of subject terms and indexing depth. Otherwise, a unilateral and even incorrect result will be obtained.

9.6 Development Direction of Computer-Assisted Information Metering Analysis

9.6.1 Development in Breadth and Depth

Continuous development from a bibliographic database that includes recording authors, book (paper) titles, and abstracts only to the current citation index database and full text database improves the depth and width of computer-assisted information metering analysis. A citation index database can offer several index

methods, such as citation, patent citation, institution, and keyword permutation indexes, which supply not only strong data support for computer-assisted information metering analysis but also a multifunctional tool. A full text database regards a single word as a metering unit, which further increases the depth of computer-assisted information metering analysis. Currently, each kind of professional or comprehensive database continuously appears simply because original data supplied by a database become increasingly comprehensive, thus widening the scope of involved disciplines, such as sociology, economics, pedagogy, management science, medical science, and agriculture.

9.6.2 Development in Practicality

The application scope of computer-assisted information metering analysis research is wide. Aside from the library and information field, the research is applied to science management, scientific decision making, scientific prediction, and even scientific technology. The indicators of bibliometrics can also be used in the evaluation of talents, scientific achievements, scientific institutions, and even the technological level of an entire country. Given that the data and research achievements of computer-assisted information metering analysis are comparatively reliable, they can offer evidence for the management and decision of related departments.

9.6.3 Development in Integration

As a classical analytical system, the citation analysis system must extract data from the input system, re-organize them in accordance with the needs of analytical treatment, and set up a unique analytical treatment environment. A data warehouse is a type of data storage and organization technology that deals with such a new analytical treatment environment. Specifically, the data conversion program extracts source data and transfers them into a target model between the input and analytical systems and then uploads them into the data warehouse; the system can extract multi-dimensional analysis data through an online analytical processing service. Analysts access multidimensional databases through online analytical service and perform citation analysis. The organization mode of the system's source data is application-oriented, whereas the organization mode of the analytical system's data is subject-oriented. The latter's level of abstraction is higher, and the exchange of these two kinds of data is the key to building data warehouses. Online analysis is a software technology related to the data warehouse and an analytical operation specially designed to support complex analytical operations. Its multidimensional data analysis mode is online data access and aims to address special problems. The multiple possible observation methods for information and the fast, stable,

unanimous, interactive access allows analysts to pervasively observe data. Multidimensional data analysis mode looks upon data analytical work as an operational process that rotates and cuts into blocks or slices a data cube that consists of variables and dimensions. Variables are of practical significance to data and numerical value metering indicators cared for by people. The dimensions are specific perspectives that people observe. Multi-dimensions and variables constitute a multi-dimensional data structure, i.e., a data cube, but the design of the cube is the key problem of multi-dimensional data analysis. The nature and features of online analytical treatment make it a powerful tool of citation analysis. Events have proven that data warehouse technology and online analytical treatment technology are particularly fit for the requirements of the citation analysis system (Qiu 2000).

9.6.4 Development in Modelling

The fundamental laws of bibliometrics offer theoretical direction and evidence for the mathematical modelling of computer-assisted bibliographic information metering statistical analysis. A computer system with strong functions is a tool of mathematical modelling. Once a mathematical model of computer-assisted bibliographic information metering statistical analysis has been set up, statistical work becomes easy. Users only need to input data into the model and set conditions in accordance with the objectives, and results will be produced soon. In this manner, computer-assisted bibliographic information metering statistical analysis is not only simple and accurate but also highly efficient.

9.6.5 Development in Intelligentization

Computer-assisted bibliographic information metering statistical analysis can be divided into three phases of computer-assisted data processing, system support, and intelligent phases.

- (1) Data processing phase. The data processing workload increases in bibliographic information metering analysis, so we need utilize computers to execute processing, editing, classification, and counting work. A universal software is adopted during this phase. This phase improves efficiency and widens the application scope.
- (2) System support phase. The major objective of this phase is producing an analytical system via a computer and database not only from a single work but the entire work to support bibliographic information metering analysis and realize automatic processing at a high level and with a wide scope. The main task of this phase is set up a special database and design analysis system or rebuild a database in accordance with the content of related databases.

- (3) Intelligent phase. As a high-level phase, the intelligent phase should not just satisfy logic inference, quantitative calculation, or fixed routine but should also be capable of flexible analytical adjustment, multipath inference, dealing with fuzzy problems, and fuzzy recognition to realize a shift from data processing to knowledge processing. Computer-assisted bibliographic information metering analysis utilizes new achievements, such as data mining technology, and looks for implicit and unknown but useful information from a large set of completed sensitive information that are unrelated through various methods, such as correlation, sequencing, clustering, and classification. It also reveals the inner complexity of data, performs an in-depth analysis, and automatically accesses highly valuable information.

Chapter 10

Application of Informetrics in Information Resource Management and Research

Informetrics has a wide application domain. The application domain mainly involves two aspects. First, informetrics can be applied to library and information science. Second, it applies to other relevant areas, such as management science and meta-science, policy study, forecast study, study of making the best use of people with special abilities, sociology, history, and science and technology management work. Informetrics application in the library and information science domain involves library science, information science, literature study, and archives theory rule research application and includes practical application in management. It is not only a powerful tool of books information science rule research; it may also serve as a library and information science work scientific management strategy for one basic theory. All laws and measurements of the analysis method of informetrics may be applied in library and information science work, such as selecting the old, collection optimization, information retrieval, reader work, information service and library integrated management system construction, appraisal and network information collection service, and other aspects in literature, such as establishments and appraisals, fund assignments, editing decision making, circulation management, books formulation, and booklist appraisal and selection, and collection policy. It has a universal significance and characteristics. The application scope is wide, and content is rich.

10.1 Informetrics and the Determination of Core Journals

Research on and determination of core journals are important components of application in the library and information science domain. Such an application is early and is still maturing. Moreover, the usage and widespread influence are very important in the application domain and have a certain representation and persuasive power. Here, we utilize literature informetrics theories and methods, discuss the core journal rationale, and form the significance of concepts, and determine the mechanism and core journals as well as the core journal literature gauging device system.

10.1.1 Theoretical Basis and Formation Mechanism of Core Journals

Intelligence has various values, which play different roles in the exchange of intelligence. The magnitude of the exchange of information value is often determined by the distribution of the papers in journals. Thus, what is the distribution? Does it follow a certain regularity?

For a long time, many scholars have carried out extensive research on these issues. The earliest and most fruitful scholar in this area has to be Professor S.C. Bradford, who is a famous British scientist of literature. He engages in science research and cultural heritage of long-term work and mentioned that for some topics, professions, or course realms, the distribution of related thesis in periodicals is very asymmetric. However, they may also be scattered in periodicals in other courses. The quantity of these “other courses” periodicals diminishes with the closeness degree between the realm and course, and the density of a related thesis of this course in each periodical increases. On the foundation of this type of qualitative analysis, Bradford performed a covariance research and found that the “virtuous blessing” furthers the fixed amount. He announced to the public that the distribution of a thesis in periodicals exhibits regulation; this is called the Bradford virtuous blessing cultural heritage dispersion law. This law not only explains the objective existence of the core periodical, but also reveals to the public the cultural heritage to distribute concentrated and long-lost regulation. In a periodical thesis’ physical distribution, this type distributes the phenomenon to have “catholicity.” Therefore, people know it as a concentrated and long-lost regulation that the cultural heritage distributes.

Given that the science cultural heritage distribution of concentrated and long-lost regulation is objective and existent and has widespread applicability, this means will cause the core periodical to come out. Obviously, the science cultural heritage distribution of concentrated and long-lost regulation is the existent theoretical foundation of the core periodical and is also the basis of core periodical measurement work. By setting out from this regulation, one can use two theories to explain the formation mechanism of the core periodical at the least. The first is checking and supervision subjected to the objective regulation of science development. This is the objective demand developed by the course because of the creation and development of a science periodical decision. Each kind of periodical has its own course and professional property, so course thesis height concentration contains a handful of core periodicals, thus becoming the cultural heritage to distribute “the heap add effect.” The second is the creation and development of the subjective factor of some. Artificial control also influences the cultural heritage to distribute the core periodical. For example, the influence of the “horse too effect” causes the cultural heritage “heap add” to have a handful of core periodicals and become the cultural heritage distribution of concentrated phenomenon. In the meantime, cultural heritage produces the exchanges process. There is so much influence that the phenomenon needs to be subjected to factitiousness to select the

factor, but an individual's choices must be controlled by "the most labor-saving rule" (beg the near law). For example, the author of a thesis generally wants to select the profession close to it, and a good reputation besides "the most labor-saving rule" can announce magazine contributions. Various periodicals want to choose the profession they agree with and whose quality is high. The result of this artificial choice causes the cultural heritage to distribute the medium heap add effect, so large quantities of related thesis are concentrated in some periodicals that have the core property. They constitute the core periodicals of a certain profession. If every choice is viewed as a success, this kind of successful accumulation will certainly easily cause new success. The mechanism of "success, creation, success" is believed to be one of the basic reasons for core periodical formation.

10.1.2 Concept of the Core Periodical and the Important Meaning of the Measurement

Core periodicals focus on one course or professional realm and publishes in great quantities such professional theses. Such important periodicals are called the core periodicals of that course (profession). The instruction indicates the thought and purpose of the measurement. Core periodicals are of two basic types: course core periodical and core periodical hidden in a library. Although these are two different concepts, they have a common characteristic and advantage. While there is no circumstance that particularly explains it, the core periodical generally means the core periodical of the course. From this characteristic, core periodicals all have common characteristics, such as objectivity and opposite and dynamic state. Cultural heritage produces information communication. The core periodical is objective and existent and subjected to checking and supervision. It influences the cultural heritage distribution of concentrated and long-lost regulation and is different because of the measurement method. These periodicals are the core periodicals for one course, and they may also be core periodicals to another course. A core periodical is not fixed and unchangeable. It presents dynamic state characteristics with the development of the course and the variety of the cultural heritage realm. Therefore, measurement of a core periodical cannot be performed once only. We need to constantly adjust or revise every five years to reflect the latest variety of the periodical. The measurement of a core periodical involves calculating cultural heritage information to learn physically applied important contents. In studying the evaluation of periodical quality, the scope of the core periodical must be ensured to acquire a high-definition intelligence report source and build a hidden book intelligence report unit. For many science workers, this professional periodical contents have an important meaning. The following three aspects are involved.

Currently, the quantity of periodicals is very large. Each unit wants to order all these periodicals. Aside from the absence of necessity, it is also impossible. This is because each course and each profession have a handful of core periodicals. As

long as we have the choice to order a handful of core periodicals, a great part of information requests can satisfy the readers. In addition, if ordering to publish up is pointless, it will only divide the time and certainly will result in hoarding, mismanagement, and poor service of the cultural heritage. Manpower, material resources, financial power, and stock are wasted. Therefore, for the quality of research and the evaluation of periodicals, assurance is given that core periodicals continuously increase their publication rationally and accurately. This is an important research topic of the book intelligence report realm and is also an important link that raises the book intelligence report unit science to management level.

Science and technology workers demand for point reading.

Incomplete statistics indicate that about 60,000 kinds of science and technology periodicals exist in the world currently. Theses annually account for 4,000,000 and above. With CA as an example, it is a periodical containing excerpts of about 15,000 kinds, and excerpts on cultural heritage value at more than 500,000. A scientist or engineering technical personnel facing this huge quantity of cultural heritage would think that this course scope inside all of the cultural heritages is impossible to browse or read. Undoubtedly, any research personnel has the choice of point reading all of these. If the core periodicals are controlled and checking is performed on the use of science and technology cultural heritage, then the effort would be reduced and the results would double. The energy spent on the task is decreased, and more intelligence reports can be acquired.

Demand for exaltation information index and cultural heritage information service efficiency.

Only by understanding the characteristics and merit and shortcoming of various periodical, controlling the emergence regulation of the important thesis and collect the path, holding a core periodical of high quality, then we can have already to specifically worked well on the information index and the literature information service. When answering the consultation and providing the selective dissemination of information service, we also have to make full use of the core periodical of related professions to improve the efficiency of literature information service. Therefore, research and measurement of core periodicals are important foundation work for building up a reasonable area for storage and for opening the exhibition of cultural heritage intelligence reports.

10.1.3 Measurement Method—The Method System of Informetrics

With regard to periodical quantity, a small part of the total amount is occupied by the periodical. Regardless of whether a course is still in a technique realm, the periodical quantities related to it are large. How to choose from many periodicals

and a certain handful of core periodicals is not an easy task to accomplish. Relevant research and objectives are important in the measurement of core periodicals. The usual choice is a quasi science method. Since the concept of the core periodical was put forward, the problem of how to measure the core periodical has persisted. Currently, the measurement of core periodicals adopts cultural heritage information to calculate the method; this has already become a method system. Although the measurement method is varied, it basically has two types. The first is the core periodical that makes use of the information to calculate the tool and the index sign to directly select each course. For example, the United States' SCI and its by-product JCR provide the quantity and influence factors of periodicals, calculate the index sign, and compare these data's nature to directly make a selection of the core periodical. The second is to use cultural heritage to calculate the method according to basic steps for measuring the core periodical. If the line is divided from the standard of the measurement, it would mainly include four types of methods (six kinds of concrete methods) as shown below. The first one regards the amount of text as the method of the standard, including the surname law method and percentage, to compensate and accumulate 100 centers.

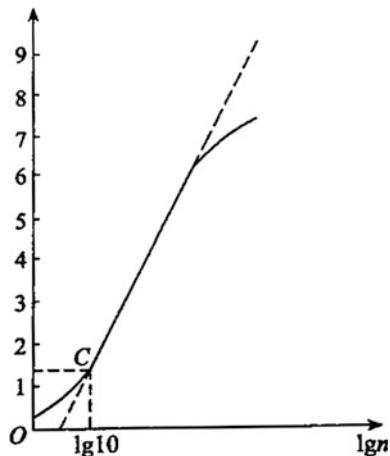
1) Bull laws method and percentage repair

Bradford's "blessing" laws for measuring core periodicals provides a scientific method. This kind of method has been extensively applied in core periodical measurement for each course. Generally, the cloth is used to implement virtuous blessing law measurement of core periodicals in three ways. The first is the district analysis method. At some point in time, one may expect a periodical to contain a course or profession and the statistics of each related thesis that publishes the text. The second one is the grade that applies the cloth to implement virtuous blessing arranges the technique, arranges the text from big to small one by one, and divides these periodicals according to the concrete circumstances into several areas (generally three areas) to make the cultural heritage quantity of each area mostly equal. At this time, the periodical of area 1 is the core periodical of the course's realm or profession.

2) Method of image analysis

This method plots according to the statistical data of rank and then uses the method of image analysis to determine the core journals. Its detailed operation is that first, the number of relevant papers in various journals are counted and accumulated into sets of data: n and $R(n)$. Second, plotting with LGN as the abscissa and $R(n)$ as the ordinate is performed to obtain Bradford curves. Third, the intersection C of the rising curve and linear part is determined, where C corresponds to $\lg nC$ and nC is the number of core journals. Finally, nC kinds of journals are listed and ranked ahead. These journals are core journals of several disciplines (see Fig. 10.1).

Fig. 10.1 Determination of core informatics journals



3) Law of Brinell and compensation of percentage

Although Bradford's law has been successfully used to determine core periodicals and create core journal directories, it also has certain limitations. One is that the rank of journals under Bradford's law is according to the absolute amount of published papers on several topics. In this kind of statistical method, the small-scale professional journals whose published papers' amount per stage is less but which meets research needs are likely to be excluded from the core periodicals. To make up for this limitation of Bradford's law in determining core journals, a researcher abroad put forward the "percentage distribution" method. This method creates statistics and calculate the papers of a topic published by every kind of journal accounting for the percentage of all papers in the journal. Then, it ranks the various journals according to the percentage. Next, it selects a suitable proportion (such as 50%) as the standard. Journals whose percentages are higher than this percentage are identified as the core journals of the discipline. Obviously, the advantage of this method is that percentage calculation has nothing to do with the size of the journals, thereby ensuring that small but closely related professional journals are not eliminated from the core journals and that core journals are in accordance with the actual usage situation in the professional field. Practice has proven that using the method of "law of Brinell and compensation of percentage" to determine core journals is effective.

(1) Cumulative percentage (80%)

This method lists the journals of a field in descending order according to the papers' amount and then accumulates the total papers' amount in n kinds of journals ranked ahead. Then it calculates the percentage between the papers' amount of the first n journals and the statistics of total papers' amount in all the journals. If it has reached the requirement of the selected percentage, the first n journals are regarded as the core journals; otherwise, it must calculate the journals ranked after the n kinds of

journals until the required percentage is achieved. Given that the general required percentage is 80%, this method is also known as “80%.”

These two methods regard the relevant papers’ amount of journals as a standard to determine the core journals. Considering the amount of journals’ published papers, to a certain extent, it can satisfy the requirements of the determination of core journals, and it is relatively simple and direct. However, Brinell’s law and the cumulative percentage method have several disadvantages, such as they simply emphasize the number and does not evaluate each kind of journal’s quality. Quality is also an important criterion to determine core journals. Furthermore, they only consider the absolute number of journals’ published papers for a topic and disregards some professional literature’s percentage in the journals’ total papers (i.e., the relative). This causes several journals that are professional but do not have a large amount of published papers or have a long publication period to not be selected as core journals. Compensation of percentage can solve this defect (the cumulative percentage method can also compensate for the percentage), thus making the two methods of determining core journals perfect and practical.

(2) Abstract

The abstract method is one of the commonly used methods for the determination of core journals. Its basic idea is that according to the frequency size of the secondary literature excerpts, journals are arranged according to directory, and it regards journals as an important degree and the choice of the basis. A journal whose papers are extracted or indexed frequently is optional for core periodicals. The abstract method determines core journals as follows:

1) Select a suitable statistical abstract magazine

The abstract method must select objects of the subject within internationally famous abstracts with authoritative magazines, citations, or indexes as the basis of statistical analysis. When a subject has two or more abstract magazines, selecting the suitable abstract magazine for statistical analysis is particularly important. Usually, choosing the medium digest magazine is suitable. In determining the period of journals and abstract used for statistical analysis, one should try to use recent volume, generally with statistics for a year. The statistical period is too short, and the core periodical lacks accuracy. With regard to core journals obtained from the statistical period, the longer they are, the more accurate the core journals are but the more time consuming. Therefore, a one-year period is relatively moderate.

2) Extract and sort a statistical journal

Statistics for a year, including the amount of various kinds of periodical literature abstract excerpts on a certain subject, are extracted and sorted according to abstract quantity.

3) Calculate the cumulative rate of abstracts

From high to low, the cumulative amount of abstracts is calculated for the proportion of the total extracted quantity of journal literature (a). If the cumulative abstract rate reaches 70% (the standard), the corresponding journal can be identified as the core of the discipline of the journal.

4) Cross comparison

If a subject has many famous abstract journals, one is selected as the main body. By using the above methods, the abstract journals' core periodical table (referred to as the main table) is created. At the same time, another famous abstract journal is selected as auxiliary, and its own core journals are created in the same way (called tables). Then, the side tables are compared and verified. If two core journals in the table contain a publication, it remains. If one is in the table and the other is not, then the publication is deleted. After verification by checking the rearrangement in the table, a serial number is assigned (if a particular issue with two serial numbers in the table is not at the same time, then it is subject to the main table). After the two working procedures of core journals, accuracy and reliability become high. The following are the limitations of the abstract method.

- ① The absolute abstract is in accordance with the journal to be sorted, so several small but professional journals are not considered as core journals.
- ② Several journals with many readers and have a high use value of science popularization, advertising, and application of technical periodicals, which are not abstract magazine excerpts, are excluded from core journals.
- ③ Abstract rates tend to be subject to an abstract source and abstract part quality, such as all kinds of limit conditions. Actually, there are more important journals and articles, which will influence the accuracy of the determination of core journals.

Nevertheless, the abstract method is still the most effective method to determine core journals. Statistical comparison of the use of the method alone produced accurate and reliable results.

(3) Quotations

The citation method is one of the commonly used methods for the determination of core journals. The basic idea is that according to the journals' size and frequency of literature citation directory, they are arranged in order. If according to the former, a large number of reference journals would be regarded as a discipline of core journals. According to its different particular way, the citation method can be applied to the following situations.

1) Citation analysis of a variety of magazines

By selecting a variety of magazines as the female parent of a subject, a paper publishes statistics on these magazines and their reference number. Then, analysis of the origin of these references is performed, and the original journals are arranged according to the citation frequency. Those at the top are classified as the core of the discipline of journals.

2) Analysis of a single journal citation method

A subject in the field of a magazine is selected as the authority and is commonly used as the female parent. According to the above method, citation data from 1 to 3 years can be classified as the core of the discipline of the journals. The standard according to the number of cited specific circumstances is provided on its own. These methods can adopt one or two steps. When two cases need to make citation statistics and sorting can be completed, this is often referred to as one step. The so-called “two steps” is based on one step; it selects a larger number of core journals as the female parent journal, citation statistics are created, and sorting is implemented to obtain more comprehensive and accurate core journals.

3) Using the method of citation tools

The United States SCI uses the core journal citation method to provide powerful tools and conditions. By analyzing SCI during a given period by a citation number and its source distribution, it is found that 500 to 1000 journals are the most important. The journals basically meet the needs of 80% of people. Therefore, SCI before 500 journals can be regarded as multi-disciplinary core journals. The multi-disciplinary core periodical table is also the reference for the determination of professional core journals and can be used directly.

JCR can analyze not only several magazines' cited frequency but also the average number of times each paper was cited and the index of the year. Based on these indicators, one can evaluate the quality of journals, determine the core journals, and form many concrete methods of citation analysis.

The use of JCR to determine single-subject core journals is extremely convenient. Accepted only from a certain subject representative of a journal source, we can obtain periodical departments of the reference cited journals. Then, addition to these journals' cited times makes the total cited times large, so they can be identified as the subject of core journals.

For public libraries and comprehensive information centers, given that they involve a wide range of professions and disciplines, the multidisciplinary core journals need to be determined. If JCR is used in the determination of multi-disciplinary core journals, one can check the cited journals and select the cited times that are higher in the journal queue table. This journal table uses the SCI database output directly by a computer, thereby avoiding the look-up table directly. Garfield, for example, analyzed 2000 journals' cited situation. Martin (Martyn) and others used this method for a full sci-tech periodical evaluation in Britain.

The citation method from the angle of periodicals by the author is used to evaluate the quality of periodicals, and the result is more objective. At the same time, this method can be used to measure professional core journals and determine multidisciplinary comprehensive core journals. If SCI, JCR, and other large tools are used, the heavy burden of manual statistics can be reduced, and the accuracy and efficiency of citation analysis can be improved. All in all using the citation to determine core journals is indeed an important method; therefore, it has been widely used. However, the citation method also has its limitations as follows:

- ① The cited rate is only a reference standard of success for journals and is not the only standard. Only high utilization rate of periodicals is an indicator of the high quality of a journal. High cited rate is not equal to high utilization rate. When only the cited rate is high, the reading rate is also high to show its high utilization rate. For example, several journals' cited rate is high partly because of high self-cited rate and high utilization rate of several practical publications, but its rate of cited is not necessarily high.
- ② New journals cannot obtain a large number of references, and some authors use them without lead. Large periodical utilization means the probability is obviously high. Journal citation data, therefore, reflect the actual usage error.
- ③ Citation analysis often only pays attention to the citation analysis of journal articles. It does not devote sufficient attention to records, proceedings, technical reports, and other types of publication.

(4) Circulation rate

The periodical frequency use (borrow to read) by readers can also produce a proper order catalogue that reflects readers' effective demands; it can be the basis of assurance of the core periodical. Concretely speaking, a magazine can be hidden in a building in a certain period; it can be lent out a number of times, read a number of times inside the building, and replicated outside. Analyzing the usage number of times and performing a statistical analysis would reveal that the one with high circulation frequency is the core periodical. When analyzing a periodical's circulation, we should also consider the brush-off rate for use outside and for replication (borrowing). The method of circulation measures its quality. However, measuring the quality of the current periodical of this building is limited by its unit and native area circumstance. In addition, the same reader could again borrow a publication to read.

(5) Comprehensive evaluation method

The above discussion shows that the periodical is quoted from the BE excerpt. The exploitation circumstance of the periodical is explained by the current number of times as the basis for the importance of the periodical and ensuring the core periodical. However, they have a shortcoming. In consideration of the short repair time

and the accuracy and usefulness of core periodical measurement, we can synthesize three methods and put forward a more ideal comprehensive evaluation method. The principle of the comprehensive method is shown in Fig. 10.1.

The P, Q, and Rs in the diagram represent leading grammar, selected essays method, and circulation that the core periodical gains, respectively. P, Q, and R mutually constitute seven areas.

- 1) $P \cap Q \cap R$ District: BE is the district that P, Q, and R mutually constitute. They represent from the source of quotation and excerpted with the current periodical. They are counts for core periods.
- 2) $Q \cap$ District of $P \cap$: Represents the source of quote and excerpts, but a few drive the current periodical.
- 3) $P \cap \cap$ District: Represents the periodical that is quoted from.
- 4) $\cap Q \cap$ District: Represents the periodical that is excerpted from.
- 5) $\cap \cap R$ District: Represents only the drive current periodical, but the time that is quoted from and excerpted is very low.

According to the importance degree of each area periodical, we can arrange in descending order the row order form. This sequence can be the basis of the selection of core periodicals.

- It is quoted from and excerpted with the current periodical.
- The periodical that is excerpted and circulated, but BE is seldom quoted from.
- The periodical that is quoted from and circulated, but BE is seldom excerpted.
- BE circulates, but since it is seldom excerpted, the periodical's BE is also seldom quoted from.
- BE is excerpted and quoted from, but a few drive the current periodical.

The other two areas are then not considered. In physical applications, the method is synthesized, and generally good results can be obtained. This is because the thesis can be announced by the periodical, and a selection process ensues. Currently, many periodicals adopt a certain quality standard of employing the rate of 10%; generally speaking, a thesis is formally announced by the periodical. The periodical thesis is excerpted and BE is quoted from, so BE is also the result of a type of selection. It is the concrete reflection that attains a certain quality standard. Therefore, the thinking that the exploitation of cultural heritage is calculated to learn the measurement method of core periodical BE "thinks greatly of the quantity but disregards quality." This standpoint is unilateral and obviously does not correspond with fact. However, we also have to awaken the ground to know that the above-mentioned measurement methods all set out from an angle and measure with a kind of standard; this will inevitably lead to limitations in sex. This is no wonder and is normal. Currently, under the circumstance that various methods still cannot completely overcome the proper limit of sex, we must work hard to investigate a new measurement method of science. Therefore, we should promote synthesis to evaluate the method and its expansion and attain a fixed-amount measurement method that combines different approaches. The two BE measure the method and

combine the qualitative analysis method to grow and repair each other. Only then can core periodical measurement be more accurate and physically match.

10.2 Informetrics and Documentation Information Collection and Management

10.2.1 Determine the Best Program for Periodical Collection

For the book intelligence report unit, how make sure the cultural heritage collects the project and how to make a building that keeps the best appearance always are the foundation of book intelligence. BE is a sexually important mission.

Brooks first presented the concept of periodical performance (utility). Given that the intelligence report value or the performance of a cultural heritage decreases with the increase in age, periodical performance also immediately deceases. However, the periodical performance's decline cannot be directly measured with the age of that periodical. Therefore, he used a year of some theses of one periodical inside the number of times that is quoted from as a performance measure. He supposed that ① the performance concept demonstrates the opposite and is only suited for the periodical of a concrete course and ② the performance of each periodical collection reduces at the same speed. According to the above-mentioned performance concept and assumptions, we can draw up a performance with an opposite calculation periodical with the dependable method of the aging degree. With P from 100% periodical total amount and for assurance (%) of the best project of the periodical, we have three approaches to adopt.

(1) The course carries the tallest control method in the text rate.

The simplest method draws out the n kind to create

$$N \ln\left(\frac{n}{S}\right) = N \ln\left(\frac{N}{S}\right) \times P\%.$$

Then, we have

$$n = S \cdot \left(\frac{N}{S}\right)^{\frac{P}{100}}.$$

From the top equation, we borrow n. For example, if N = 600, then S = 1.5, P % = 50% hours, and n = 30. If P% = 76%, then n = 90.

In addition, collecting a course of 50% is enough as long as the related thesis carries the most amount of text for 30 kinds of periodicals. The advantage of this

kind of method is can order P% of the periodical budget and requires the least control. However, it does not consider the old, and the past bookcase utilization is low.

(2) Select and remove the time control method

Compared with periodicals with a large collection quantity, the total performance of some periodicals is low. These periodicals are selected and removed through the calculation

$$a^t = \frac{U(t)}{U(0)} = 1 - P\% = f.$$

ϕ BE does not make use of the mutual benefit value. For example, if $\alpha = 0.87$ and $P\% = 0.5$, then $t = 5$ (year). If $P\% = 0.25$, then $t = 10$ (year).

The advantage of this kind of method is that it collects as a whole. Its weakness is that it requires a time standard to select and remove periodicals. It could select and remove another periodical of the core periodical, but regardless of the surplus performance ratio of the core periodical, the performance is still high.

In addition, according to author surname laws, a scientist needs a reference from A and B with a periodical thesis that reads C (three types of periodicals). For each kind of periodical, the collection strategy should be according to the unit of property but should exhibit differentiation. If an organization or library demonstrates a comprehensive temperament for a natural science, then we should lay particular emphasis on collecting author surname laws in the C periodical and A periodical of each course. The C periodical is mainly consist of important periodicals of comprehensive and technical science. The A periodical denotes the core periodical of each course. An important C periodical of the A and B periodicals should emphasize the collection of this course. The B periodical can be used as long as it is acceptable. The B periodical mainly denotes the non-nuclear periodical of this course realm. Certainly, this is just a principle demarcation. In actual work, the best project is collected according to the concrete cultural heritage of circumstances and the above-mentioned method.

10.2.2 Select the Best Means to Collect Literature

The cultural heritage collection work of a library and intelligence report organization varies and involves ordering, exchanging, making duplicates, etc. Under this circumstance, determining which kind of method is more economic and more reasonable is worthy of research. For example, book intelligence report units can order a relevant course as a part of the periodical. If a reader needs another thesis in the periodicals, he/she can adopt the replication approach. The problem is to request assurance, which periodicals have to order. However, detemining which periodicals

the thesis has to pass through for replication is more economic and valid. We can make use of the author to reveal the blessing laws to resolve this problem.

A_N is established as the average list price of a certain periodical. P increases for this periodical concerning several N articles of one course at the average price of replication. Thus, $A > P_r$ makes duplicating these articles for publication in the periodical of these articles worthwhile. The problem lies in the type of periodical and the distribution grade ordinal number. This kind of periodical increases then grows each time the periodical's annual ascends. The proportion of the article amount in the total amount is small in r . If N amounts of periodicals exist, then N/r is the relevant article that the entire annual publication needs to accumulate. The number is not smaller than the r of that kind of periodical with the lowest series. The R value can be from $r = A/P$ type, and N is the predestination. However, the N/r value may require some corrections because it is released by the straight line segment. It should satisfy the condition of $N/r \geq c$, in which c is the core periodical to count. We use

$$S = 1 - \frac{1}{r} - \frac{1}{N_r} \left[R(N) - R\left(\frac{N}{r}\right) \right].$$

The N and r values in the equation are already known, $R(N)$ and $R(N/r)$ can show the method to calculate via a diagram. In the diagram, the inclined rate k of the straight line must be part of the diagram. $R(N)$ and $R(N/r)$ can be adopted. S would economize the budget for comparison. In actual work, S can look like the formula calculation.

$$S = 1 - \frac{1 + \ln r}{r}$$

For example, if $N = 2000$ and $r = 5$, the diagram shows the method to calculate $S = 0.48$ then economize the 48% budget. The adoption looks like the formula calculation. If r approaches 1:00, S approaches 0. Ordering the periodical budget is related to making a duplicate among articles. The specific value of the budget is not very large, so adopting this kind of method is not useful. Currently, foreign countries adopt the above-mentioned method. Locally, technical improvement must be served and replication expenses must be reduced along with the intelligence report. Only then will this kind of method achieve an expansion in application.

10.2.3 Using Bradford's Law as a Literature Purchasing Strategy

When a library or information center gathers information, how to reasonably distribute the limited funds for purchasing is an important subject worthy of studying.

In terms of periodicals, there are two problems to solve. One is how to determine the quantity of periodicals, and the other is how to select and buy high-quality periodicals preferentially. That is, a reasonable purchasing strategy must be formulated. A periodical acquisition strategy should be in accordance with the first unit of funds and conditions to determine a possible collection target (such as the proportion of the total cover paper). Next, the number of journals to be ordered is verified, and Bradford's law is applied to select those that are most able to achieve this goal and have the highest density of journal information.

(1) Determine the appropriate requirement proportion

To obtain 100% of literature about a topic, a large number of information sources is needed. For most libraries and information centers, meeting the requirements of this collection is impossible. Even if the proportion is reached, it is not worthwhile from the perspective of cost effectiveness. For example, to make the collection range from 80% to nearly 100%, one needs to spend much money. This 20% behind the collection scope is likely to require the same cost as much as the 80% and could even be more than it costs. Literature purchasing should therefore be according to the needs of information service, and funds may determine the appropriate requirements. Generally, the accounts of literature in this field should be 90 to 50%.

(2) Calculate the quantity of periodicals

According to the requirements of the predetermined proportion of papers, we can calculate the number of journal literature by using the formula. For example, there are 500 kinds of journals in a field. Half of related publications in this field need to be collected, namely, $\phi = 1/2$ if $S = 5$. It should be booked as follows:

$$n = S \left(\frac{N}{S} \right)^f = 5 \times \left(\frac{500}{5} \right)^{\frac{1}{2}} = 50 \quad (\text{species}).$$

The subscription rate of the top 50 journals can meet the requirements. If the average annual subscription fee for each journal is 11 yuan, then ordering 50 journals needs 550 yuan. Meanwhile, if you know the amount of subscription fee and journals, using the formula can also determine the proportion of papers.

(3) Priority order core journals

According to the principle of Bradford's law, the vast majority of journal articles are taken from a small number of core journals. The basic principle of literature information gathering is to collect the highest density of intelligence information. Therefore, to achieve a predetermined collection goal, we must priority purchase core journals related to areas of interest. In addition to using the method described in the first quarter i to determine core journals, we can also use the Bradford distribution curve to determine the scope of prior purchasing according to the journal articles. In this regard, Bradford's law has practical guiding significance.

10.2.4 Using Literature Obsolescence Law to Guide Book Weeding Out

The obsolescence law of scientific literature has instructional significance for weeding out old data in time and building the best collection.

Periodical collection needs cost. To save money, many foreign libraries have begun to weed out journals when they reach a certain fixed number of years.

(1) Using the literature aging index as a guide for document removal

Literature aging data are an important basis for drawing out old data. For a subject, book intelligence has the following uses.

1) Half-life is used to determine the time of weeding out

Literature half-life is an important measure of literature aging. It provides the time basis for weeding out literature. Literature aging speed, namely, half-life is not easy to change. The literature aging law presents a long-term trend conservation. In determining the collection for a fixed number of years, it can be measured using the obtained half-life of related disciplines' concerned data. Book intelligence can determine the half-life of literature according to the half-life method. Then, according to the collection of key factors, such as inventory capacity, the collection time can be determined.

2) Aging coefficient is used to determine removal plan

The literature aging coefficient is an important indicator of literature aging because the aging coefficient is conserved. When book intelligence working in literature is rejected, we can formulate a relatively good calculation for sex and aging degree according to previous studies on journal aging coefficient. This can guide the weeding out of books. Developing a relatively good sex and aging degree of journals and a reliable method has important practical significance for the work of book intelligence.

3) Price index is used to determine the removal plan

The price index is an important measure of literature aging. It can be used in all literature of a particular field. It can also be used to evaluate several journals, an institution, a particular author, and the aging characteristics of an article. Distinguishing between "present" and "archive" literature will guide the weeding out of literature.

4) Use the remaining useful indicators to determine the removal plan

Journals are weeded out only after the journal profits decrease to a certain extent. Otherwise, the old periodical collection cost savings will not be enough to pay for the cost of replication. If they are not eliminated but saved as a miniature journal, it must also be after the profit has declined to a certain value. The remaining useful indicators provide a measure of a certain type and content of the journal in terms of

intelligence requirements. Book intelligence can determine the remaining useful literature and then make a decision on whether they would be removed.

Either way, it involves the optimization problem of the periodical management. Using the literature aging law to identify the best solution of periodicals collection. Therefore, the literature aging law has a great role in the periodical management.

(2) Other literature measurement indicators are used to guide the elimination

In addition to literature aging data, many indexes can be used as a basis for weeding out old data.

1) Published paper amount

Periodical management needs a benefit index. The reduction in journal benefits cannot be directly measured by the age of each journal, namely, the number of years since publication. A simple example is as follows. Journal a published 100 articles related to a subject all throughout the year. Journal b published similar 50 articles throughout the year. Due to the same subject article's aging speed, after several years, the benefit of journal a is twice that of journal b. Therefore, in identifying the old, we must consider their amount. That is, core journals that published the paper amount should be saved longer.

2) Citation amount

Periodical benefit evaluation is measured by the number of citations. Using journal citation data to determine the shelf life is a commonly used method.

3) Circulation data

The time distribution of science and technology personnel using literatures is regular. From the perspective of the statistics of periodical lending, 1 to 3 years is the most. Then, it gradually decreases. This rule and circulation data can be an important basis to guide the library work of weeding out.

4) Feedback data of the reader

The readers' utilization of and opinions on books have a reference value.

Based on the above statistical data of the measurement index, the warehouse area size of the unit, the need for basic collection, and the specific situation of future development, we can develop a reasonable solution for weeding out old literature. Thus, the best collection can be created.

10.2.5 Best Allocation for Literature Purchasing Funds

In a book intelligence unit, how to allocate funds is a practical problem that must be solved. Especially now under the condition of increasing literature, increasing price, and tightening of budgets, this problem is becoming more prominent. In general,

the principle of literature purchase fund allocation is as follows: spend less, ensure the literature purchase quantity, and satisfy the information demand of readers in the maximum limit. The best scheme of literature purchase fund, can be solved by the method of bibliometrics. The basic ideas include the following:

(1) Reasonable control of the proportion of new literature

There should be a reasonable proportion of new and original literature. That is, the collection update (including weeding out of the old) speed should be appropriate. For a certain subject or profession, the renewal speed of its collection of literature should adapt to the discipline development speed. As we know, the development speed of literature is closely related to the increase in the amount of literature. Therefore, literature measurement data can provide a quantitative basis for determining the proportion of new literature.

(2) Determine the proportion of interdisciplinary literature

The determination of the proportion is based on various literature measurement index statistics. For example, interdisciplinary literature circulation statistics, analysis of all kinds of readers, and reader comparison can illustrate the difference in the demand for interdisciplinary literature.

(3) Ensure quality and focus on selection and buying

Literature purchasing is based on the premise of need, but interdisciplinary literature and all kinds of literature have a difference in priorities and quality. With regard to the budget allocation, both should be considered, and the key points should be ensured. The determination of key literature is based on qualitative research. For example, one should focus on selecting and buying core journals.

In the study of the optimal allocation of literature funds, two concepts apply. One is the “law of diminishing marginal utility.” In the field of literature, it means that with the increase in literature, the diminishing marginal utility presents a certain regularity. According to this principle, we can reasonably determine several copies of books and periodicals and grasp the optimal quantity of literature.

The other concept is “the journal of commodity value.” This pertains to the ratio of the journal’s price to the number of usage times in a year (called the commodity value). A journal with a low commodity value should be ordered continuously. A journal with a high commodity value should cease to be ordered. The concept involves combined journal use time and price. At the same time, evaluating journals from the two aspects of usage and economic benefits provides a comprehensive measure for periodical ordering and funding.

The problem of how to build an optimal allocation mathematical model of literature purchase funds has been discussed since the 1960s, but no thorough and practical model has been established. This area requires further research.

10.2.6 Calculation Method of Book Shelf Placeholder

In book shelves, how large a position should each type of book hold? This is actually the growth forecast in the future for each type of book collection. Similar to the study on literature growth changes in bibliometrics, the basic relationship is the relationship between literature quantity and time. However, the unit of literature quantity is thickness (cm). Basic data involve the thickness of all kinds of books over the years. On this basis, the relation curve and equation of literature quantity growth over time can be obtained. Then, we can forecast the future trend according to it. The placeholder size of all kinds of books can then be determined.

At present, the reserved space for each type is mostly without basis given that many books and materials are classified. Most situations reserve empty shelves behind the library. Thus, the phenomenon of moving books from the bookshelf occurs often. It causes management confusion. To avoid large shelves, the method of reserving empty space behind each type of book is adopted. The length of the reserved space can be set in two ways.

(1) Simple calculation method

- 1) The proportional constant K is used.

$$k = \frac{\text{the length of the shelf space (cm)}}{\text{the length of the already stored books (cm)}}$$

- 2) The reserved space length of all kinds of books is equal to the product of K and the thickness of the books of all kinds.
- 3) Individual adjustments are made for the calculation results. Then, the placeholder of all kinds of books is determined.

(2) Regression analysis method

- 1) The thickness of all kinds of books changes over time; the former is the solid variable, and the latter is the independent variable. Statistical data on the two variables over the years are obtained, and a figure is drawn. This will show that thickness increases with the increase in time. The relationship is linear. Regression analysis can be performed.
- 2) According to the statistical data, the regression coefficients are obtained to determine the regression equation of certain types of books.
- 3) The required reserve position value of this type of books in the coming years is calculated from the regression line equation.

10.2.7 Evaluation of Literature Collection Work

Literature collection and management are important tasks in the library and information sector. The quality of these tasks directly affects the quality and efficiency of library and information service. Therefore, a scientific evaluation of the work of literature collection is necessary. Literature collection should be adequate, reasonable, economical, and can meet the information needs of readers or literature utilization. These are the general principles to evaluate literature collections. In particular, the content and method for evaluation are as follows:

(1) Literature readers' occupancy

The standards of literature collection decision optimization are an important research topic in literature management theory. Literature readers' occupancy can be used as a standard for quantitative evaluation.

Literature collections make the possible degree of systematic literature collections meet certain standards and make reader literature satisfaction meet certain requirements. The purpose of literature editing and collection is to use and occupy a certain audience. Both the possible degree of literature collection and readers' satisfaction with literature can be described with literature readers' occupancy. The changes in literature readers' occupancy demonstrates a Markov property. Using the changes in reader occupancy year by year for forecasting can improve periodical collections and develop the literature editorial optimization decision scheme.

The following departs from literature use frequencies and gives several concepts of literature occupancy.

1) Full occupancy

The unit of time (e.g., one year), the ratio of the number of certain types of literature readers, and the total number of readers can be used to evaluate and adjust the direction of literature collections in book intelligence agencies.

2) Part of the occupancy

The unit of time (e.g., one year), the ratio of the number of readers with a certain literature, and the number of readers with similar types and text type's literature can be used to select core journals.

3) Text types of occupancy

The unit of time (e.g., one year), the ratio of the number of readers of a certain literature, and the number of readers with similar types and different text types of literature can provide the basis for journal editors to select text types.

The evaluation standard of literature readers' occupancy is constrained by intelligence economics. The expert evaluation method can be applied, and the statistical data of readership require a survey to determine a standard value. If it is higher than this value, they can edit and collect primary literature. If it below this value, they can collect secondary literature by using a copy from other documentation centers.

(2) Literature utilization

Literature utilization data can be used to evaluate. How literature is used is an objective evaluation of the status of literature collections. Therefore, the statistical data of literature utilization, the lending rate of library or intelligence agencies, and the copy amount of literature can be used as basis for literature collection work evaluation.

(3) Collection structure

Applying document information analysis and evaluating collection measurement methods mainly use the established table of core journals in various disciplines and the best library catalog table with the library's holdings directory; comparison is performed one by one. Then, a comment on the integrity and quality of the collection can be provided. This also allows for the evaluation of a city. In a regional or national library for a particular subject or area of expertise, the circumstances of book collection, the scope and extent of repeated collections, and the characteristics and causes of books that are not collected are considered.

10.3 Informetrics and Information Retrieval

10.3.1 *Determination of the Integrity of Search Tools*

Quality retrieval tools are directly related to retrieval efficiency. Thus, in literature information retrieval, the retrieval tools used should have a basic valuation. The determination of the bibliographic integrity of search tools of abstracts and indexes can be carried out by using the laws of Bradford. The specific approach is as follows:

- (1) Count the actual number of abstracts and indexes to be tested and the number of the quoted periodical.
- (2) According to the obtained actual statistics from the subject of a periodical n and the $R(n)$ of a set of data, use $R(n) = K \log n$ to determine the total number ($N = K$). Then, according to the formula $R(N) = K \log N$, obtain the total paper journals in the discipline.
- (3) Compare the theoretical calculations to the actual value to determine the integrity.

For example, according to Bradford's law formula, an abstract magazine should extract at least 643 kinds of journal literature published in a particular year. It should extract 2784 relevant papers. However, in fact, the abstract magazine only extracts 374 journals and 2284 papers in a year. Therefore, the abstract magazine lacks 269 kinds of journals. The periodical shortage rate is 41.8%. It lacks 500 papers, so the shortage rate is 18%. In other words, the abstract magazine has 58.2 and 82% of journal paper integrity.

In the management of intelligence agencies in the library, testing papers' index integrity is an important task. The general practice is to extract some text containing high amounts of journals and statistically examine its capacity and the indexed rate of related papers to estimate the integrity of the index. This method is time consuming, laborious, and involves randomness; thus, it is difficult to use for accurate estimation. Using Brandt's law is more convenient. The specific approach involves arranging journals according to the grades of Bradford. If the ratio of the number of journals that should be verified and all the relevant papers journals is ϕ , we have

$$N \log \frac{n}{S} = fN \log \frac{N}{S} \quad (0 \leq n \leq N), \quad \text{then } \frac{n}{S} = \left(\frac{N}{S}\right)^f,$$

$$n = S \cdot \left(\frac{N}{S}\right)^f. \quad (10.1)$$

By using Formula (10.1), we can determine the index of a known size (known n), which is the coverage ratio of the total number of papers (seeking ϕ). Thus, we check whether the index has reached the original coverage requirements. If we require an index (or abstract) to report the amount of papers of a certain percentage, we quote the number of journals to determine the scope of the abstracting and indexing of excerpts. For example, if the requirements of a reported index is half of 2,000 journal papers ($S = 5$ is already known), then what is the number of journals to be extracted at the least? In this case, $N = 2000$, $S = 5$, and $\phi = 0.5$ may be incorporated into Formula (10.1).

$$n = 5 \times \left(\frac{2000}{5}\right)^{0.5} = 100 \text{ (Species)}$$

Quoting only the 100 kinds of journals with the highest paper rate can report 1/2 of 2,000 journals containing all relevant papers.

The index actual data can be compared with the theoretical value of the index according to Brandt's law to determine the index integrity. If the end of the actual value are very far from the theoretical values, the integrity of the index is poor. A foreign scholar studied the computer output on muscle fibers by referring to the MEDLABS catalog and computer science directory, as shown in Figs. 10.2 and 10.3.

In Fig. 10.2, the actual value of the number of journals N and the calculated values E is 313. $R(N) = 1402$, $R(E) = 1346$, and $S = 3.6$, $C = 18$. Figure 10.3 shows that $N = 750$, $E = 522$, $R(N) = 4050$, $R(E) = 3436$, $S = 3.4$, and $C = 13$. The visible, actual, and estimated values vary greatly, so the catalog is incomplete.

Fig. 10.2 Comparison of the actual and calculated values for the MEDLABS catalog

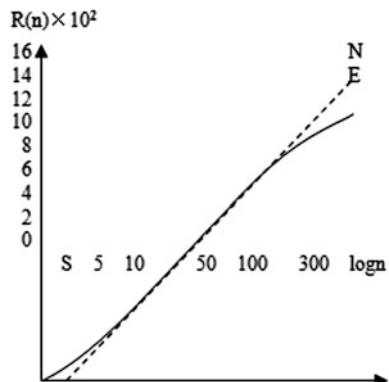
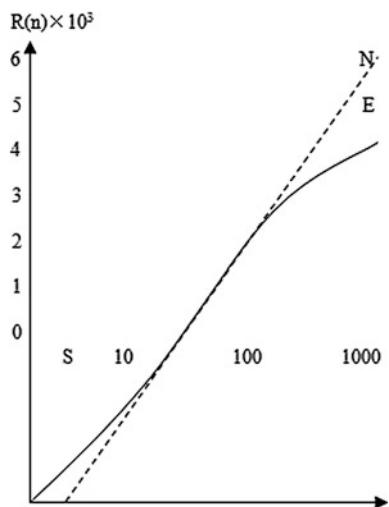


Fig. 10.3 Comparison of the actual and calculated values for the computer science catalog



10.3.2 Bradford's Law for Information Retrieval

Bradford's law for information retrieval mainly involves three aspects.

(1) Total number of full papers to be checked

In literature searching, this number is often estimated in advance. If n types of journals are fully searchable, the number of related papers would be determined. In this manner, we can conveniently evaluate retrieval effectiveness. Such problems can be calculated by applying Bradford's law formula.

According to Bradford's law, we can expect a full inspection rate corresponding to the relationship of n types of journals. When N is large enough, $K \approx N$. Therefore, the expected fully searchable number of n type of journal papers of in N journals is

$$R(n) = N \log n \quad (c \leq n \leq N).$$

Similarly, the total number of fully retrieved paper journals for all N is

$$R(N) = N \log N.$$

(2) Estimation of the search range

From the two previous formulas, we can conclude that the ratio of the number of fully searchable papers in n journals and the number of fully searchable papers in all N journals are

$$\frac{R(n)}{R(N)} = \frac{N \log n}{N \log N} = \frac{\log n}{\log N}.$$

In Formula 10.2, we set up $R(n)/R(N) = \phi$, so we have

$$n = N^f. \quad (10.2)$$

If it is extended to form Bradford's law that is revised by parameter S, we obtain

$$n = S \left(\frac{N}{S} \right)^f.$$

These two formulas have a practical application. In literature searching, they can be used to estimate the scope of the search. For example, a subject has about 400 journals. If you plan to retrieve 1/2 or 1/3 of the total relevant papers, then the number of journals you need to retrieve at the least is calculated as follows:

$N^{\frac{1}{2}} = 400^{\frac{1}{2}} = 20$ (species); $N^{\frac{1}{3}} = 400^{\frac{1}{3}} = 7$ (species) (rounded to the nearest integer). That is, as long as we find the top 20 journals with the highest paper rate, we can obtain half of the subject literature. If we need to obtain one third of the literature of a subject, we can meet the requirements as long as we search on the previous seven journals' range. Formulas (10.1) and (10.2) can also be used to guide the work of the journal subscription.

(3) Calculate the full retrieval rate

To retrieve data on a group of related journal articles on a topic, we can examine whether it complies with Brandt's law. If so, the total number of periodicals on the subject can be determined and retrieved together with the actual number of journals to compare. We can check the full rate determined with respect to the theoretical values. Below is an actual example for illustration.

Table 10.1 shows the retrieved data and calculated values on journal literature about vitamins.

Table 10.1 Retrieved data and calculated values on journal literature about vitamins

n	R(n)	
	Actual value	Calculated value
5	269	268
10	386	383
15	463	452
20	509	500
25	537	537 reference point
38	562	567
40	597	615
60	638	682
80	664	730
100	684	768
146	730	830
(167)	—	(853)

In the table, n is the serial number of journals sorted according to the decrease in paper citation rate and R(n) is the cumulative number of papers before n journals. The actual retrieved data listed in the table are in line with Brandt's law because $R(5) = 269$ (found number); $2R(5) = 538$ and $R(5^2) = R(25) = 537$ (found number).

Therefore, the choice of n = 25 as a reference point can determine the total number of full-retrieval journal vitamins (N). According to the laws of mathematical formulas by Brandt, $R(25) = K \log 25 = 537$ Then $K = 537 / \ln 25 = 537 / 3.22 = 167$ (the following are the natural logarithm) there $N = K = 167$ (species) Corresponding to R(N) value is $R(167) = K \log 167 = 167 \times 5.12 = 853$ pieces).

Therefore, for the data in Table 10.1, 167 species-related periodicals are expected to be found in a complete a full retrieval, but the actual number is 146 species. The number of fully retrieved papers is 853, but in fact, 730 are found. With respect to the theoretical value, we can determine the full retrieval rate of journals and papers as follows:

Journals full retrieval rate = $146/167 \times 100\% = 87.4\%$; paper full retrieval rate = $730/853 \times 100\% = 85.6\%$.

10.4 Informetrics and User Research

Readers are the object of library and information services and are also the user of documents and absorber of intelligence. The reader is a medium related to documents, and informetrics has a connection with reader research. Measurement research information can use "quantity" as a concept to promote its readers, service readers, and quantitative description of the law of various documents. Conducive

for in-depth reader research on particular quantitative research, several of the methods of informetrics can be used in reader studies. Depth information on metrology research greatly improves reader work to improve the efficiency of serving readers.

10.4.1 User Distribution in Line with the Law of Bradford

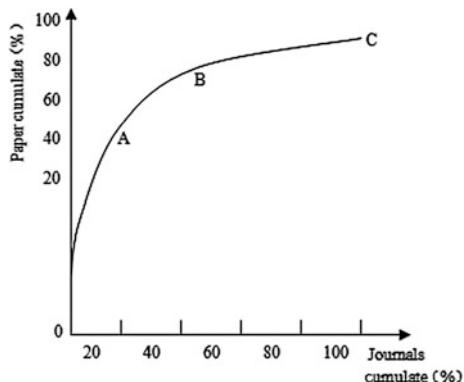
The Reader Services of Library and Information Center was established based on experience. The concept cannot be quantified. Bibliometric methods are used, although bibliometrics and reader's statistics can produce several regular conclusions. For example, reader flow and the number of times each reader uses journals in a certain period can be counted. If Bradford's tiering technology is utilized according to the order, the number of reader arrangements and packets can be reduced, so that each group's times of using the journals are approximately equal to the number of successive groups of readers in a geometric sequence. That is, the reader flow distribution is in line with the laws of Bradford. For example, journal statistics show that readers of the United States Allen (Allen) Memorial Medical Library in March 1908 constitute a minimum core of 13 people who use the journal 110 times, 18 people who use it 108 times, 24 people who use it 107 times, and so on. Evidently, the geometrical distribution of the eight successive groups of readers [1, (1.4):(1.4) 2: ...: (1.4) 7] is consistent with the distribution of Bradford. Also, for 86 readers, 76 kinds of journals were in circulation in 1, 2, 3, and 4 groups in March. According to the distribution of readers, we can determine the minimum core readers (i.e., major users). With the help of "medical index," we can identify the areas that this core audience is interested in to identify relevant publications. In addition, by establishing the distribution of journal articles, we can determine the core journals that users are most interested in. By studying readers' distribution and periodical distribution, the collection of the library or information center is maintained in an orderly state, thus providing readers the most suitable literature intelligence information.

10.4.2 Guide Users in Using Journals

Using Bradford's law to guide reader's literature has a practical significance. Specifically, by using the principle of Bradford's law to guide readers in selecting their literature focus saves time and improves the efficiency of obtaining information.

For the statistics and processing of papers for a given subject and period, if the y-axis represents the number of papers published in the cumulative percentage and the x-axis represents the cumulative percentage of journals published in a number of these papers, we can obtain the discipline of the papers' Bradford profile. Figure 10.4 is a subject of 155 kinds of magazines publishing 375 papers' Bradford profile.

Fig. 10.4 Papers' Bradford profile



As shown in the figure, there are 155 kinds of journal-published papers with a percentage of 100%, including 30 kinds of journal-published papers with 66%, and 5 kinds of journal-published papers with 33%. This means that if only five kinds of core journals are read, we can have 33% of this thesis subject. If 90 kinds of journals are read, then we can get 66% of all papers. With the “core journal” concept to guide the reader, the reader can greatly improve the efficiency of the use of the journal.

10.4.3 *Guide the Reader in Buying and Reading the Best Books*

From the perspective of bibliometrics, the quality of publications can be evaluated and identified. It can provide a scientific basis for the procurement of books and guiding the reader in reading. A library and information unit guides the reader to buy and read key books, primarily for readers to recommend and provide the “best books” and the corresponding “core publishing house.” Therefore, we must determine the optimal inventory of books in various disciplines. When bibliometrics is used to determine the core books, we arrange according to journals or citations from the original magazine of a subject, the statistics of journal articles that published or quoted the discipline, and the statistics on the number of books each document is cited in. Then, we sort according to the number of books cited. Several books ordered in the front are one of the best books in the subject area. For example, Baughman in “Social sciences and humanities index” (Social Sciences and Humanities Index Vol. 24 (1970–1971 year) extracted 446 journal articles in sociology and determined the total number of 11,130 cited literature papers behind it, including library materials (books, technical reports, degree in languages, government documents, etc.) with 6,840 articles. In the 6,840 books and materials, 759 are cited twice or more than twice, and 99 are cited five or more times. Baughman identified 99 kinds of books as the “best sociological books.” These are books and

materials related to sociology that the library must purchase and collect. At the same time, he also found that “Glencoe Free Press” (Free Press of Glencoe) is the main publisher of these best books. The best book published accounted for 26% of the total. With this method, we can determine the best books of various subjects. The best list of books that the library and information sector should procure and the guidelines for readers to buy and read these books should be focused on.

10.5 Concentration, Dispersion Laws, and Examples of Document Information Flow

The concentration and dispersion law of documentation and information flow is one of the basic characteristics of literature distribution. Bradford's law can be used for revelation and description. Below is an example to illustrate the basic processes and methods of such studies. In practice, the reader can learn by analogy.

In 2002, we conducted a scientific quantitative analysis for “Information Learned Journal” based on its issued papers and citations in the founded 20 years. An important aspect is the extensive literature on statistics according to Bradford's law using the tiering and image analysis technique to study the distribution of intelligence literature concentration and dispersion. Based on the image analysis results, we determined Chinese and foreign core journals in the intelligence profession.

10.5.1 Research Methods

“Information Learned Journal” is undoubtedly a core information science journal, and the distribution of cited documents' source journals reflects the distribution of the literature intelligence information source. Hence, we created statistics and analyzed the cited documents' source journals in the 20 years of “Information Learned Journal” that are cited and published by the China Science and Technology Intelligence Society, China Science and Technology Information Institute.

The statistics data are hierarchically arranged in Tables 10.2 and 10.3. Table A is the number of periodicals species, B is the citation amount, C is the cumulative number of A, D is the cumulative number of $A \times B$, and E is $\lg C$. With logn as the abscissa, R (n) for the vertical axis is drawn for intelligence professional literature with a Bradford distribution (Figs. 10.5 and 10.6).

Table 10.2 Distribution of Chinese source journals' ranking

A	B	C	D	E
1	926	1	926	0
1	485	2	1414	0.301
1	320	3	1736	0.477
1	218	4	1949	0.602
1	153	5	2102	0.699
1	132	6	2234	0.788
1	127	7	2361	0.845
1	103	8	2464	0.903
1	80	9	2544	0.954
1	69	10	2613	1
1	67	11	2680	1.041
1	47	12	2727	1.079
1	46	13	2773	1.114
1	36	14	2809	1.146
1	35	15	2844	1.176
1	34	16	2878	1.204
1	33	17	2911	1.231
1	31	18	2942	1.256
2	29	20	3007	1.301
1	28	21	3025	1.322
1	24	22	3052	1.342
1	22	23	3074	1.362
1	19	24	3093	1.38
4	18	28	3165	1.447
1	15	29	3180	1.462
3	13	32	3219	1.505
1	12	33	3231	1.519
6	11	40	3297	1.602
8	10	48	3377	1.681
4	9	52	3413	1.716
2	8	54	3429	1.732
6	7	60	3471	1.778
10	6	70	3531	1.845
13	5	83	3596	1.919
24	4	106	3692	2.025
31	3	137	3785	2.137
56	2	191	3897	2.281
214	1	412	4118	2.615

Table 10.3 Distribution of foreign source journals' ranking

A	B	C	D	E
1	183	1	183	0
1	155	2	338	0.303
1	96	3	434	0.477
1	78	4	512	0.602
1	54	5	566	0.699
1	41	6	607	0.788
1	36	7	643	0.845
1	29	8	672	0.903
1	23	9	695	0.954
1	18	10	713	1
3	15	13	758	1.139
1	14	14	772	1.146
2	12	16	796	1.204
3	11	19	829	1.279
3	10	22	859	1.342
4	8	26	891	1.415
7	7	33	940	1.519
4	6	37	964	1.568
10	5	47	1004	1.672
14	4	61	1070	1.785
16	3	77	1118	1.887
45	2	122	1208	2.086
163	1	285	1371	2.455

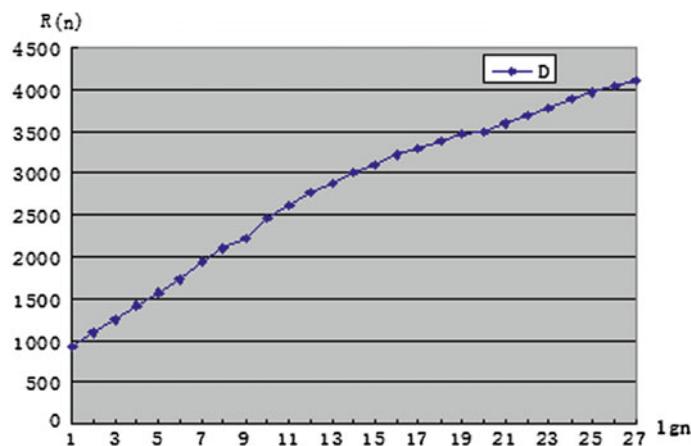
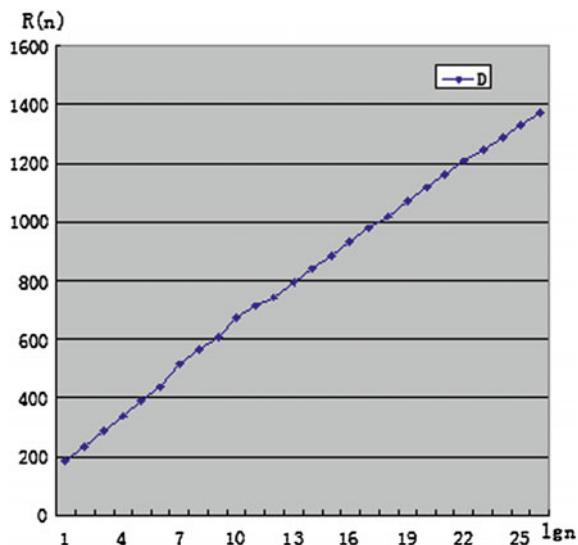
**Fig. 10.5** Distribution of Chinese source journals' ranking

Fig. 10.6 Distribution of foreign source journals' ranking



10.5.2 Research Results

By statistics, we found that 1982–2001 “Information Learned Journal” citations of Chinese sources have 412 kinds of journals. The top seven kinds of source journal citations’ amount accounted for 57.5% of the total amount, and the top 33 kinds of source journal citations’ amount accounted for 78.6%. We determined it according to the ranking in Table 10.4. Citations of foreign sources have 285 kinds of journals. The top 10 kinds of source journal citations’ amount accounted for 52.1% of the total amount, and the top 22 kinds of source journal citations’ amount accounted for 62.66%. We determined it according to the ranking in Table 10.5.

10.6 Examples of and Research on the Law of Literature Information Utilization

The law of using document information is conducive for the development of scientific information resources and improves the relevance and effectiveness of intelligence services. It is one of the common topics in the fields of library science and information science. By using the documentation information measurement method in the quantitative study of the law of literature information use can produce good results. The following examples illustrate such research methods and analysis contents, as well as several regular conclusions.

Table 10.4 Chinese important source journals

No.	Journal name	Citations	Cumulative citations
1	Information Learned Journal	926	926
2	Information Science	485	1411
3	Journal of Information Science	320	1731
4	Technology Information Work	218	1949
5	Library and Information Service	153	2102
6	Modern Library and Information Technology	132	2234
7	Intelligence Theory and Practice	127	2361
8	Information Professional Research	103	2464
9	Computer World	80	2544
10	Chinese Information Technology	69	2613
11	Chinese Library Journal	67	2680
12	Knowledge of Library and Information Science	47	2727
13	Intelligence Science and Technology	46	2773
14	Journal of Academic Libraries	36	2809
15	China Information Review	35	2844
16	Foreign Intelligence Science	34	2878
17	Computer and Library	33	2911
18	Computer Science	31	2942
19	Information and Documentation Services	29	2971
20	Library and Information	29	3000
21	Journal of Information	28	3028
22	Chinese Information	24	3052
23	Journal of Nanjing University	22	3074
24	Science and Science Technology Management	19	3093
25	University Library Science Communication	18	3111
26	Library Journal	18	3129
27	Theory and Practice of Systems Engineering	18	3147
28	Library	18	3165
29	China Science and Technology Journals	15	3180
30	Computer Research and Development	14	3193
31	World Book	13	3206
32	Library Tribune	13	3219
33	Information System Engineering	13	3231

Table 10.5 Foreign important source journals

No.	Journal name	Citations	Cumulative citations
1	J. Am. Soc. Information Sci.	183	183
2	Information Process & Manage	155	338
3	J. Doc.	96	434
4	J. Inform. Sci.	78	512
5	Hmu Cep (俄)	54	566
6	Commun. ACM	41	607
7	Online	36	643
8	情報管理 (日)	29	672
9	Annual Review of Information Science	23	695
10	International Classification	18	713
11	Am. Doc.	15	728
12	J. ACM	15	743
13	Database	15	758
14	Scientometrics	14	772
15	ACM SIGIR FORVM	12	784
16	IBM J. Res. & Doc.	12	796
17	Computer	11	807
18	Decision Support System	11	818
19	Alib Proc.	11	829
20	Nature	10	839
21	Library and Information Science Reseach	10	849
22	The Electronic Library	10	859

10.6.1 Research Methods of the Law of Literature Information Use

(1) Direct statistics and survey

This method involves statistically examining the readers' literature resource utilization in the library and information department or performs readers' group survey to arrive at a conclusion, which is used to guide library and information service. Statistics and surveys to classify readers generally involve the following:

- 1) In accordance with the range of disciplines that the readers need to classify;
- 2) In accordance with the readers' occupation to classify;
- 3) In accordance with the expression of reader information needs to classify;
- 4) In accordance with the readers usage information to classify;
- 5) According to the abilities and level of readers to classify;

- 6) According to the level of reader information to classify;
- 7) According to the services way of readers to classify.

For example, China Science and Technology Information Institute conducted a comprehensive survey on the literature utilization situation of scientific and technical personnel. The survey was performed according to the abilities and level of scientific and technical personnel. According to their title, they were divided into high, middle, and junior officers, and the current situation of scientific and technical personnel's utilization of journals was surveyed. From the scientific and technological personnel's common journal classification, their general periodical utilization, Chinese periodical utilization, utilization of foreign periodicals, utilization of journals with a comparison of various types, utilization of journals with different title, periodicals' cross utilization, and so on can be determined to analyze journals' utilization of scientific and technical personnel. The survey showed that journals play an important role in the literature sources of scientific and technical personnel utilization, in which the utilization of the public offering of Chinese journals is high. The use of foreign journals in English, Japanese, and Russian publications is highly common. The high, middle, and junior staff utilization rates of foreign periodicals are declining, whereas Chinese periodical utilization shows the opposite.

(2) Citation analysis methods

With the citation analysis method, we can grasp the law of literature utilization. It is one of the most important bibliometric methods.

For example, in 2002, we conducted a scientific quantitative analysis on "Information Learned Journal" based on its papers and citations issued throughout its 20 foundation years. Information Learned Journal is undoubtedly a core information science journal, so a scientific analysis of its citations reflects the literature utilization of the information science researchers to a certain extent.

- 1) Analysis of the number of citations. First, in its 20 years, Information Learned Journal's number of articles, number of theses with a reference to documents, number of reference articles, the average of citations, and statistics for such items were used as reference every year. Statistics results show that in information science thesis, the number and proportion of theses with reference documents presented a yearly growth trend. The total number of literature references also increased, indicating that the researchers in information science research increasingly focused on learning and absorbing existing achievements and respecting the work of others. It also indicates that Information Learned Journal has applied increased specification of draft and editing for better quality. Interestingly, the average number of each citation did not increase but showed a slight downward trend in 1993. This result shows information on past studies' introduction, presentation, learning, and succession to innovative research and development. The changes in the entire situation can be seen in Fig. 10.7.

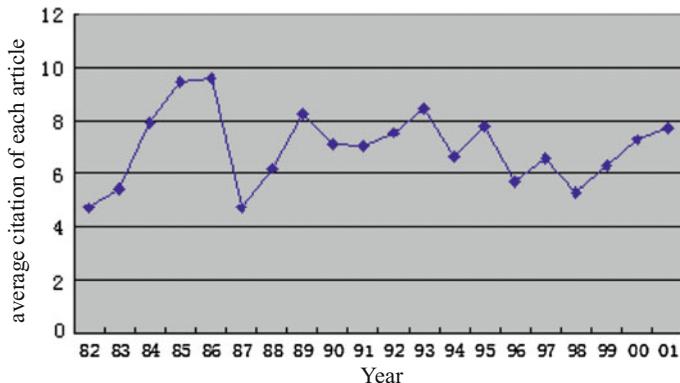


Fig. 10.7 Citations' yearly distribution

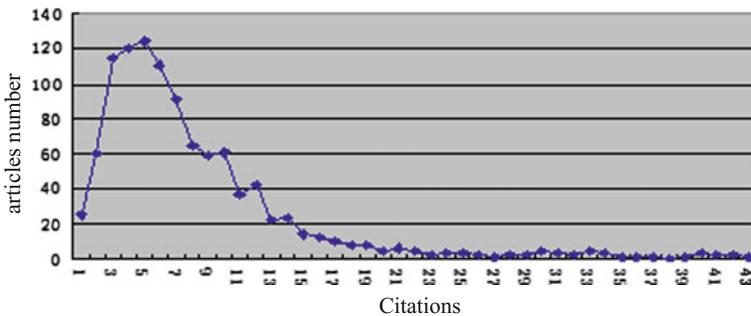


Fig. 10.8 Citations' distribution

Figure 10.8 shows that in the theses with references to documents of Information Learned Journal, 124 articles contain 5 citations, 120 articles contain 4 citations, 115 articles contain 3 citations, 111 articles contain 6 citations, and 92 articles contain 7 citations. All these articles account for 40.87% of the reference articles. That is, based on the midpoint 5, both sides expand at an approximately normal distribution compared with the average of per paper citations with 9.01 as the midpoint of expansion. The amount of some disciplinary citations' distribution law is normal or partially normal based on the average number of citations in articles as the midpoint to expand both sides. There are complex reasons for this difference, which may be caused by the largely subjective characteristics and journal differences.

2) Statistical analysis of citation language

Statistics show that the document citation languages are mainly English, Japanese, Russian, French, and German. Chinese literature accounted for 69.28% of the total citations, and English (the main foreign language) accounted for 28.13%. Together,

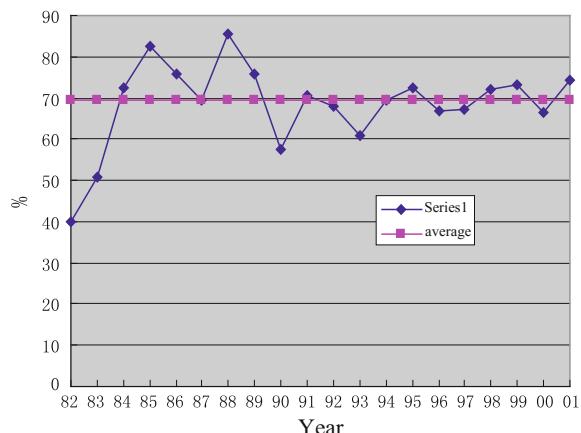
they account for 91.57% of foreign language citations. Regardless of the number of references in Chinese, foreign literature shows an upward trend. For foreign language, the proportion of English gradually increases, and that of other languages gradually decreases (concentrated on English). The proportion of Chinese citation is more stable. In particular, it reached a stable state after 1991, indicating that after 1991, our intelligence research entered a stable stage of development; either the country's literature and other intelligence services were able to provide more to meet the research needs or in the early 1980s, or we were mainly dependent on foreign materials (for example, in 1982, 60% of citations are in foreign languages). From another aspect, foreign literature's use of Information Learned Journal author declines. A certain gap exists when compared with other disciplines' author who take advantage of foreign language materials. The question is worth considering.

3) Statistics and analysis of the type and amount of citation

We created statistics for the citation type of Information Learned Journal. The statistics show that the major Chinese citation types of information science research are journals, books, conference papers, newspapers, dissertations, technical reports, and network standards. Among them, Chinese journal articles account for 62.10%, and books account for 21.40%. Both types of citations altogether account for 83.50%, indicating that journals and books are the major document types that information science researchers take advantage of and also their main sources of information.

From these statistics, we can see that the foreign language citation sources for information science research are journals, books, networks, technical reports, conference papers, dissertations, newspapers, and standards. Among them, journal articles account for 46.62%, and books account for 26.83%. Both types of citations altogether account for 73.45%, indicating that journals and books are the main document types that information science researchers take advantage of and their main sources of information.

Fig. 10.9 Proportion of citations in Chinese

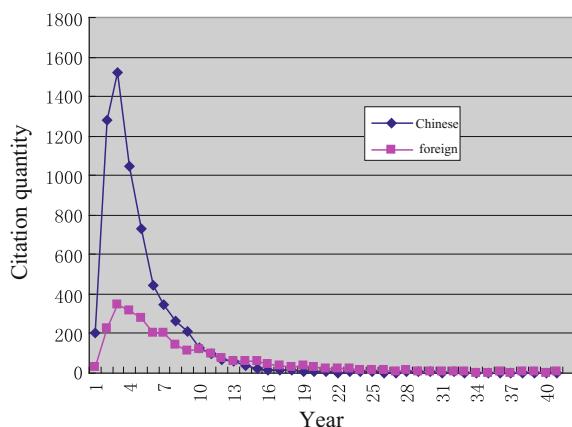


Between the two, for foreign citation sources, the two types of network literature and science and technology reports are more important than their Chinese counterparts. This indicates that science and technology reports in our country are studied less, and network information resource construction is insufficient. Network documents, whether in Chinese or foreign languages, are new and an increasingly important source of information. (Fig. 10.9)

4) Statistics and analysis of citation years

Figure 10.10 shows the distribution of the total number of Chinese and foreign citations each year. The largest average citation life is 2 years. This finding is similar to the conclusion of D. Price, who studied English citations with a maximum age of 2 years. It also shows that the apparent peak of Chinese citation distribution is higher than that of foreign citation distribution and decreases faster. The Chinese citation amount of more than 10 years is not as good as the amount of foreign language citations. This result shows that Chinese literature spread and are used fast, but it cannot explain the rapid aging of Chinese literature and the slow aging of foreign literature. On the one hand, it is determined by the particularity of the historical stages of development of our country's information science research. For our country's information science, its short history, rapid development, and scope of statistics basically include the entire period of the restoration and development of information sciences. The time range of each year of citation in Chinese sources is small, resulting in the relative concentration of aggregate data. On the other hand, foreign literature are carefully screened, more important, more famous, more classic, and the cited authors are relatively concentrated in a few well-known experts in intelligence. In addition, statistical data are only from domestic Chinese journals. This is also one of the reasons. Comparison of the aging of Chinese and foreign literature requires certain preconditions.

Fig. 10.10 Chinese-foreign citation age distribution



5) Statistics and analysis of citation discipline

Citation discipline is affected by the distribution of the subject classification system. We should adopt a crude, clear, and easy-to-distinguish principle according to the actual citation situation and classify citation literature and order by size. From the statistics, we see that citation discipline involves 29. Information science self-reference accounts for half, which is 50.78%. In related disciplines, citation literature of computers, library science, economics, linguistics, science, and philology account for 83.12% of related disciplines' citation literature, which are mainly related disciplines that are closely related to information science. In particular, the citation rate of computer science and library science is 12.8 and 11.35%, respectively. The other subjects involved systems theory, law, mathematics, philosophy, sociology, statistics, management, education, publishing, industrial technology, information theory, biology, archival science, psychology, cybernetics, communication, chemistry, futurology, physics, medicine, geography, military, and so on. Information sciences is an open and comprehensive discipline.

In related disciplines, social and natural sciences account for almost half, but in citation literature, the citation amount of social science disciplines accounts for 66.49%, which indicates that information sciences and social sciences are closely linked, with a strong social science disciplinary nature. At the same time, it can be that the Information Learned Journal author has a main background in social sciences and is familiar with social science literature, so his citations are more skilled.

6) Statistics and analysis of citation source

Through statistics and analysis of source journals, we can grasp the scope of utilization literature and the core journals of information science research. The findings have been presented in the previous section. The study results reflect that information science researchers use literature regularly. There is a certain significance, and research methods exhibit certain exemplary performance.

10.6.2 Study of the Law of Scientific Researchers Using Literature Information

Scientific researchers are one of the main service objects of library and information service. Collecting, sorting, and providing documents is an important task for a scientific researcher. Therefore, it is helpful to improve the service quality of library and information service and to carry out knowledge service. Research shows that the utilization law of scientific researchers can be analyzed from the following aspects.

(1) Purpose

Researchers use literature with a clear purpose because they generally have their own research and task. The task is clear, so the purpose of literature demand is very

strong. Research problems often involve finding information or seeking answers from the literature or inspiration. This is one of the characteristics of researchers.

(2) Comprehensiveness

For any scientific work, we need to possess sufficient information. To study scientific and technical issues, we must collect and possess sufficient materials to provide a reliable, sufficient basis for analysis to reflect the nature of the problem. To this end, researchers are generally aware of the following:

- 1) First, consider what documentation needs to be collected and how to collect it to achieve comprehensive requirements.
- 2) Only objective facts are considered.
- 3) Documentation must ensure maximum reliability and accuracy.
- 4) A calm, impartial attitude must be exhibited toward all scientific literature analyzed. There should be no preconceptions, stereotypes, and prejudice.
- 5) In the analysis, the shortage of documents or materials must be actively addressed.

(3) Systematic

Researchers should not only be purposeful, but should also be systematic. As for the contents of literature, we need professional related literature as well as the literature of other disciplines. The text, type, time, school, etc. are necessary to Chinese and foreign literature. Currently, the trend is to read Western literature, both books and journals, as well as patent documents, conference papers, and other special literature. The latest and previous literature are both necessary, while various factions, different points of view, and documentation require the system to search Eris. In this manner, they can comprehensively and systematically analyze problems and conduct research work based on a relatively broad basis.

(4) Timing

When researchers are engaged in a research, the most urgent need is to master the dynamic national research of this topic as soon as possible. Given the very rapid development of science and technology, the cycle from research to invention application becomes short, while the aging of scientific literature constantly accelerates. According to foreign statistics, the average life of scientific literature used by researchers is only five years or so. Some argue that a delay of a year and a half to two years in the publication and use of scientific and technical literature means that its intelligence value decreases by approximately 30%. Therefore, researchers' requirements on timing for literature are high, and they focus on using the latest scientific literature.

(5) Stage

Another feature of researchers' utilization of literature is the stage. In general, a research work has to go through topic selection, research, experiments, summary,

and appraisal stage. The specific requirements of each stage differ, so their demand and utilization of literature also have obvious stages.

Based on a literature survey to select research projects, after determining the subject, an in-depth literature review is provided. Its purposes are as follows:

- 1) Understand what their predecessors have achieved on this subject to avoid duplication or repeating the mistakes of their predecessors;
- 2) Understand the reference information, data, techniques, and methods of research for the subject;
- 3) From similar previous human studies, learn the research methods to establish hypotheses and analyze the ways of thinking;
- 4) Learn about the latest trends and progress in the field of science to expand their horizons and to consider the issue fully. For example, for Chinese scientists studying synthetic insulin, inspection of a large number of documents is needed to solve the problem of research trends.

In summary, in the appraisal stage, researchers also need to quote the results of others and reference information from their peers to write a high-quality paper or report.

(6) Skill

Researchers using and reading literature generally take advantage of search tools for scientific retrieval and strive to achieve precision, recall, and browsing of related literature, abstracts, book reviews, research on the dynamics of the type of information. They aim to find out what literature is available and create their own index data. Next, they read the literature they found. When reading, they often use three methods, namely, reconnaissance, browse, and intensive.

In summary, the basic features of researchers using and reading literature are as follows: a very clear purpose, comprehensive and systematic requirements, time, new content, and certain stages and skills. By researching and clarifying these laws, library and information services departments will be able to effectively carry out work to improve service quality.

Chapter 11

Application of Informetrics in Science and Technology Management and Forecasting

Science and technology management is a new research subject regarding the basic principles and methods of modern science and technology management. It is a comprehensive discipline that was developed based on modern management science, science of science, science and technology history, talent studies, system theory, natural dialectics, and so on.

A very close relationship exists between informetrics and science of science and technology management. From the management point of view, there is no evaluation for management, so there is no scientific evaluation of scientific management. Therefore, it is a new and effective means to study and evaluate the science of science, talent studies, and science and technology history from the perspective of informetrics. This chapter mainly discusses the application of informetrics in the aspects of scientific research, talent evaluation, and science and technology forecasting.

11.1 Informetrics and Science of Science

11.1.1 Basic Principles

Science is a discipline that studies the law of scientific development and the structure of an organization. The development law of science, to a large extent, is reflected by the growth and decline of talent, money and results, growth rate, subject structure, and its proportion. These aspects of the development and changes in scientific literature should be the focus of a variety of changes. This is because scientific literature is an objective record of scientific and technological knowledge and achievements, and it is a manifestation of science. Any scientific research and technological creation must be based on necessary scientific documents for its final stages. At the same time, science and technology involve the use of scientific

literature to inherit and develop. Therefore, the quantity and quality of scientific literature are measures of the level of science and technology. According to the content and quantity of scientific literature, the history and present situation of science and technology can be summarized, analyzed, and evaluated. Specifically, several aspects are involved.

- (1) The amount of scientific literature can reflect the extent and stage of the development of science and technology.
- (2) The national or linguistic distribution of the amount of scientific literature reflects the strength and technical advantages of scientific and technological research in different countries.
- (3) The change in the amount of scientific literature reflects the speed of scientific and technological development, and doubling of the amount of documents can be used as a measure of the development of science and technology.
- (4) The change in the amount of literature reflects the turning point of the development of science and technology: from growth to decline or from the development process to the recession process.
- (5) The amount of literature published by scientific research institutions can reflect the technical capabilities and research results of the institutions.

At the same time, the informatics reveals and describes the law of scientific literature, but also fully reflects the scientific development of a certain appearance, characteristics and laws. Therefore, using the method of information measurement, we can study the structure and development of science, scientific and technological achievements, talents, regional and institutional evaluation, scientific and Technological Forecasting, etc. In short, information measurement is not only applied to the field of science, but also provides a new way for the scientific research.

11.1.2 Research Contents

The application of information measurement in science research is mainly reflected in the following four aspects.

- (1) Research on the characteristics of scientific development

The development of modern science has many outstanding characteristics, which fully reflect the amount of scientific literature and its changes. Therefore, it can be used to study the characteristics of scientific development from the perspective of measurement. Research on information measurement shows that the development of science has the following characteristics.

1) Scientific development speed

The study of information measurement shows that the amount of modern scientific literature refers to the growth of the index. American Chemical Abstracts, from its founding in 1907, published the first one million pieces of abstracts for 32 years and the second million in 18 years. One million was published for 3–4 years every five months, and the last one million was published in only two years. The rapid growth of scientific literature is the concrete embodiment of the speed of scientific development.

2) Scientific development is inherited

The inheritance of scientific development from the content of literature and the use of the content is fully reflected. From the science citation, the content is coherent. Scientific and technological personnel not only use the latest literature, but also review the past. That is, the development of science has the obvious characteristic of inheritance, and the integration and development of science and technology require the use of scientific literature.

3) Scientific development is a stage

A close relationship exists between the number of documents, the number of subject words, and the development of the subject. In the initial stage of science, the amount of literature was small; with the gradual development of the subject, the amount of related literature increased rapidly. When a subject develops to a more mature stage, the number of documents exhibits stability then a slight decline until the subject shows a comprehensive trend. Then, the number of documents increases, and a new cycle begins. The relationship between the number of subject words and the development stage of the subject is similar to that of the subject. Yu A. Shiliejjeer used the “thesaurus” to define the information volume. According to his theory, information can change the subject structure, such that the number of subject words that reflect the new subject of the new branch of the subject also decreases or increases. This shows that the number of subject words and subject development has a certain link.

4) Subject to cross permeability

In scientific research, the cross penetration of various disciplines is increasingly serious, and the edge of the discipline is constantly emerging, which makes it difficult to complete the knowledge of a subject. In fact, a single discipline is very rare. According to a survey on 1129 kinds of foreign periodicals, those involving more than four subjects accounted for more than 60%. Subject literature on the content of cross penetration is a reflection of the characteristics of cross penetration between subjects.

5) Transfer of scientific development focus

An interdisciplinary literature's quantity, abstract number, and word frequency statistics often reflect the proportion of each subject and the pace of development

(the transfer of scientific development focus). For example, in recent years, the United States' chemical abstracts accounted for 28–41% of the number of biological chemical abstracts from 1967 to 1979, which is far higher than the proportion of organic chemistry and other chemical branch. This is mainly due to the rise in biochemistry in recent years as a result of an important subject.

6) Collective nature of scientific labor

The increase in the number of literature authors reflects the strengthening of scientific labor collective tendency. Before and after the industrial revolution, a person may be able to conduct research on one subject, which is published in the name of the individual. Many research projects must rely on collective strength for completion, and the results are the crystallization of collective wisdom. Therefore, the performance increases. The statistical analysis of D. Price of American Chemical Abstracts (CA) showed that in 1910, an author of a chemical literature accounted for more than 80%; in 1963, an author of a chemical literature accounted for only 32%, two authors accounted for 43%, three authors accounted for 15.5%, and more than three authors accounted for 9.5%. Similar trends were observed in mathematical literature. The co-authors of a scientific literature embody the characteristics of scientific labor.

(2) Research on scientific structure

The structure of science as a system is hierarchical and dynamic. Bernard, a famous scientist, pointed out that the scientific structure of a model is to be reestablished. From the angle of information measurement, bibliography (table of contents) analysis, citation analysis, and word frequency method can elucidate the structure of scientific development and the rules of scientific development, thus providing a basis for the scientific research management and scientific decision making.

1) Using the method of bibliographic analysis to study the scientific structure

Bibliography (table of contents) analysis is a statistical data analysis method based on information. Broadly speaking, bibliographic information lists all the descriptive projects and symbols but mainly refers to the classification number, title (title), author, publication, subject headings, languages, audit, etc. Given that the bibliographic information on the content of the literature of that role and the retrieval literature indicates a marked effect and external bibliography includes the intelligence phenomenon, the social phenomenon that exists between natural link functions shows that interlocking network links actually exist between bibliographic information. Therefore, a certain relationship exists between the number and content of bibliographic information, which is the basis of the analysis of the information and scientific structure of research. This can be understood from two aspects of the study.

The structure of a subject can be revealed. A certain connection exists among the various fields of scientific knowledge, the various subjects, the carrier of knowledge, and the form of writing, such as books, periodicals, subject words, key words, and classification number. The semantic nature of human knowledge makes allows bibliographic information to reflect knowledge itself to some extent because bibliographic information is the knowledge of several concepts of language and symbol expression. The structure of bibliographic information (view, genus, coordinate, group, etc.) reflects the relationship among these concepts but is a more concise reflection of the structure of knowledge (or subject). For example, each subject in the field of journal title and the key words' number ratio can reflect the degree of subject knowledge increment (amount of information) or the ratio of people's demand for each subject knowledge. At the same time, the closeness degree between theme words to express all kinds of knowledge differs, so subjects with proprietary subject words often appear in the same literature. The proprietary theme word with different disciplines is not easy to meet in the document. As the saying goes, "not a family, not into the door." Similarly, in a large comprehensive abstract, the same subject journals, authors, publishers, and others are clustered together in text classification and indexing. Therefore, according to the classification number, the article can reflect the specific gravity and the development speed of the subjects or branches.

The change in the number of books often reflects the dynamic characteristics of the development of the discipline. Given that the "human society has some kind of mechanism to evaluate scientific information," the result of this kind of evaluation makes the development of the subject urgent. This phenomenon is reflected in the list of new words, the emergence of new classification numbers, the number of popular subjects, the number of key words, the rapid increase in the number of words, etc. At the same time, the number of documents, the number of key words, and the number of subject words often reflect the rise and fall of scientific development. The amount of literature and subject development stage also have a close relationship. In addition, the usual sentence patterns, commonly used words, the author's research topics, the study of the author of the subject, and the habit of hobbies are related. As for commonly used words and collocation, different authors have different subjects, different experiences, different habits, and different frequency of use, resulting in the unique "word scientific guide." From these statistics of the amount of bibliographic and content information and through an overall analysis, we can infer that scientific research has several meaningful conclusions.

2) Using citation analysis to study the scientific structure

First, a relationship exists between scientific citation and the citation. A scientific paper is not isolated from the large environment of the scientific system, but is a part of the network structure of scientific literature. Therefore, a correlation exists between the citation and the cited. An article can be found through the mutual citation relationship between the literature and the subject so as to find a scientific paper of the subject connection, scientific literature structure, and scientific structure.

Second, the frequency of a paper is an objective measure of its academic level and value. Scientific papers are cited to illustrate the information content at the end of the scientific communication process, and the use amount is a measure of the academic level and value of the paper.

Therefore, the structure of the science can be analyzed from the number and structure of citations. Citation analysis is used to study the scientific structure and history of science, improve scientific management, formulate scientific policies, and avoid the uncertainty caused by subjective judgment and the traditional evaluation method, which reflects objective reality. At present, citation analysis methods, such as citation and clustering analyses, have become an effective method to study the scientific structure.

The application of citation analysis in scientific structure research mainly includes static research, the study of dynamic research, and research on the structure of research.

In static research on the scientific structure, the formation of the cluster in the network can be used to study the static structure of microcosmic science. When professional literature in the same is cited, through the same cited clustering, we can obtain a number of clustering. If a node (circle) is represented by a cluster, the number of nodes is the cluster number, and the number of nodes is two. Thus, we can derive the relation network diagram of clustering. This kind of network graph can be used to analyze the relation between subject and specialty.

Dynamic comparative study of the scientific structure is based on static research. Through a comparison of a few years, we can find out the trend of its change. The methods used are as follows:

- A. Contour map method: If used on behalf of the literature, aside from the point of the text for the authors, the distance between the points is determined in at the same citation intensity. The co-citation strength is great, the distance between points is close, and the contours of the height are proportional to each literature citation. A contour map can be used to analyze the dynamic structure of a research progress of a few years.
 - B. Block diagram method. The principle involves the combination of the network graph and the dynamic analysis of the network diagram. From a cluster of several small clusters, we remove the co-citation strength as a certain clustering threshold, plot the diagram to block on behalf of small clusters, and create a digital marked box contained within the document number. The box is similar to the link, which is the same as the two clusters with the cited intensity. The block diagram of such a form can reflect the degree of the clustering structure of the discipline (or specialty) and the degree of the discipline.
- 3) Microstructure study of the super structure.

This involves the use of the so-called two-dimensional space map to describe cross disciplinary macro clustering. Clustering analysis reveals the macro clustering of several disciplines. This situation is common in the formation of the sub structure of a large profession, and this method of literature is often involved in various

disciplines. In the point representation, the distance between points is inversely proportional to the intensity of the points. This can be obtained by the method of two-dimensional space map of a subject.

Analysis of the development of the discipline structure by using word frequency

Word frequency analysis is actually a special case study analysis. In recent years, scholars studied and evaluated discipline and field dynamic development trends and analyzed characteristics through a subject or a field of keywords in the title, abstract, keywords, and text appearance frequency.

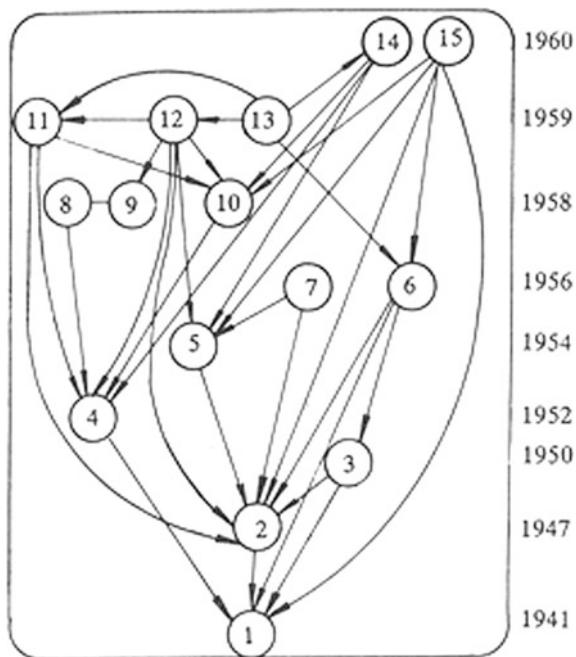
Research on the history of science and technology.

For the development of science and technology, information measurement has become an important research technique because aside from bibliographic data, citation data can also provide historical information. At the same time, each paper is a record of specific events in the process of scientific development, and each historical event occurs at different time points. The citation of scientific literature can reflect the origin and development of these events and reveal the development process of several scientific ideas or experimental techniques. Citation by distribution using historical figures and citation networks can be used to prove a subject background and provide an overview of the development, breakthrough achievements, and future development direction.

The time sequence of scientific progress is an important content and way of scientific and technological history research. The historical figure vividly presents the overview of scientific progress. It is based on the different times of the citation relationship. One of the basic forms of the citation relationship among papers is the time series, which is the time distribution of the citations. If each paper is an important problem in the development of science and cited many times as the key problem, then in accordance with the time series, the literature citation relationship describes the origin and development of the key issues. Thus, we can draw a history map if a small circle represents these key issues and the time sequence, and a number of lines represent the key issues of citation association with an arrow to indicate the relationship. To test the accuracy of this historical map, two historical maps have been compiled from the book of the genetic code. The first book is to find out the relationship between important historical papers and general papers. The second picture is the name and subject of the research provided by the books "Chemical Abstracts," "Medical Digest," etc., and then the citation index and citation relationship are established. Comparison of the two charts shows that the relationship between the cited network and the "genetic code" is 65%. When the problem in the citation network is weighed against the amount and manner of the citations, the highest number of times cited is in agreement with the judgment in the book. This shows that this kind of map can depict scientific progress in the specific process of a certain period and can also realize the development of scientific history. Therefore, the citation network map has created a new form of research on the history of science and technology.

In 1960, according to the above idea, Alan used papers and their mutual citation relationship according to the time sequence of the nucleic acid staining method in the 1940–1960s and clearly revealed the relative importance and interaction of

Fig. 11.1 Fifteen citation networks for nucleic acids



waste paper. Figures 11.1 and 11.2 (Michaelis 1947) is clearly the most important in this field because it has been cited repeatedly at different times.

In addition, some of the problems of the citation network map can be corrected through the omissions in the history of science. For example, in the history of science, people thought that the important papers published in 1865 by Mendel Michaelis G. were not paid attention to until the academic community was discovered in 1900. However, according to the genetic map of the citation network (Fig. 11.2), it is clear that the paper was cited by at least four scientists before 1900. Mendel's achievement was even in the ninth edition of the Encyclopaedia Britannica, and the article title was cited as "hybrid." The scientific community was aware and paid attention to it. If the above citation network diagram can be found, the process of scientific development of an abnormal phenomenon can be established. Darwin, in his 1876 article, cited the article (1869) but did not cite Mendel's article. Hoffman's article cited Mendel's article five times. This is clearly an issue of interest among scientific historians. The above examples illustrate the validity and popularization of the citation analysis in the research of scientific development.

(4) Research on science and technology policy

The direct application of information measurement in science and technology policy research is extensive. For example, the distribution of scientists, the determination of scientific productivity, etc. can provide a basis for the development of science and technology policy.

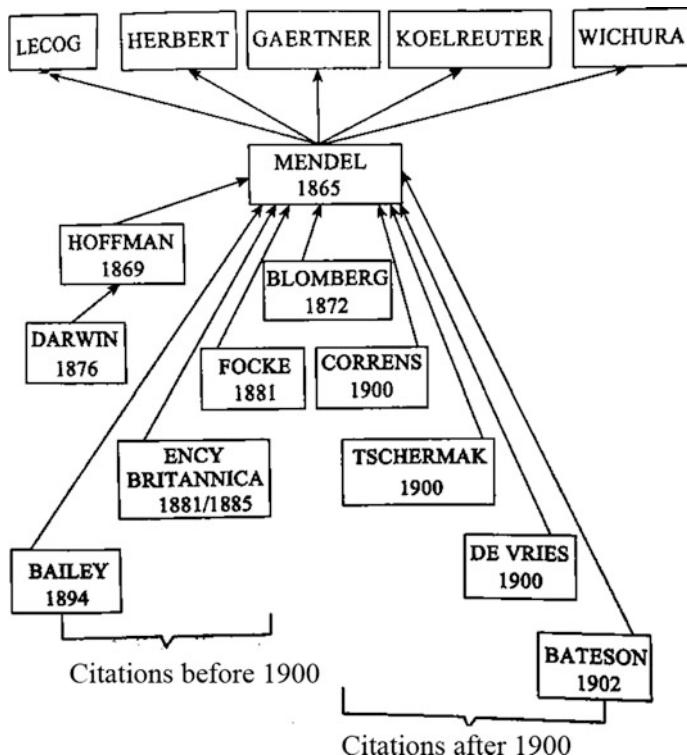


Fig. 11.2 Citation network of the paper in 1865 by Mendel

D. Price in his statistical analysis of the number of scientists all over the world found that 90% of the 1967 world scientists lived in 14 most developed countries in the world. The scientists in the 40 most developed countries account for 99% of the total number of scientists around the world. Therefore, his conclusion is that the distribution of scientists is closely related to the development of a country.

Recently, Kovach used 1967–1976 scientific thesis author directory, which involved statistics on the published number of scientists for the determination of national scientific productivity, and analyzed the seismicity trend of scientists all over the world. The results showed that the growth rate of scientific productivity in Spain was the fastest, and the average growth rate of South Africa, New Zealand, Iran, Nigeria, and Brazil significantly increased. He also found that in most Western European countries, the number of people engaged in research and development was increasing, but their scientific productivity was significantly lower than the average growth rate. He believed that this situation was mainly related to the serious political and social unrest in these countries.

Inhaber divided world cities into five categories according to their population and provided statistics on the number of scientists per 1000 people in various cities.

The results showed that scientists are mostly concentrated in large cities. Therefore, he proposed the use of the countries where scientists are geographically positioned to calculate reasonable policy issues.

11.2 Informetrics and Talent Evaluation

Science and technology talents are the wealth of a country. The quantity and quality of scientific and technological talents are important symbols of scientific and technological capabilities. How to identify and evaluate talents is an important issue to be studied.

Given that people of integrity are revealed through a variety of ways, we evaluated and identified talents to provide a variety of ways and means. The general method proceeds as follows: evaluation and identification of scientific and technological achievements, periodic assessment, and practice for all kinds of competitions and other activities to select and identify talent. Talents can also be found through various academic conferences and academic exchange activities of outstanding papers. Various academic periodical editorial departments can be assessed through the review of manuscripts to identify the ability to create and put forward new theories, ideas, and talents. Although these methods are feasible, their common flaw is the lack of quantitative analysis. At present, the method of information measurement is widely used in foreign countries to evaluate scientific talents. Many studies have indicated that from the angle of information measurement, the evaluation of talents is more objective and accurate. In particular, the United States' Science Citation Index and Social Science Citation Index for information measurement in the application of talent evaluation have opened up a very broad prospect.

11.2.1 *Talent Evaluation Theory of Informetrics*

In scientific activities, scientific literature and scientific talent have a certain inner link. The objective is to evaluate the basis of the information from the perspective of measurement. The principle of using the method of information measurement is from the following two aspects.

- (1) Achievements of scientific and technological personnel and number of academic achievements in the number of papers published

The famous information scientist A. Mikhailovich from the former Soviet Union once pointed out that “for every scientist, the number of his articles reveals his scientific labor efficiency sufficiently” (only a relative measure). This is because the achievements of scientific and technological personnel must be expressed in the form of literature to promote their application and obtain social recognition. In the

application technology, the ratio of the number and the number of patent literature is almost 1 to 1. Literature and scientific discovery are not one, but the proportion of the two is small. From a macro perspective, the number of the entire literature and the development of social productive forces are proportional. From the micro perspective, the number of published literature is a measure of authors' scientific and technological achievements and contributions. In general, the more scientific and technological personnel, the greater their achievements in literature, especially patent literature, and the more published they are in important journals. Therefore, the number of published literature can be used as a basis for the evaluation of scientific talents.

- (2) Achievements of scientific and technological personnel and academic attainments are related to the number of cited papers published in literature

The papers published by scientific and technical personnel are not always cited as references. According to foreign statistics, the history of scientific literature is cited more than 10 times for 50%, 2 times for 9%, 1 time for 10%, and about 6% of the literature has never been cited. In addition, science and technology research shows that for a scientist in a normal situation of about four papers published in a year, these papers are nearly 1/4 after publication and no one cited them. Generally, if a paper is cited four or more times a year, it can be classified as "classic literature." Although citation motives and behaviors vary, regardless of how, they always vary because of the viewpoints and materials. This has certain value to assist authors in solving several problems. Therefore, the number of citation times, to a certain extent, reflects the quality and value of a literature and reflects the influence and position of the author in the academic circle and his contribution to the society. However, for several authors who publish many literature, the number of citations may also be large. Therefore, author literature citation rate reflects the quality and value of literature, which can measure the academic level of the author. Thus, the cited rate of scientific and technical personnel is one of the important indicators to evaluate scientific talents.

11.2.2 *Talent Evaluation Methods*

According to the above principle, two main methods can be adopted to evaluate talents through informetrics.

- (1) Bibliographic analysis

The certain period of time all scientific and technical personnel are in publication, the number of articles, and the tables of contents are counted. Then, the various contents are compared, and having many published articles is generally perceived as having many achievements. This method is simple and direct but is insufficient; it is usually used with other methods.

(2) Citation analysis

Using the citation analysis method for the evaluation of scientific research personnel and their achievements is generally implemented with four indexes. The total is the total number of citations, and each paper's citation is an introduction to the number. This is one of the most common methods and simple and easy. The specific approach is as follows: use of direct statistics or Science Citation Index database statistics in a certain period to determine who is cited and who was elected. This method is objective and reliable and has been widely applied in the selection of personnel.

Since the 1980s in Europe and America, the use of the citation analysis method in the evaluation of scientists and their achievements has become the traditional peer review and an effective auxiliary means; it could be replace peer review because compared with peer review, the citation analysis method is more quantitative, objective, and involves less human factors. Cost and time are also saved. In the United States, the citation analysis method and peer review were used in a comparative study on 467 students in a university to assess who should be awarded a doctoral degree in biochemistry. It is a conventional method, and 152 were used as peer review experts. The other is to use the method of citation analysis. Striking similarities have been found between the two, but the citation analysis method is more convenient, objective, accurate, and easy to use.

11.2.3 Selection of Outstanding Scientists

The use of large databases, such as the Science Citation Index, provides a wide range of disciplines, and the selection of the world's leading scientists is an important application of the citation analysis method in the evaluation of scientists. Garfield presented three types of large-scale citation statistics to select the world's outstanding scientists.

(1) Citation statistics

In 1977, Garfield first used the citation index in SCI and selected 250 scientists from nearly 30,000,000 citations from 1961 to 1975. The selected papers need to be cited more than 4000 times, and the average author in 15 years is cited only 50 times. The characteristics of the selection is the first author of the statistics only and does not involve the co-author. The average age of 250 people was 63 years; 42 of them were Nobel laureates (17% of 250). A total of 151 of them belonged to at least one Academy of Sciences, 60% of the total, and accounted for 1/4. These data are sufficient to explain the effectiveness of the selection of the citation analysis method.

In 1978, Garfield used the database of SCI from 1961 to 1976 for the second year. To make up for the deficiencies in the first time, the statistical characteristics merged the record of the first author citation indexes and the co-author of "source

index” statistics. The selected were cited 5496 times. The first author average citation was 1794, and the co-author average citation frequency was 3702 times. The co-author citation is much higher than that of the first author, so the co-author of the computation is reasonable. A total of 300 persons had an average age of 54 years, and 160 Academy of Science academicians accounted for more than half. Twenty-six won the Nobel Prize (8.66% of the total), and 177 people had various awards (accounted for 59%). Therefore, among the 300 Academy of Science academicians, the winners are few. Citation analysis is thus a feasible method to evaluate the results of talents.

In 1981, Garfield conducted a third citation statistics analysis. He used 1965 to 1978 (14 years) data in the SCI database and selected living scientists' written papers (including first author and co-author). He calculated the citation times according to the citation counts and the number of ranking. Among the selected, the most often cited by thousands of scientists accounted for one thousandth of the world's millions of technology researchers. In 14 years, the average of each author cited 3811 times was selected. Then, each of the communication address of authors was determined through a letter to investigate the profession, birth year, service unit, works, awards, scientific institution, and full name. The results were as follows:

- ① Citation number. Thousands of scientists were cited in the 14 years per capita. A total of 121 papers have 32 articles with the first author's name published and 89 with the co-author's name published. The average number of citations is 3811 times per year and 272 times a year.
- ② Gender and age. The number of women scientists in the table is about 23. The average age is 53 years, of which 42 to 61 years of age accounted for 77%. The youngest is 33 years old. It is generally considered that the best age for invention is about 37 years old, but with the improvement of scientific level and scale, the best age has a tendency to decrease.

Nobel Prize winners and Academy of Sciences. Among thousands of people, 44 are Nobel Prize winners, and more than half of the total number of Nobel Prize winners in 1965 to 1977 are in medicine, chemistry, and physics. We can see that the winners have written more papers and are cited more often. Among them, 378 people received the title of academician, and the organic chemist R.B. Woodward has the title of 12 academicians. The average age of academicians is 58 years old and that of non-academicians is 51 years.

Comparison of different subjects. Thousands of scientists in the table were cited by 38 professional classifications of their papers. The number of academicians, Nobel Prize wins, birth year, and so on indicate a difference between the large and the largest number of physicists, chemists, immunologists, endocrine scientists, etc.

Country and institutional analysis. Investigation of and statistics on the number of scientists should be conducted with a country's comprehensive national strength and science development. Strong and powerful countries and the number of

Table 11.1 Statistics on thousands of outstanding scientists

Country	Number of institutions	Number of scientists
US	147	736
UK	28	85
Sweden	6	42
France	7	26
Canada	9	23
West Germany	9	21
Switzerland	10	13
Australia	6	12
Japan	8	11
Israel	4	10
Denmark	3	4
Italy	3	4
Belgium	2	3
Holland	3	3
Argentina	1	1
Czech	1	1
Finland	1	1
Hungary	1	1
Norway	1	1
Spain	1	1
Soviet Union	1	1
Total	252	1000

scientists must also be indicated. For example, there are 147 units in the United States, with 736 scientists. The number of countries and scientists is shown in Table 11.1. The former Soviet Union has only 1 scientist selected, which is completely due to the SCI citation. The SCI citation statistics for the former Soviet Union and Japan are very inadequate. Similarly, in the same country, talent is concentrated in the famous units of all countries; the more famous institutions are, the more scientists there are.

(2) Citation analysis

The following conclusions can be obtained from the above analysis.

- 1) Using the citation method to choose talents is feasible. Judging from the results of the three citation statistics, selecting talents through extensive investigation is appropriate, and the results are in line with the actual general trend. The citation method can also be used to solve disputes in academics, whether it is the evaluation of the individual or unit.
- 2) The citation method can be used to predict future winners.

- 3) The citation method can be used to evaluate the scientific research ability and performance of a group. In general, the more columns with the highest number of citations, the higher the prestige of the unit. According to SCI data, the US National Institutes of Health's biomedical research literature account for 3.4%. In SCI, the authors of 12 kinds of biomedical journal reviews of 7.5% are scientists of the Institute. This shows that the National Institutes of Health's pool of talents is very effective.
- 4) The citation method can be used to determine citation classics. From the table, every selected scientist in literature cited a number of the most famous article analysis, and 250 articles having 13 journal contents (CC) were as "cited classic literature" (citation classics). By July 1981, 750 papers have been published. Classical works are one of the key points of citation analysis, which can obtain many useful results to evaluate the best journals.
- 5) The co-citation rate is high. Every scientist in the cited literature most frequently selected 300 papers from 300 people in the table. Only 35 are writers. The average number of authors in each paper is 3, and the frequently cited authors are mostly co-authors. Among them, 2-4 author literature account for the majority. Therefore, the phenomenon of co-authors is indeed worthy of attention, that is, in the table of 300 people, more than 250 can reflect the objectivity of things.

(3) Limitation of the evaluation of talents by citation

Although the citation method is an effective method for the evaluation of talents, it is limited by several factors. For example, the use of the original data and citation habits result in the problem of the occurrence of the watch and co-author scoring problems. The development trend for the future will be to increase the statistical object, expand the scope of the profession, increase patents, monographs, standards, and other varieties, and reasonably deal with the phenomenon to make the citation method perfect.

11.2.4 Forecasting of Future Winners

(1) Relation of literature citations to authors' honor

Garfield's citation analysis showed that outstanding scientists, Nobel Prize laureates, and academicians of the Chinese Academy of Sciences accounted for a large proportion rather than a few winners. According to statistics, 1961 to 1971 Nobel Prize winners were cited 222 times; a was elected as an academician of NAS a year ago, and the average was 99 times. This statistic shows that the higher the honor, the more the number of references cited. This phenomenon reflects the literature citations and authors of the honorary title link to a certain extent.

(2) Nobel Prize winner prediction

The number of papers published by scientists is an objective evaluation criterion that can measure the achievement of scientific progress and predict the future of scientific development. Garfield successfully predicted the 1969 Nobel Prize winner by using the data provided by SCI in 1968. Table 11.2 shows the number of citation times in 1967 by the number of the ranking of the first 50 scientists. In the table, two scientists (H.R. Barton Derk (41) and Gell-mall Muray (6)) received the Nobel Prize in 1969. At the same time, among the 1981–1982 Nobel Prize winners of 13 people, 6 have long appeared in Garfield's statistics. Sixty are from the National Academy of Sciences in 1982, 17 people are from the United States, and thousands of people are in the table. These data are sufficient to prove the accuracy of the method of citation analysis. Through the analysis of the research work of these scientists, we can predict the trend and future of research and development.

Table 11.2 Sorted according to citations before 50 scientists' ranking (1967)

Rank order	Name	Citation numbers	Rank order	Name	Citation numbers
1	LOWRY OH	2921	26	ELIELEL	721
2	CHANCE B	1374	27	STREITWIESERA	717
3	LANDAU LD	1174	28	MULLIKEN RS	712
4	BROWN HC	1150	29	JACOB F	711
5	PAULIN GL	1063	30	BORN M	710
6	GELLMANN	942	31	BRACHET J	706
7	COTTON FA	940	32	WINSTEIN S	702
8	PEPLE JA	933	33	ALBERT A	687
9	BELLAMY LJ	906	34	LUFT JH	674
10	SNEDECOR GW	904	35	DEDUVE C	673
11	BOYER PD	893	36	VONEULER US	668
12	BAKER BR	876	37	FIESER LF	666
13	KOLTHOFF	853	38	HUISGEN R	661
14	HERZBERG G	842	39	NOVIKOFF AB	655
15	FISCHER F	826	40	GOODWIN TW	643
16	SEITZ F	822	41	BARTON DHR	632
17	DJERASSI C	801	42	FISHER RA	631
18	BERGMAYER HU	754	43	BATES DR	627
19	WEBER G	750	44	FLORY PJ	626
20	REYNOLDS ES	748	45	STAHL E	626
21	MOTT NF	741	46	DEWARM JS	619
22	ECCLES JC	737	47	GILMAN H	618
23	FEIGL F	729	48	FOLCH J	618
24	FREUD S	727	49	DISCHE Z	614
25	PEARSE AGE	726	50	GLICK D	609

Garfield initially developed the “Science Citation Index” to provide a new type of search tool. However, recently, the development of more proof has been provided regarding the use of “Science Citation Index” is in the evaluation of science and technology management application. In addition to the evaluation of scientific and technological talents, people have started using citation analysis as a powerful weapon for the evaluation of scientific research institutions and national and regional, scientific research ability and academic activities. Good results have been obtained.

11.3 Informetrics and Regional and Institutional Research Evaluation

In the process of carrying out scientific management, the problems of how to scientifically evaluate a scientific work based on its own achievements and efficiency and how to scientifically evaluate the scientific research level and academic status of national and regional scientific research institutions emerge. Famous Hungarian scientometrics scientist Professor Braun used SCI data and the citation analysis method to evaluate the scientific research level and academic status of national and regional research institutions. The data were sorted and a global scientific schematic was derived. Very good results were achieved. Braun created a scientific quantitative evaluation method that has been widely used. Below is a brief introduction of Blauwen (T. Braun) and Schubert (A. Schubert) in this area of research and evaluation work.

11.3.1 Scientific and Quantitative Study of Braun

One of the most attractive targets in the scientific measurement is to create an index system for evaluating scientific institutions and to study the development of various scientific organizations and groups. Professor Braun opted to remove States, the former Soviet Union, Britain, Federal Germany, France, and Japan, which are six countries with 32 national natural science literature citations. Through a comparative analysis, he presented a scientific quantitative evaluation index system for the evaluation and analysis of the department of all countries in the world to learn their status and development trends.

(1) Data processing principles

Braun included SCI data from 1976 to 1980 and paid special attention to the statistical analysis of 1980 from 1978 to 1979 in the citation rate.

In this paper, only measurement journals published papers and publications on the statistics, reports, monographs, reviews, communications, and technical notes of

several publications. For ease of measurement and screening, classification of each paper involved determining the date of publication, publication type, author nationality, and subject area. Given that processing 200 million is difficult, they were classified not one by one and according to their respective disciplines. According to KiffaPince, journals were classified according to subject and theses according to the journal they belong to. In the process of periodical classification, the core journals of various disciplines are listed first. Then, in the premise of statistical calculation of the interaction between journals, to conduct subject classification of the collection of all journals, several of the more comprehensive journals are divided into individual disciplines. Finally, all are merged into the following fields.

- Clinical medicine
- Biomedicine
- Biology
- Chemistry and physics
- Earth science and space science
- Engineering science
- Psychology
- Mathematics
- Others

(2) Scientific measurement indicators for scientific evaluation

Blauwen defined the following 12 measurement indicators to comprehensively evaluate the status quo of the 32 countries of science and technology.

- 1) Number of first authors: number of first authors of various countries and relevant papers in various fields from 1978 to 1979.
- 2) Number of papers: number of papers in each country and the related papers in each subject area.
- 3) Subject distribution of the paper: calculate the distribution and percentage of papers published in various countries during from 1976 to 1980.
- 4) Number of papers not cited: SCI gives the number of papers published from 1978 to 1979, each country, and its number of papers published in each subject area.
- 5) Percentage of papers not cited: percentage of the total number of papers published from 1978 to 1979 that are not cited in the papers.
- 6) Number of highly cited papers: according to the state and subject, the papers that were cited more than 10 times the number of papers published from 1978 to 1979.
- 7) Percentage of highly cited papers: the index indicates that the highly cited papers account for the percentage of the total number of papers published in 1978 to 1979.
- 8) Actual citation rate: number of papers published between 1978 and 1979 and according to state statistics, those published in 1980.
- 9) Expected citation rate: number of papers published in a certain country or a subject in different journals multiplied by the impact factor of the

corresponding journal; the sum of the product is the expectation of the country or the subject.

- 10) Relative citation rate: ratio of actual citation rate to expected citation rate, that is, the relative citation rate. The relative citation rate can be a cross disciplinary comparison.
- 11) Average citation rate (impact factor): number of papers published from 1978 to 1979 in 1980 divided by the number of papers published by the total number of papers published during the year.
- 12) Average influence factors: by country and subject, the number of relevant papers published from 1978 to 1979 divided by the total number of papers. The business is not only an average impact factor. The average impact factor provides information on the quality of the papers, which is the quality information of various periodicals published in the papers. The ratio between the average citation rate and the average citation rate is the relative citation rate.

Figure 11.3 shows the total number of scientific papers and their subject distribution in China. Table 11.4 shows the actual citation rate of Chinese science and technology papers for 1980 SCI, the expected citation rate, and relative citation rate. Table 11.3 shows the impact of China's scientific and technological papers for 1980 SCI.

11.3.2 Schubert and Other Scientists' Measurement Research

After Braun created the citation analysis method for the evaluation of national and regional scientific research activities, Schubert (A. Schubert) used SCI statistical data on 96 countries around the world with 114 kinds of main disciplines and

Fig. 11.3 China science and technology papers and their subject distribution (1976–1980 SCI)

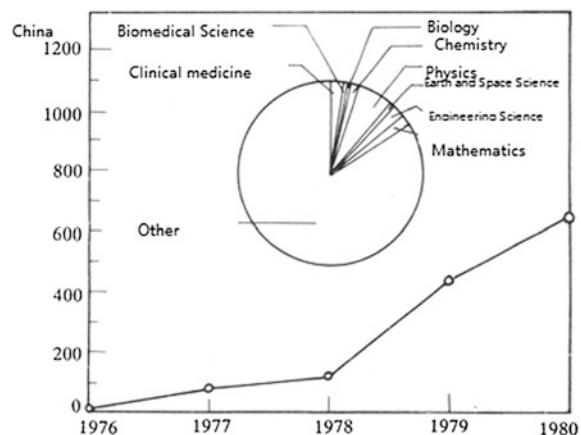


Table 11.3 Actual citation rate, expected citation rate, and relative citation rate of Chinese scientific and technical papers (1980 SCI)

Index number	VIII	IX	X
	Citation rate		Relative citation rate
	Actual value	Expected value	
Clinical medicine	9	26.72	0.34
Biomedical science	1	18.48	0.05
Biology	0	1.43	0
Chemistry	3	6.32	0.48
Physics	131	71.06	1.84
Earth science and space science	0	1.73	0
Engineering science	6	5.64	1.6
Psychology	0	0	0
Mathematics	2	1.13	1.77
Others	57	73.81	0.77
Total	209	206.23	1.01

Table 11.4 Citations of Chinese science and technology papers (1980 SCI data)

Index number	XI		XII	
	Average citation rate	Standard mean error	Mean influence factor	W—statistics
Clinical medicine	0.692	0.231	2.005	-5.907
Biomedical science	0.167	0.167	3.08	-17.48
Biology	0	0	0.283	*****
Chemistry	1	1.283	2.077	-0.839
Physics	4.226	2.039	2.292	0.948
Earth science and space science	0	0	0.577	*****
Engineering science	0.75	0.473	0.705	0.096
Psychology	0	*****	0	*****
Mathematics	0.5	0.589	0.283	0.368
Others	0.118	0.02	0.153	-1.375
Total	0.376	0.066	0.371	0.075

evaluated their level and corresponding position using the scientific measurement index system. The years from 1981 to 1985 were used together with the scientific research activities and literature communication of countries and areas in the world.

Schubert's use of SCI and Braun mainly includes science and technology literature publishing quantity, citation, the cited, and the absolute number and variety of relative quantity.

- 1) Amount of literature. It mainly refers to the SCI included in research papers, reviews, academic essays, and letters for the four types of literature.
- 2) Publishing share. It mainly refers to the publication of a percentage of the total number of publications in a country or region (according to SCI's collection of statistics).
- 3) Number of citations. It pertains to the number of papers cited in a country or region in the SCI citation index.
- 4) Citation shares. This refers to the citation volume percentage of the total citation of a country or region in the world.
- 5) Expected citation rate. This rate can be obtained through the calculation of the impact factor of a country or a region or a journal average citation rate (see Braun index in the expectation of citation rate).
- 6) Actual citation rate. This is the average citation rate for the actual statistics of the SCI database.
- 7) Relative citation rate (RCR). RCR is the ratio of the actual citation rate to the expected citation rate.

$$PRC = \frac{X}{Y},$$

where X means the actual citation rate and Y means the expected citation rate.

- 8) Posting index (AI). Its specific definition is as follows:

$$AI = \frac{X}{Y},$$

where X denotes a country in a given field of papers published in the world's total share of the paper and Y means the share of the papers published in the entire research area of the country in the world's papers.

Through a comprehensive evaluation of the above indicators, we can roughly estimate and speculate the scientific and technological development level of a certain country or region and the status of literature exchange trends.

The AI values in Table 11.5 are 1 because the statistics include all areas of natural science and is not specific to a particular given subject. If only one particular discipline is included, the AI and AAI values would generally not be 1 (Table 11.6).

11.3.3 Evaluation of Scientific Research Institutions

The statistical evaluation of scientific research institutions with the method of citation analysis is also very objective and effective; it has been increasingly used to evaluate the scientific research status and performance ranking of various scientific research institutions in recent years. Table 11.7 shows the China Science and Technology Information Research Institute and the State Education Commission in

Table 11.5 Ninety-six countries and regions' (1981–1985) main scientific measurement index list (excerpt) (including all fields of natural science)

Country and region	LPQ	WL (100%)	CQWC	WC (100%)	ECR	ACR	RCR	PI	CI
USA	706114	36.81	3029147	50.82	4.08	4.29	1.05	1	1
UK	171858	8.96	572999	9.61	3.18	3.33	1.05	1	1
USSR	139501	7.27	97285	1.63	0.79	0.7	0.88	1	1
Japan	134107	6.99	343723	5.77	2.84	2.56	0.9	1	1
Germany-FR	112625	5.87	345696	5.80	2.84	3.07	1.08	1	1
France	89538	4.67	250460	4.20	2.9	2.8	0.97	1	1
Canada	80001	4.17	235231	3.95	3.1	2.94	0.95	1	1
Indian	50581	2.64	48680	0.82	1.62	0.96	0.59	1	1
Italy	43706	2.28	101079	1.70	2.84	2.31	0.81	1	1
Australia	42775	2.23	120697	2.03	2.89	2.82	0.98	1	1
Netherlands	32657	1.70	114787	1.93	3.31	3.51	1.06	1	1
Sweden	31543	1.64	115554	1.94	3.16	3.66	1.16	1	1
Switzerland	23454	1.22	106958	1.79	3.75	4.56	1.22	1	1
Israel	20422	1.06	54889	0.92	3.29	2.69	0.82	1	1
Poland	16879	0.88	22721	0.38	1.92	1.35	0.7	1	1
Germany-DR	16732	0.87	23622	0.40	1.45	1.41	0.97	1	1
Belgium	16395	0.85	50523	0.85	3.03	3.08	1.02	1	1
Denmark	15798	0.82	54844	0.92	3.04	3.47	1.14	1	1
Spain	15660	0.82	25828	0.43	2.57	1.65	0.64	1	1
Czechoslovakia	14624	0.76	17619	0.30	1.44	1.2	0.84	1	1
Finland	12066	0.63	32748	0.55	2.82	2.71	0.96	1	1
South-Africa-R	10439	0.54	15508	0.26	1.91	1.49	0.78	1	1
Austria	10297	0.54	20524	0.34	2.1	1.99	0.95	1	1
Norway	9785	0.51	27113	0.45	2.82	2.77	0.98	1	1
New-Zealand	9424	0.49	18432	0.31	2.26	1.96	0.86	1	1
Hungary	8988	0.47	16063	0.27	2.13	1.79	0.84	1	1
PR-China	8347	0.44	4716	0.08	1.24	0.56	0.45	1	1
Brazil	6987	0.36	10116	0.17	2.51	1.45	0.58	1	1
Argentina	5396	0.28	8289	0.14	2.6	1.54	0.59	1	1
Egypt	4788	0.25	3561	0.06	1.42	0.75	0.52	1	1
Bulgaria	4687	0.24	3553	0.06	1.28	0.76	0.59	1	1

Notes LPQ: literature publication quantity; WL(100%): world literature(100%); CQWC: citation quantity WC(100%); ECR: expected citation rate; ACR: actual citation rate; RCR: relative citation rate; PI: posting index; CI: citation index

1993 through the SCI database, including the number of papers in various scientific research institutions and the output of the world university rankings. This is very important for the understanding of the academic level worldwide and the position and gap among universities in the world (Table 11.8).

Table 11.6 Life sciences' (1981–1985) main science citation index list (excerpt)

Country and Region	Number	Proportion (100%)	CQ	WC (100%)	ECR	ACR	PostI	PostI	CI
USA	434145	40.76	1917483	53.18	4.26	4.42	1.04	1.11	1.05
UK	113084	10.62	388497	10.78	3.22	3.44	1.07	1.19	1.12
Japan	62683	5.89	171567	4.76	3.18	2.74	0.86	0.84	0.83
Germany-FR	58456	5.49	172259	4.78	2.72	2.95	1.08	0.93	0.82
Canada	48759	4.58	153232	4.25	3.26	3.14	0.97	1.1	1.08
France	46348	4.35	128116	3.55	2.92	2.76	0.95	0.93	0.85
USSR	32339	3.04	17516	0.49	0.69	0.54	0.79	0.42	0.3
Australia	28281	2.66	81273	2.25	2.9	2.87	0.99	1.19	1.11
Sweden	24344	2.29	88942	2.47	3.19	3.65	1.15	1.39	1.27
Italy	23861	2.24	52659	1.46	2.83	2.21	0.78	0.98	0.86
Netherlands	20439	1.92	73620	2.04	3.45	3.6	1.05	1.13	1.06
Indian	16876	1.58	16499	0.46	1.17	0.98	0.57	0.6	0.56
Switzerland	13238	1.24	56045	1.55	3.67	4.23	1.15	1.02	0.87
Israel	12530	1.18	31473	0.87	3.26	2.51	0.77	1.11	0.95
Denmark	12108	1.14	39426	1.09	2.95	3.26	1.11	1.38	1.19
Belgium	10244	0.96	35160	0.98	3.13	3.43	1.1	1.13	1.15
Finland	9091	0.85	25832	0.72	2.9	2.84	0.98	1.36	1.3
Spain	7445	0.70	13080	0.36	2.89	1.76	0.61	0.86	0.84
German-DR	7418	0.70	10173	0.28	1.43	1.37	0.96	0.8	0.71
Norway	7400	0.69	22196	0.62	2.92	3	1.03	1.36	1.35
South-African-R	7420	0.68	10365	0.29	1.81	1.43	0.79	1.25	1.1
Czechoslovakia	6742	0.63	7110	0.20	1.31	1.05	0.8	0.83	0.67
New-Zealand	6666	0.63	13076	0.36	2.21	1.96	0.89	1.27	1.17
Austria	6266	0.59	12553	0.35	2	2	1	1.1	1.01

Note Q: citation quantity; WC(100%): world citation(100%); ECR: expected citation rate; ACR: actual citation rate; PostI: posting index; PubI: publication index; CI: citation index

11.3.4 Evaluation of Scientific Research in China

(1) Statistics and analysis of Chinese scientific papers

With the continuous development of scientific management theory and application in the world, research work on scientific research evaluation was also carried out in China. Since 1987, the Institute of Scientific and Technical Information of China has analyzed the number and cited times of papers published by Chinese scientific and technical personnel (including Taiwan Province) at home and abroad. Every year, the statistical results were released to the public. This has become a major event in the scientific community and has received much attention in the community.

Table 11.7 Top 50 universities that published SCI papers in 1989

Rank	University names	Country	Number of papers
1	Harvard	U.S.A	5548
2	University of Washington	U.S.A	5132
3	University of California, Los Angeles	U.S.A	3963
4	University of Tokyo	Japan	3430
5	Columbia University	U.S.A	3293
6	Cornell University	U.S.A	3199
7	University of Michigan	U.S.A	3188
8	Stanford University	U.S.A	2925
9	University of Toronto	Canada	2917
10	University of California, San Francisco	U.S.A	2910
11	Johns Hopkins University	U.S.A	2751
12	University of Wisconsin-Madison	U.S.A	2737
13	Kyoto University	Japan	2722
14	Yale University	U.S.A	2568
15	Osaka University	Japan	2514
16	Massachusetts Institute of Technology	U.S.A	2410
17	University of California, Berkeley	U.S.A	2410
18	University of Minnesota	U.S.A	2378
19	University of California, San Diego	U.S.A	2374
20	University of Pennsylvania	U.S.A	2346
21	University of Florida	U.S.A	2320
22	Moscow University	Russia	2262
23	University of California, Davis	U.S.A	2177
24	Ohio State University	U.S.A	2128
25	University of Cambridge	Britain	2108
26	University of Illinois at Champaign-Urbana	U.S.A	2037
27	Duke University	U.S.A	2028
28	University of Arizona	U.S.A	1938
29	University of Oxford	Britain	1928
30	University of Pittsburgh	U.S.A	1927
31	University of Southern California	U.S.A	1901
32	Tohoku University	Japan	1871
33	University of Texas at Dallas	U.S.A	1864
34	University of North Carolina	U.S.A	1789
35	University of Chicago	U.S.A	1788
36	University of British Columbia	Canada	1758
37	Lund University	Sweden	1754
38	Iowa State University	U.S.A	1730
39	Leiden University	Holland	1691
40	Pennsylvania State University	U.S.A	1680

(continued)

Table 11.7 (continued)

Rank	University names	Country	Number of papers
41	McGill University	Canada	1670
42	University of Texas	U.S.A	1645
43	Sapienza University of Rome	Italy	1615
44	Ludwig Maximilian MuenchenUnitversitaet	Germany	1605
45	Northwestern University	U.S.A	1593
46	Texas Agriculture and Mechanic University	U.S.A	1587
47	KarolinskaInstitutet	Sweden	1578
48	Purdue University	U.S.A	1553
49	University of Alberta	Canada	1534
50	California Institute of Technology	U.S.A	1517

After years of improvement and accumulation, a relatively strict and complete system was formed. International papers were retrieved from SCI, EI, and ISTP, which were three authoritative search tools in the world. They can represent the highest level in today's world of scientific research. Domestic papers were obtained directly from the 1214 kinds of major scientific and technical journals. The focus was on the most important research papers and technical development achievements in China, which made the statistical results possess academic authority.

Currently, the methods and data of Chinese scientific papers are widely available. This has become an important basis and a powerful tool for the majority of colleges and universities, scientific research institutions, medical departments, and various enterprises and institutions to conduct scientific research management, discipline, and talent assessment. It improves the level of China's science and technology management and promotes the development of China's scientific and technological undertakings.

Under the encouragement and guidance of this evaluation management approach, many universities and research institutions in China have formed a more active atmosphere of academic competition. This improves the overall level of research significantly, and outstanding talents come to the fore. The international influence and role of scientific papers increase gradually. This forms a good development trend. With the statistical results in 1996 as an example, we briefly explain scientific research evaluation.

1) Chinese scientific papers' impact and location in the world

In 1996, in accordance with the statistical data of SCI, EI, and ISTP, science and technology personnel in China published a total of 27,569 journal articles and conference papers internationally, which showed an increase of 4.4% compared with the previous year. Sorted by the number of papers, China ranked 11 in the world. If the number of papers cited by SCI is to be counted, there were 14459 papers in China, which ranked 14 in the world. If the number of papers cited by EI is to be counted, there were 9147 papers in China, which ranked 6 in the world.

Table 11.8 Papers of 50 universities in 1989 SCI

Rank	University	Country	Number (papers)
1	Harvard University	US	5548
2	University of Washington	US	5132
3	University of California, Los Angeles	US	3963
4	University of Tokyo	Japan	3430
5	Columbia University	US	3293
6	Cornell University	US	3199
7	University of Michigan	US	3188
8	Stanford University	US	2925
9	University of Toronto	Canada	2917
10	University of California, San Francisco	US	2910
11	Johns Hopkins University	US	2751
12	University of Wisconsin-Madison	US	2737
13	Kyoto University	Japan	2722
14	Yale University	US	2568
15	Osaka University	Japan	2514
16	Massachusetts Institute of Technology	US	2410
17	University of California, Berkeley	US	2410
18	University of Minnesota	US	2378
19	University of California, San Diego	US	2374
20	University of Pennsylvania	US	2346
21	University of Florida	US	2320
22	Moscow University	Russia	2262
23	University of California, Davis	US	2177
24	Ohio State University	US	2128
25	University of Cambridge	UK	2108
26	University of Illinois at Champaign-Urbana	US	2037
27	Duke University	US	2028
28	University of Arizona	US	1938
29	University of Oxford	UK	1928
30	University of Pittsburgh	US	1927
31	University of Southern California	US	1901
32	Tohoku University	Japan	1871
33	University of Texas at Dallas	US	1864
34	University of North Carolina	US	1789
35	University of Chicago	US	1788
36	University of British Columbia	Canada	1758
37	Lund University	Sweden	1754
38	Iowa State University	US	1730
39	Leiden University	Holland	1691
40	Pennsylvania State University	US	1680

(continued)

Table 11.8 (continued)

Rank	University	Country	Number (papers)
41	McGill University	Canada	1670
42	University of Texas	US	1645
43	Sapienza University of Rome	Italy	1615
44	Ludwig Maximilian MuenchenUnitversitaet	Germany	1605
45	Northwestern University	US	1593
46	Texas Agriculture and Mechanic University	US	1587
47	KarolinskaInstitutet	Sweden	1578
48	Purdue University	US	1553
49	University of Alberta	Canada	1534
50	California Institute of Technology	US	1517

Table 11.9 Citation situation in 1996 of China's science and technology papers that were collected in SCI from 1991 to 1995

Year	Number of papers collected (represented by A)	Number of papers cited(represented by B)	Cited rate: B/A	Times cited: C	C/B	C/A
1991	5408	1196	0.221	2121	1.773	0.392
1992	6224	1582	0.254	2790	1.764	0.448
1993	6645	1923	0.289	3570	1.856	0.537
1994	6721	2227	0.331	4242	1.905	0.631
1995	7980	1898	0.238	3077	1.621	0.386
Total	32978	8826	0.268	15800	1.79	0.479

There were 3963 conference papers, which ranked 11 in the world. The countries whose total papers exceeded China's are the United States, Japan, Britain, Germany, France, Canada, Italy, Russia, Australia, and the Netherlands.

The citation times of papers in the world reflects the international influence of papers. From 1991 to 1995, there were 32978 papers collected by SCI in China, among which 8826 papers were cited 15800 times in 1996. The cited rate is about 0.27, and each cited paper was cited 1.79 times on average. The average citation rate in the world is 2.23.

2) Chinese scientific papers' sorting of times cited in the world

Cited statistics here is the number of times papers published by scientific and technical personnel in China and collected by SCI are cited internationally. Table 11.9 shows the citation case in 1996 of scientific papers collected by SCI from 1991 to 1995. The 10 universities and 10 institutes whose papers were published from 1993 to 1995 and collected by SCI and were most cited in 1996 are listed in Tables 11.10 and 11.11.

Table 11.10 Top 10 universities whose papers were cited most in 1996 among China's science and technology papers that were collected in SCI from 1993 to 1995

Name of colleges and universities	Number of papers included from 1993 to 1995 (represented by A)	Number of papers cited in 1996 (represented by B)	Times cited in 1996 (represented by C)	B/A	C/A	C/B
Nanjing University	1009	378	790	0.375	0.783	2.09
Beijing University	707	237	455	0.335	0.644	1.92
Tsinghua University	551	173	321	0.314	0.583	1.855
Fudan University	542	185	372	0.341	0.686	2.011
University of Science and Technology of China	602	178	334	0.296	0.555	1.876
Lanzhou University	485	145	232	0.317	0.507	1.6
Jilin University	385	128	249	0.332	0.647	1.945
Nankai University	361	107	200	0.296	0.554	1.869
Shandong University	295	99	148	0.336	0.502	1.495
Zhejiang University	328	102	165	0.311	0.503	1.618
Total	5238	1732	3266	0.331	0.624	1.889

3) Co-authored situation of domestic papers

In 1996, 1227 kinds of domestic scientific and technical journals included a total of 116,778 papers. These papers relate to 340,473 authors, and each paper has an average of 2.9 authors. A total of 27,271 papers have only one author, accounting for 23.3% of the total papers. A total of 76.7% of the papers were co-authored. The specific circumstances are shown in Table 11.12.

4) Age, job title, and sex distribution of domestic authors

According to the statistical analysis of 20,502 papers attached with authors' age, we found that 12453 papers' authors were all under the age of 35, which accounted for 49.8%. A total of 5132 papers' authors were between 30 and 45 years old, which accounted for 20.5%. A total of 5940 papers' authors were between 46 and 59 years old, which accounted for 23.7%. The number of papers whose authors were all under the age of 45 accounted for 70%.

Table 11.11 Top 10 institutes whose papers were cited most in 1996 among China's science and technology papers that were collected in SCI from 1993 to 1995

Name of institutes	Number of papers included from 1993 to 1995 (represented by A)	Number of papers cited in 1996 (represented by B)	Times cited in 1996 (represented by C)	B/A	C/A	C/B
Institute of Physics CAS	623	232	453	0.372	0.727	1.953
Shanghai Institute of Organic Chemistry, CAS	330	202	483	0.612	1.464	2.391
Changchun Institute of Applied Chemistry Chinese Academy of Sciences	381	147	272	0.386	0.714	1.85
Institute of Metal Research, CAS	269	93	158	0.346	0.587	1.699
Institute of Chemistry, CAS	230	86	187	0.374	0.813	2.174
Dalian Institute of Chemical Physics, CAS	158	75	174	0.475	0.101	2.32
Fujian Institute of Research on the Structure of Matter, CAS	114	55	89	0.482	0.781	1.618
Institute of Theoretical Physics, CAS	182	73	151	0.401	0.83	2.068
Shanghai Institute of Optics and Fine Mechanics, CAS	189	54	83	0.286	0.439	1.537
Lanzhou Institute of Chemical Physics, CAS	125	54	98	0.432	0.784	1.815
Total	2601	1071	2148	0.412	0.826	2.006

According to the statistical analysis of 30,732 papers attached with authors' job title, we found that 14,685 papers were written by scholars with senior titles, which accounted for 47.8%. A total of 5258 papers were written by graduate students, which accounted for 17.1%.

According to the statistical analysis of 25,180 papers attached with authors' sex, we found that 20,993 papers were written by males and 4187 papers were written by females. The ratio of male to female is about 5:1.

The brief data above show that the statistical analysis results have a variety of evaluation function. We can clearly understand the situations of various universities and research institutions, and the situations include scientific development, personnel structure, institutional academic level, and subject exchange. Thus, we can provide quantitative data and a decision-making basis for scientific research and technological innovation activities.

Table 11.12 Analysis of co-authored domestic papers

Co-authored type	Number of papers	Percentage of total co-authored papers (%)
Co-authored by scholars in one unit	61112	69.00
Co-authored by scholars in one province or city	14868	16.80
Co-authored by scholars in different provinces or cities	10773	12.10
Co-authored by scholars in different countries	1844	2.10
Total number of coauthored papers	88597	100.00

(2) Study on the measurement index of Chinese scientific journals

As mentioned in Chap. 10, JCR plays a large role in the evaluation of scientific journals. To achieve a more objective and accurate evaluation of Chinese scientific journals' function and quality in terms of scientific activities and exchange of literature, the Institute of Scientific and Technical Information of China developed the Chinese Science and Technology Journal Citation Reports (CJCR), which fills the gap in this area. It assists the majority of scientists, journal editors, and science and technology management departments in evaluating, selecting, and utilizing journals scientifically. Table 11.13 lists the partial instantiation of the report.

11.4 Informetrics and Science and Technology Forecasting

11.4.1 *Informetrics and Science and Technology Forecasting*

(1) Basic concept of science and technology forecasting

According to the basic principles of forecasting science and its history and current status of technological development, science and technology forecasting is used to analyze and speculate about the development prospects of science and technology and the extent of its impact on social progress to arrive at a predictable conclusion. Correct analysis and scientific forecasting of the future development of science and technology is an important prerequisite and guarantees that the right decision and scientific management are applied. It is significant for promoting scientific and technological progress and innovation ability in China and enhancing the comprehensive national strength.

Table 11.13 Chinese science and technology journal citation reports (part of chemical journal)

CODE	Journal	Total times cited	Number of papers published in 1995 and 1996	Times cited in 1997 of papers published in 1995 and 1996	Impact factor	Number of papers published in 1997	Times cited that year when papers were published in 1997	Immediacy index	Regional distribution of papers	Fund papers	Cited half-life
T003	Engineering Plastics Application	163	175	113	0.646	96	1	0.01	17.2	13	0.95
T014	Plastics Industry	179	156	95	0.609	73	2	0.027	19.2	17	1.81
T022	China Plastics	127	163	71	0.436	95	2	0.021	17.6	26	1.55
T005	Journal of the Chinese Ceramic Society	364	239	78	0.326	127	3	0.024	17.8	63	4.69
T074	Natural Gas and Chemical Industry	123	158	47	0.298	84	5	0.06	16	20	3
T007	Journal of Chemical Engineering	272	227	61	0.269	116	9	0.078	15.4	76	4.24
T002	Polymer Bulletin	72	68	17	0.25	36	1	0.028	10.8	11	3.5
T071	Carbon	54	74	17	0.23	42	0	0	14.4	2	4.2
T009	Chemical Reaction Engineering and Technology	77	122	27	0.221	68	19	0.279	14.4	14	1.25

(continued)

Table 11.13 (continued)

CODE	Journal	Total times cited	Number of papers published in 1995 and 1996	Times cited in 1997 of papers published in 1995 and 1996	Impact factor	Number of papers published in 1997	Times cited that year when papers were published in 1997	Immediacy index	Regional distribution of papers	Fund papers	Cited half-life
T008	Chemical Metallurgy	73	127	28	0.221	73	11	0.151	11.4	26	1.82
T010	Ion Exchange and Adsorption	144	187	41	0.219	102	4	0.039	17.2	40	3.77
T013	Journal of Synthetic Crystals	89	141	28	0.199	40	4	0.1	14.2	20	3.08
T070	China Surfactant Detergent & Cosmetics	127	168	29	0.173	82	1	0.012	20.2	5	5.86
T053	Photosensitive Material	42	101	17	0.168	53	0	0	14.2	1	2.8
T075	China Adhesives	54	132	21	0.159	82	1	0.012	20.8	5	2.33
T067	Synthetic Fiber in China	66	91	14	0.154	65	2	0.031	13.6	1	5.67
T025	Chemical Industry	143	167	25	0.15	72	5	0.069	18.8	19	6.66
T018	China Synthetic Rubber Industry	111	245	34	0.139	130	5	0.039	15.8	20	3.92

(continued)

Table 11.13 (continued)

CODE	Journal	Total times cited	Number of papers published in 1995 and 1996	Times cited in 1997 of papers published in 1995 and 1996	Impact factor	Number of papers published in 1997	Times cited that year when papers were published in 1997	Immediacy index	Regional distribution of papers	Fund papers	Cited half-life
T051	Glass Enamel	61	131	18	0.137	71	2	0.028	16.2	4	4.5
T015	Carbon Technology	52	117	16	0.137	60	2	0.033	14	9	2.88
T021	Journal of East China University of Science and Technology	133	264	35	0.133	138	6	0.044	2.6	64	4.03
T060	Coal Chemical Industry	21	79	10	0.127	41	0	0	11.6	1	2.13
T072	Inorganic Chemicals Industry	126	174	22	0.126	105	2	0.019	22.8	2	6.08
T016	Journal of Chemical Engineering of Chinese Universities	42	136	17	0.125	74	2	0.027	12.8	44	2.33
T054	Sea-Lake Salt and Chemical Industry	48	156	19	0.122	84	1	0.012	15.4	1	2.5

(continued)

Table 11.13 (continued)

CODE	Journal	Total times cited	Number of papers published in 1995 and 1996	Times cited in 1997 of papers published in 1995 and 1996	Impact factor	Number of papers published in 1997	Times cited that year when papers were published in 1997	Immediacy index	Regional distribution of papers	Fund papers	Cited half-life
T004	Bulletin of the Chinese Ceramic Society	87	157	19	0.121	101	1	0.01	17	20	3.6
T017	Chemical Processing of Forest Products	68	112	13	0.116	57	0	0	12.6	21	4.5
T019	Chinese Journal of Pharmaceutical Industry	261	464	52	0.112	216	3	0.014	23.8	9	4.33
T020	Journal of Beijing University of Chemical Technology	46	137	15	0.11	76	7	0.092	1.6	18	2.33
T063	Modern Chemical Industry	108	299	33	0.11	161	1	0.006	22.6	10	3.39

Science and technology forecasting mainly includes the following five aspects.

1) Scientific predictions—predictions about the prospects for the development of science

Scientific prediction is mainly used to forecast the development trend of the entire scientific system to analyze the differentiation, cross, penetration, and evolution direction of various disciplines; forecast the development prospects of existing and new disciplines that may arise, especially interdisciplinary and integrated subjects; and forecast the practical value of certain scientific theories, the development trend, and the evolution law of the cycle of science and technology. By applying the basic principles of informatics to science and technology forecasting, we can evaluate and predict the development trends and prospects of a particular discipline or field of knowledge.

2) Technology predictions—predictions about the prospects of technological development

Technology prediction is mainly used to forecast the development prospects of several major technology areas and technical inventions; forecast new materials, new techniques, new equipment, and new methods that may arise; and forecast the applications of new technologies.

3) Product predictions—predictions about the prospects of product development and application by using information measurement methods

Senlong, a Japanese intelligence expert, predicted the product structure and development prospects of three major polymer materials, namely, plastic, rubber, and fiber, by analyzing the occurrences of keywords. His prediction was consistent with the actual situation and the production output of chemical industry. Therefore, the predicted results are essential, accurate, and convincing.

4) Scientific and technological undertakings predictions—predictions about the future prospects of science and technology

Predictions about scientific and technological undertakings are mainly used to predict the development prospects of scientific research institutions, scientific and technological personnel, science and technology books, intelligence information, scientific and technological exchanges, technology transfer, and the like.

5) Science and technology's prediction about the economy and social impact

It is mainly used to predict the impact of science and technology on all aspects of social and economic development.

(2) Basic principle of science and technology forecasting by using information measurement methods

Science and technology forecasting is achieved through an exploratory analysis of historical and existing scientific and technical development paths. Most of these

development paths exist in the form of scientific literature. The development of technology will inevitably lead to increased amounts of scientific literature. The main output of fundamental research is in the form of scientific papers, and patents represent a country's innovation capability. The numbers of scientific papers and patent ownership have become the main basis to evaluate the research and innovation ability of a country or region, and they are important indicators used to measure a country's comprehensive national strength. We can roughly understand the development trend of world science and technology by analyzing the number of scientific papers and patent ownership of each country in a given period. Therefore, forecasting the development prospects and trends of science and technology by informetrics is a new and important means of science and technology forecasting.

The growth of a discipline generally experiences the process of budding, development, maturation, and differentiation. In this process, papers, on behalf of the research results of this discipline, must be changed according to the number and content of the composition. That is, the embryonic stage only has a handful of papers, the contents of which are mostly the discussion of experimental facts and subject concepts. In the development stage, the number of papers increase significantly, and the contents become complete and mature. Theoretical papers increase significantly. When a discipline enters the mature stage, the growth rate of papers slows down and gradually reaches saturation, and the proportion of application papers increases. This indicates that the discipline has matured, and new developments will rarely appear. When differentiating a new field of knowledge, the number of papers in the subject rapidly decreases. This indicates that the growth process of a discipline is closely linked with changes in the number and contents of papers. This linkage is an important basis for predicting the development trends and prospects of a discipline by informetrics. In addition, a significant characteristic of the development of modern science and technology is the mutual penetration between disciplines. Mutual penetration between disciplines is closely related with mutual citation among papers. The citation phenomenon reflects the linkages between disciplines and marks the inheritance and development of the relationship between science and technology. The citation chain between scientific literature specifically and vividly reflects the structure of science. Hence, through these relationships, we can track and predict the production, development, differentiation, mutual penetration, and trends of a discipline. Through the relationship among literature, technology, and research projects, we can also conduct appropriate science and technology forecasting.

(3) Methods and basic steps of science and technology forecasting

The methods of science and technology forecasting can be divided into three categories: qualitative, quantitative, and integrated. Examples include expert consultation method (key technology method), in-depth study method, Delphi method, scenario analysis, analogy, trend extrapolation, brainstorming, AHP, smoothing, regression analysis, correlation matrix, and decision tree method. The information

measurement method is one of the important newly developed methods of science and technology forecasting.

The basic steps of science and technology forecasting are as follows:

- 1) The prediction target is analyzed, and a predicted target is determined.
 - 2) Relevant information is collected, and literature data are quantified.
 - 3) Through induction and analysis, predictive models are built, and conclusions are drawn through the analysis.
 - 4) The reliability and accuracy of conclusions are examined.
- (4) The science and technology development trend of countries is forecasted by using information measurement methods.

According to statistics, in the five years from 1997 to 2001, the world's scientific papers increased annually. The five-year change in the total number of scientific papers in the world obtained by counting SCI is shown in Fig. 11.4. The changes in the number of papers of 12 countries and regions are shown in Fig. 11.5. In addition to Russia, whose number of papers increased because of significant changes in the national political system, other countries and regions all had different degrees of growth.

By analyzing the total number of papers in various countries in the five years, we find that a technologically advanced country has an absolute advantage. As shown in Fig. 11.6, the total number of papers of the United States, Britain, Japan, Germany, and France in five years has reached 4,787,483 and accounts for 63.06% of the world's total number of papers. The total number of papers of the United States is 1,549,388 and accounts for one-third of the world's total number of papers. By contrast, developing countries account for a lower proportion. Although the number of Chinese papers increased rapidly recently, they account for only 2.66%.

The five-year average annual growth rate in the world's papers is 2.21%. Figure 11.7 shows that the average annual growth rates of the United States, France, and Russia, three technologically developed countries, are lower than the world's average annual growth rate. Meanwhile, the average annual growth rate of developing countries and regions are higher than the world's average annual growth rate. The average annual growth rate of China's scientific papers is 20.62%, which is nearly 10 times the world's average annual growth rate. Developing countries whose growth rate is over 10% include Singapore, South Korea, and Brazil.

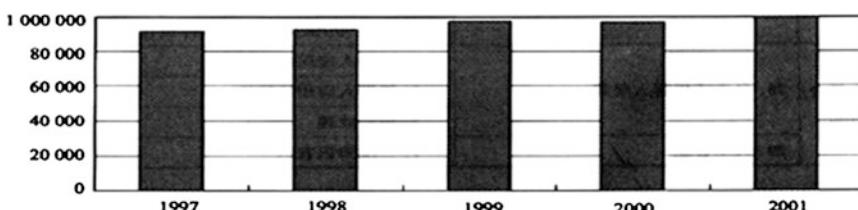


Fig. 11.4 Change in the total number of scientific papers in the world from 1997 to 2001

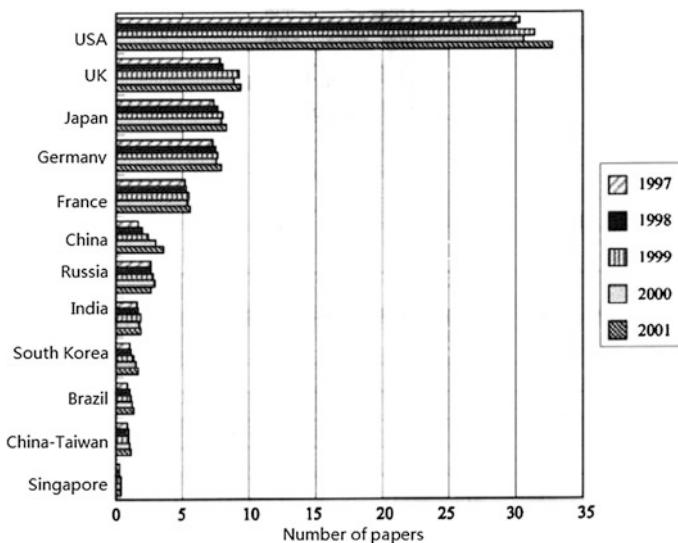


Fig. 11.5 Change in the number of scientific papers of 12 countries from 1997 to 2001

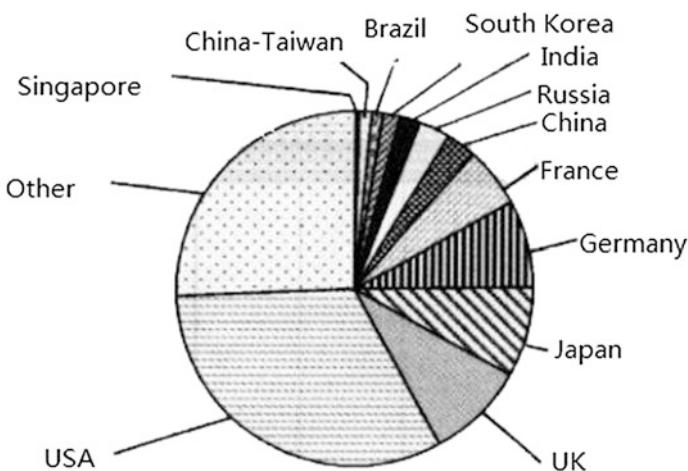


Fig. 11.6 Total number of scientific papers of 12 countries from 1997 to 2001 and their proportion in the world's total papers

Overall, for the number of world output of scientific papers, developed countries account for a very large share, but the growth rate is slow. Only the growth rate of Britain and Germany is higher than the world's growth rate. The scientific paper output of developing countries is still low, but the growth rate is very high.

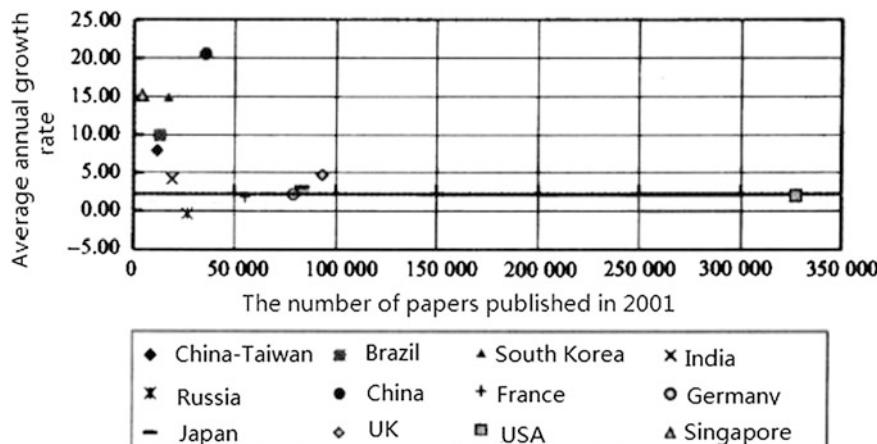


Fig. 11.7 Comparison of the total number of scientific papers of 12 countries and regions in 2001 and the average growth rate from 1997 to 2001

11.4.2 Conducting Technical Evaluation and Prediction by Using Patent Documents

(1) Important source of technical intelligence

A patent document is the actual recording of the various application values of technical inventions; it contains a wealth of detailed technical information. Therefore, the number of national patent documents represents the level of a country's innovation capability. Reports indicate that 90–95% of technical inventions in the world are published in patent literature. Currently, the number of patent documents is large, and the growth rate is quite high. The cumulative amount of patents in the world has reached 37 million, and about one million patent documents are published annually. The patent literature database that is composed of a large number of patents not only collects a variety of science and technology intelligence in the world within the maximum range, but also stores the major technological inventions of all fields of technology chronologically. According to statistics, 80–95% of the world's patent literature intelligence cannot be obtained by other technical literature. Thus, in modern science and technology intelligence work, patent literature is an extremely important source of technical intelligence.

With the rapid development of science and technology, patent documents, which symbolize the wealth of knowledge contributed to society by inventors, have received increasing attention from scientific and intelligence communities. Activities, such as obtaining technical intelligence needed from the patent literature, have become increasingly frequent. Practice has proven that patent literature not only tells people about new inventions or improvements in a particular expertise,

but also contains a large number of technology trends and economic intelligence. Through an analysis of the volume growth of patent documents as well as regrouping and statistical analysis of various patent documents, meaningful data can be obtained to evaluate and predict certain expertise. New technology areas actively explored by inventors and technology areas that have achieved breakthroughs in recent years can also be revealed. We can understand the technical level of each country and compare the pace of development and technology trends of countries. Likewise, we can compare the technical activities of major manufacturers in various professional fields.

(2) Profile of the US Office of Technology Assessment and Forecast

To better utilize patent literature and explore technical evaluation and forecasting methods by using existing patent literature, the United States Patent and Trademark Office set up the Technology Assessment and Forecast Office (OTAF) in 1971. The main tasks of the office are to collect, aggregate, analyze, and organize information and publish publications on technical evaluation and forecasting. Another key task is writing and providing special reports according to the requirements of government agencies and private enterprises or companies. The contents of thematic reports mainly involve conducting technical evaluation and prediction by using patent document data.

The working methods and procedures of OTAF are as follows:

- 1) The subject to be studied is determined, and the relevant categories are identified based on the subject. Determining the category is the basis of technical evaluation and prediction and the first step in initializing technical evaluation and forecasting.
- 2) After the category is determined, the analysts propose data requirements, and computer patent literature databases provide data required for the relevant category.
- 3) Analysts will count, analyze, compare, and identify these data, thereby obtaining several meaningful conclusions. Analytical reports are subsequently written.

The main results of the technical evaluation and forecast planning of the US Patent Office are reflected in OTAF publications. These publications can be divided into two categories: reports openly published and having a unified title and number and special reports prepared for several users. The first type of publication has been reported in the US Government Reports Announcement and Index (GRA & I), thus providing retrieval convenience.

- 1) Open publications mainly contain the following:

- ① Technical evaluation and forecast report

The technical evaluation and forecast report is a set of relatively complete materials with consecutive numbering found in 1973. Its contents can be divided into three

sections: overview of the statistical evaluation, analysis of specific areas, and method-related patent activity analysis and discussion of the data selection criteria.

② Patent overview

In the beginning of 1983, the publication reported a total of four areas, namely, synthetic fuels, solar energy, microelectronics, and biotechnology. Each patent overview focused on the situation within a technical field of patent activity. Its contents included activity overview, patent owner analysis, the first page of the patent in one year (which reflected all the bibliographic contents, abstracts, and illustrations), and citation analysis.

③ Industrial patent activities

The report was first published in 1981 and republished in 1982; the content has been updated. It is the outcome of the Office of Evaluation in cooperation with the National Technical Information Service (NTIS). It mainly lists the institutions that have many patents and analyzes them in a chronological or alphabetical order.

- 2) The special report is only for the requesting user. A special report prepared according to standard procedure has the following forms:

① Technology overview report

It lists the number and date of the patent invented within a certain number of years in the technical field in accordance with the nation and company. Patent numbers and names are then listed according to the company. Such reports show the status of nations or companies within a specific technical field to determine their technical strength.

② Agency overview report

This report lists the patent numbers and names of the patent obtained by a country, a company, or several companies within a certain period.

③ Overview report of multiple companies' patent activity

④ Product report determined by Standard Industrial Classification

For more than a decade, the US Patent Office implemented technical evaluation and forecast planning and achieved certain results. Several of these results can help us directly conduct research and foreign trade. Several provide a reference for our country to carry out patent activity analysis and promote the use of patent literature to gain increased technical intelligence.

- 3) Patent activity analysis

Two basic methods can be used for patent documents. One is to obtain information by analyzing the patent literature content, which focuses on the technical content of a specific patent specification. This method is familiar to the majority of scientists, engineers, and technicians. The other method is to obtain information through an analysis of patent activity, which mainly involves studying the patent document as a

whole to understand the status of technology development and predict trends of technological development with the method of quantitative analysis of information. The information obtained includes who (or which company, country) and when published how many patents in which field; how many patents related to a specific technology were published within a particular field; utilization of patents and predictions of this situation; which technical field is particularly active or which technical field has had breakthrough development; and predicting what new products will emerge in the market. At the same time, the information can be used to infer about research strength, technological strength, and academic level according to the company's patents. Analysis of a wide range of patents can also reveal the technical ability and level of a country. For example, analysis of the science and technology capacity in the Science Indicators Report proposed by the US National Science Board (NSB) classifies patent activity as an important indicator. This intelligence is very important for us to carry out technology transfer, technology trade, and improve competitiveness.

The patent activity analysis method is feasible for the following reasons.

- 1) A patent document is novel, advanced, and practical.
 - 2) Research and development work in society can be directly, comprehensively, and timely reflected in patent literature, and this intelligence function is not possible with any other source of information.
 - 3) The number of patent documents is large, and these patent documents form a complete patent literature database.
 - 4) Patent specification contains many standardized bibliographic data, the application date, the date of approval, transfer cases, priority, and classification, so an in-depth statistical analysis can be conducted for different bibliographic contents.
 - 5) Many countries have a network version of the patent literature database, which creates the conditions to analyze increasingly frequent and active patenting activities by using a computer.
- (4) Launching technical evaluation and prediction

The method of using patent literature to carry out technical evaluation and forecasting is implemented through patenting activity analysis. On the basis of the data provided by the patent literature database, the following technical evaluation and predictive analyses can be performed.

1) Evaluation of technological advances and trends

A very close link exists between technological progress and patent literature, which is the basic premise of technical evaluation and prediction by using patent literature. Analysis of the volume growth of patent literature in different categories reflects the technological advances, the inventor's attention, new technology trends, and so on in the field of science and technology. Analysis of the equivalent number of patents can reflect the technology trends in several other countries to a certain extent.

2) Identification of the most active areas of technology

The growth rate of patent documents in each category reflects trends in the contemporary world of technological development to determine the most active technology fields. First, we discharged the 50 most active documents according to the chemical, electronic, and mechanical areas. We found that the general purpose programmable digital computer system in the electrical field increased by 693 new patents in the three years of 1976–1978, and the hybrid digital data processing system increased by 552 new patents within three years. The increased amount of patent literature of the two most active categories within three years in the electric field accounted for about 16% of the increased amount of literature of all 50 of the most active categories. In the mechanical field, the three categories of solar heater, solar cooker, and solar furnace and heating system with the sun heating source (related to solar energy) have the largest increase in the amount of patent literature. Foreign patents account for only 10.7–14.5%. During the same period, the total proportion of foreign patents is 37% on average. Fifty categories increased by 6,014 new patents in three years, of which about 30% were related to energy technologies. Hence, energy technology is a rapidly developing field of technology. US energy technologies, particularly solar energy, are ahead of that of other countries.

(5) Examples of technical evaluation and forecasting

- 1) Evaluate and predict the innovation capacity of countries in the world from changes in number of US patents.

America is possesses not only basic research output powers, but also a large amount of patents. US patents and those of other developed countries epitomize the creative ability of the world. The number of patents obtained by the United States and other countries or regions released in the Technology Assessment and Forecast Report of the United States Patent and Trademark Office in April 2002 is shown in Table 11.14. Analysis of Table 11.14 results in the following conclusions.

- ① In the five years from 1997 to 2001, the number of US patents was 2,548,929. A total of 450,390 of them were obtained by Americans, accounting for 55.15% of the total. In other words, nearly half of the patents were obtained by other countries or regions aside from the US. Thus, the worldwide technology competition in the United States is very intense. A considerable part of America's technological superiority is also occupied by other countries.
- ② From the number of patents in one to five years, we can find that states or regions, including the United States, all show a positive growth trend. This indicates that the world's ability to innovate is becoming increasingly powerful.
- ③ According to the number and proportion of US patents, the technologically advanced countries have significant advantages. The number of patents in Japan accounted for nearly 20% of the total number of US patents. Germany, France, and Britain also have more than 2% of the share. By contrast, developing countries have a small proportion, but South Korea's share of US

Table 11.14 National and regional patent numbers granted by the US Patent Office (1997 to 2001)

Area	1997	1998	1999	2000	2001	5-year cumulative	Proportion (%)
World total	124146	163209	169146	176087	184051	816636	100.00
USA	69922	90701	94090	97014	98663	450390	55.15
Japan	24191	32118	32514	32923	34890	156636	19.18
Germany	7292	9582	9895	10822	11894	49485	6.06
France	3202	3991	4097	4173	4456	19919	2.44
UK	2904	3726	3900	4090	4356	18976	2.32
China-Taiwan	2597	3805	4526	5806	6545	23279	2.85
South Korea	1965	3362	3679	3472	3763	16241	1.99
Singapore	100	136	152	242	304	934	0.11
Brazil	67	88	98	113	125	491	0.06
China	66	88	99	163	266	682	0.08
India	48	94	114	131	179	566	0.07

patent number is close to 2%. The number of patents in China is small and accounts for a small share, showing that its technological innovation capability is still relatively weak. The technological competition with the United States is also at a disadvantage.

2) Technical evaluation and predictive analysis (with semiconductor technology as an example)

In the 150 most active small categories in the US patent literature database from 1976 to 1978, the semiconductor technology is particularly impressive. In the 50 most active small categories in the electric field, 19 of them are related to semiconductor technology, such as solid state devices, systems, and circuits. The increased number of literature accounts for about 45% of the total increased number of literature in 50 small categories. Moreover, five small categories in the mechanical field and two small categories in the chemical field are related to semiconductor devices and technology. Thus, the semiconductor technology is still one of the world's most active modern technology at present. It is a comprehensive technology, and its high-speed development is reflected in the substantial increase in the amount of literature.

Through the above-mentioned types of chart analysis, we can draw the following conclusions.

- ① Speed development of semiconductor processing technology. From the relative growth rate and the actual growth rate of the amount of patent documents' application and approval, we can find that the majority of semiconductor processing technology had a new breakthrough in the late 1960s and always maintained the momentum of high-speed development.
- ② As shown by all types of U.S. patent literature database, the amount of patents obtained by Americans accounted for more than half. In most semiconductor

processing technology categories, the number of patents obtained by Americans accounted for 70–80%, and some even reached 90%. The US absolutely controls this field of technology, but the US patents acquired by other countries are increasing.

- ③ Most semiconductor processing technology patents are controlled in a few dozen companies. The technology competition among these companies is very intense. The technical activity trends can be seen from the amount of acquired patents and the change in the contents of various companies.

11.4.3 Using Informetrics to Predict the Development Trend of a Discipline

By applying the basic principles of informetrics to science and technology prediction, we can estimate the development trends and prospects of a particular discipline or field of knowledge. Below is an example to explain the basic methods and processes.

(1) Examples of the basic method

1) Determining the target

In the field of chemistry, chromatography is a new method for separation analysis of substances developed rapidly in the past two decades. It uses different extents of adsorption of the adsorbent for various components to be tested to achieve the purpose of identification and separation. We analyzed changes in the number and content of chromatography literature from 1965 to 1978 in foreign countries by informetrics to depict the current development outline and trend of the subject.

2) Statistics and analysis of literature

To master all the documentation sources related to chromatography, one must first select a search tool that reports the entire contents of chromatography literature as the object of analysis. The Analytical Chemistry of *Chemical Abstracts* reports only the categories of “Inorganic Analytical Chemistry” and “Organic Analytical Chemistry.” Analytical chemistry’s applied technology is scattered in various other parts, which caused great difficulties in document statistics. Therefore, we used the British *Analytical Abstracts* and the bibliography of *Chromatography Magazine* as part of the study. We counted the number of excerpts in *Analytical Abstracts*; the number of abstracts in 1965 was about 6640, about 9820 in 1970, and about 8980 in 1975. Table 11.15 lists the percentage of documents in various analytical methods.

The above figures show that in the past decade, the proportion of classical analytical methods (gravimetric, volumetric method, and others) decreased by nearly half. Spectrometry was the most widely used and occupied the first place. Chromatography occupied the third place in 1965 and jumped to the second place in 1975 because of significant development.

Table 11.15 Percentage of documents in various analytical methods

Analytical method	1965 (%)	1970 (%)	1975 (%)
Chromatography	24	30	27
Liquid chromatography	8	11	9
Spectroscopy (including electronic paramagnetic, NMR, spectrophotometry)	36	37	36
Electrochemical	10	13	20
Gravimetry, volumetric method, and others	30	20	17

To understand the development of the various branches of chromatography, we counted all bibliographies in Chromatogram Magazine published in 1970, 1975 to 1976 (the first half), and 1977 to 1978 (the first half). They were arranged in accordance with gas chromatography, liquid chromatography, paper chromatography, thin layer chromatography, and the four other branches. The statistical results are shown in Table 11.16.

Table 11.16 shows that in the last eight years, the number of papers on chromatography significantly decreased. The number of liquid chromatography documents increased and occupied the first place.

We also counted the distribution of the amount of literature of various chromatographic methods by experimental techniques, and the results are shown in Table 11.17. The table shows that documents related to column efficiency and pillar padding accounted for the largest proportion. This reflects the prevailing research priorities of foreign chromatography.

3) Conclusion

By using informetrics to analyze the results of foreign literature related to chromatography, we delineated the current development outline of chromatography. The study of foreign liquid chromatography had made significant progress, and its technological focus was to improve column efficiency and column filler over the last decade. This is also the direction of the development of recent research on chromatographic analysis.

Table 11.16 Percentage of the various branches of chromatography literature

Branches of chromatography	1970 (%)	1975–1976 (%)	1977–1978 (%)
Gas chromatography	29	26	25
Liquid chromatography	26	39	45
Paper chromatography	13	7	5
Thin layer chromatography	32	28	25

Table 11.17 Distribution of the amount of literature of various chromatographic methods by experimental techniques (%)

Experimental techniques	1970		1975–1976(first half of a year)		1977–1978(first half of a year)	
	Gas chromatography (%)	Gas chromatography (%)	Liquid chromatography (%)	Gas chromatography (%)	Liquid chromatography (%)	
Detector	21	19	1600.00%	19	17	
Column efficiency and pillar padding	17	22	3300.00%	28	33	
Equipment and Materials	22	24	1400.00%	18	20	
Physical and chemical properties measured	10	9	400.00%	5	5	
Automation	13	16	8	17	–	
Others	17	10	25	13	25	

(2) Using informetrics to evaluate and predict the world's two hot spots—gene and nano technologies

The 21st century is the era of rapid development and wide application of life science and information technology, and nano science and technology facilitate the development of the technology containing the two areas. Governments and enterprises in technologically advanced countries have a large investment in nanotechnology in an attempt to seize the science and technology strategic high ground in the 21st century. Many western countries and companies view nanotechnology as state and commercial secrets and strictly control exports. Developing countries, including China, are not far behind. They have carried out some research on the nano domain and are undertaking research and exploitation.

Currently, in the study of life sciences, genetic research is very active. The research results emerged in large numbers. On June 26, 2002, scientists from the United States, Japan, Germany, France, Britain, and China announced that they completed a working draft of the human genome sequencing. The major findings indicate that human beings have taken a crucial step in the process of study themselves.

Many new theories are to be explored and discovered in the two hot research fields, and the research results are still mainly presented in the form of research papers at present. Thus, by analyzing the number and change trend of nano and genetic research papers in various countries and regions, we could determine the science and technology research capacity and status in the international arena of these countries and regions as well as the development trend of the subject.

1) Statistics and analysis of gene papers

① Output analysis of the world gene papers from 1997 to 2001

As a hot research topic in the field of life sciences, gene research achievements are rich, and a large number of papers were published. As shown in Table 11.18, with the total number of SCI papers increasing year by year, the number of gene papers also increases every year. The proportion of the total number remains substantial at 2.1%, and growth has stabilized.

Table 11.18 World gene papers and their proportion

Year	Gene papers	Total number of SCI papers	Percentage (%)
1997	15994	745819	2.1
1998	16557	770591	2.1
1999	16542	785222	2.1
2000	16371	778453	2.1
2001	17098	815463	2.1

Table 11.19 Number of SCI gene papers and circumstances in several countries and regions

Year	1997	1998	1999	2000	2001	Total	Share (%)	Average annual growth (%)
World	15994	16557	16542	16371	17098	82562	100	1.71
USA	7254	7080	7096	6833	7295	35558	43.07	0.22
UK	1436	1603	1577	1338	1303	7257	8.79	-1.94
Germany	1356	1446	1455	1427	1539	7223	8.75	3.30
France	1055	1192	1156	1098	1156	5657	6.85	2.56
Japan	2131	2299	2319	2234	2469	11452	13.87	3.90
Russia	161	180	146	204	225	916	1.11	10.73
China	145	196	239	330	422	1332	1.61	30.77
India	76	106	77	91	133	483	0.59	19.11
Brazil	69	92	136	140	142	579	0.70	21.38
Korea	201	254	256	313	365	1389	1.68	16.51
Singapore	27	30	39	42	47	185	0.22	15.18
China-Taiwan	135	134	123	191	210	793	0.96	14.07

② Output analysis of gene papers in different countries and in Taiwan Province of China

In the five years from 1997 to 2001, the number of gene papers increased year by year. Table 11.19 shows that the proportion of gene papers published in the five scientific and technological countries (USA, Japan, Britain, Germany, and France) has reached 81.33% of the total gene papers. By contrast, the number of gene papers produced by China and South Korea, two developing countries, accounts for 1.61% and 1.68%, respectively, and is slightly higher than that of Russia. Other developing countries and regions account for less than 1%.

Table 11.19 shows that China had a rapid growth trend, and the annual percentages were 0.9%, 1.2%, 1.4%, 2.0%, and 2.5% from 1997 to 2001, respectively. The United States showed a downward trend, and the annual percentages were 45.4%, 42.8%, 42.9%, 41.7%, and 42.7% from 1997 to 2001, respectively.

From 1997 to 2001, the average annual growth rate of the world's gene papers was 1.71%. The growth rate of Japan, Germany, and France was higher than this value, and the growth rate of the United States and Britain was lower than this value. The United Kingdom experienced negative growth. The situation in China, India, Brazil, South Korea, and Singapore is close. They all had a low proportion of the number of papers, but the annual growth rate of the number of papers was higher than the world annual average. Growth in Taiwan Province of China was also more than the world annual average. China's annual growth rate was the highest and reached 30.77%.

Table 11.20 World nano papers and their proportion

Year	Nano papers	Total number of SCI papers	Percentage (%)
1997	2798	745819	0.38
1998	3599	770591	0.47
1999	4866	785222	0.62
2000	5439	778453	0.70
2001	7656	815463	0.94

2) Statistics and analysis of nano papers

① Output analysis of world nano papers from 1997 to 2001

As another research hotspot in the world, nanotechnology is developing very rapidly and exhibits broad application prospects, so countries provide increased investment in nano technology. The number of nano papers showed an increasing trend, and the proportion of nano papers in the total number of SCI papers also showed an increasing trend. Table 11.20 shows that in the five years from 1997 to 2001, the number of nanotechnology papers increased by nearly three times from 2798 to 7656. The share in the total proportion increased by 2.5 times from 0.38% to 0.94%, and the growth continued.

② Output analysis of nano papers in different countries and Taiwan Province of China

In the five years from 1997 to 2001, the number of nano papers increased annually, and the total number was 24358. Table 11.21 shows that the number of nano papers published in the five technologically advanced countries (the USA, Japan, Britain, Germany, and France) was 14673 and accounted for 60.23% of the total papers. The number of papers published by China was 2482 and accounted for 10.19% of the total. According to the sorting of number of nano papers in five years, China ranked third in the world, only after the US and Japan. From 1997 to 2001, the number of Chinese nano papers accounted for 8.22%, 7.83%, 9.57%, 10.86%, and 11.92% respectively. The growth trend is evident. Compared with the situation in 1997, the share of US nanotechnology papers increased by 3.97% points from 28.31 to 32.38% in 2001. The share of Japanese nanotechnology papers increased by 0.59% points from 12.76 to 13.35% in 2001, China's share increased by 3.70% points.

From 1997 to 2001, the average annual growth rate of the world's nano papers was 29.09%. Except for France, the growth rate of the other four technologically advanced countries is higher than the world average growth rate. For developing countries and regions, including China, their average annual growth rate is higher than the average growth rate of major powers and the world average. In particular, the average annual growth rate of Brazil, South Korea, Singapore, India, and Taiwan Province of China, which have a lower proportion of the number of papers, is not only higher than the world average annual growth rate, but also higher than the Chinese growth rate.

Table 11.21 Number of SCI nano papers and circumstances in several countries and regions

Year	1997	1998	1999	2000	2001	Total	Share (%)	Average annual growth (%)
World	2798	3599	4866	5439	7656	24358	100	29.09
USA	795	1041	1269	1333	2479	6917	28.40	35.97
UK	103	130	165	188	337	923	3.79	36.58
Germany	258	332	412	475	733	2210	9.07	30.60
France	209	220	288	334	531	1582	6.49	27.78
Japan	357	450	564	648	1022	3041	12.48	31.00
Russia	134	144	200	215	318	1011	4.15	25.44
China	230	282	466	591	913	2482	10.19	42.29
India	42	62	95	109	224	532	2.18	55.27
Brazil	8	10	28	36	85	167	0.69	92.42
Korea	27	48	116	151	329	671	2.75	91.87
Singapore	8	15	34	50	85	192	0.79	82.81
China-Taiwan	23	29	45	52	122	271	1.11	57.86

(3) Conclusions

From the above analysis, we can draw the following conclusions.

- 1) From 1997 to 2001, the number of papers in the two hot research fields increased.
- 2) In the two hot research fields, the five-year cumulative number of gene papers was 82562, and the number nano papers was 24358. The former is 3.4 times the latter, showing that the life sciences era has arrived. The annual average growth rate indicates that the world's average annual growth rate of gene papers was 1.71%, and the world's average annual growth rate of nano papers was 29.09%. The latter is 17 times the former, showing that the pace of development of nanotechnology research exceeded the pace of the development of genetic research.
- 3) By analyzing the number of papers of various countries and regions in the world in the two hot research areas, we can see that the share of United States, Britain, Japan, Germany, and France is about 80% in the field of genetics and about 60% in the nano field. This percentage indicates that a large number of relevant research results are still mastered by science and technology power. However, China has witnessed a breakthrough in the field of nanotechnology research with an increasing number of papers, and its share of the world's share has exceeded 10%.
- 4) By analyzing the average annual growth rate of papers in the two hot research areas, we find that the share of the absolute number of papers published by developing countries and territories is low, but the growth rate is higher than the world's average. The growth rate far exceeds science and technology development, thus showing the development potential of developing countries and territories.

11.4.4 Prospects of Product Development and Application Using Informetrics

In early 1980, Small Senlong from Japan Information Center Science and Technology studied and forecasted the prospects of polymer materials by using the information quantitative analysis method. He proposed the analysis report, examined the polymer industry in the eighties from intelligence in the seventies, and obtained good results.

By using the JOIS-S line retrieval system, Small Senlong surveyed the number of papers related to polymer materials published in Scientific and Technical Literature Breaking News from April 1978 to December 1979; then he examined about 32000 papers related to rayon, rubber, plastics (including coatings, adhesives), and various aggregates from JOIS-S Science and Engineering Database. Afterward, he selected keywords in accordance with JICST thesauri and counted the number of times plastics, rubber, fiber, and other major nouns were used. The results are shown in the following three tables (Tables 11.22, 11.23, 11.24).

According to the statistics of keywords about plastic, rubber, and fiber, Small Senlong predicted the product structure and prospects of the three polymer materials and reached the following conclusions.

The situation of the product varieties of the plastics industry in the eighties was as follows: thermoplastics was still dominant; polyolefin (including polyethylene), acrylic resin, and polystyrene were still remarkable products; and PVC was a “hot” variety.

Table 11.22 Number of major plastics keywords

Plastic nouns	Number of times
Thermoplastics	10314
Acrylic resin	1777
Acetal resin	206
Fluorine resin	658
Polyamide	1129
Polyimide	305
Polyolefin	2991
Polyethylene	2071
Polypropylene	851
PVC	1391
Polycarbonate	418
Polystyrene	1539
Thermosetting plastics	2893
Amide resin	216
Epoxy	1429
Unsaturated polyester	449
Reinforced plastics	1527

Table 11.23 Number of major rubber keywords

Rubber nouns	Number of times
Styrene—butadiene rubber	268
Acrylonitrile—butadiene rubber	177
Isoprene rubber	130
Urethane rubber	134
Ethylene—propylene rubber (including EPOM)	173
Neoprene	116
Silicon	173
Butadiene rubber	174
Natural rubber	286

Table 11.24 Number of major synthetic fiber keywords

Fiber nouns	Number of times
Acrylic fiber	315
Polyamide fibers	705
Polyester	184
Polyolefin fibers	189
Rayon	333
Acetate	65
Triacetate	22

The trend of rubber varieties in the eighties is not yet clear and continues to be evaluated. The data in Table 11.23 show that natural rubber products have attracted people's attention, and scholars are actively studying natural rubber. We conclude that natural rubber technology will exhibit considerable development in the future.

Polyester fibers were overwhelming in the eighties. Polyamide fibers also elicited people's interest, whereas acetate was increasingly neglected.

With the statistics on the number of papers related to rubber and plastic, we can predict the prospects of these materials (Table 11.25).

Table 11.25 Number of papers on rubber and plastics

Use	Number of papers
Packaging materials	1098
Of which: food packaging	337
Building materials	820
Of which: composites	127
Auto industry	584
Aviation, aerospace industry	218

According to the data in the table, we infer the following:

- (1) Rubber and plastic, as packaging materials, will remain useful in the future; they will be used as building materials mainly for exterior work, piping, soundproofing, and insulation.
- (2) Reinforcement of plastics and other composite materials' applications as structural materials will undoubtedly increase.
- (3) Given that many rubber and plastics are used as packaging materials, special attention must be devoted to their recovery and utilization.

The above analysis shows that when conducting scientific and technical forecasting via informetrics, the statistics of papers are the basis for the development of predictive intelligence. However, the data must also be combined with related scientific and technological knowledge. Then, through flexible use and in-depth analysis, the forecast results can be expanded. For example, Table 11.22 shows that the frequency of epoxy resin is high. This high value indicates that research and production departments are attracted to epoxy resin. Epoxy resin is the main raw material for reinforced plastic, so it promotes and enhances the development of plastics. Meanwhile, the production of glass fiber should accordingly follow. To produce reinforced plastics with high strength, heat resistance, fire resistance, and other special properties, carbon, boron, and polyamide fibers also need to be developed.

The study conducted by Small Senlong through the use of informetrics also presents several problems, such as false or missed detection and keyword representativeness. However, the study still correctly predicted the general trend of polymer materials. Senlong's conclusion is also supported by evidence from the production statistics of related cultivars and is broadly consistent with the actual production of the Japanese chemical industry in the early eighties. Therefore, the predicted results are essentially accurate and compelling.

Bibliography

- Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to ‘webometrics’. *Journal of Documentation*, 53(4), 404–426.
- Bai, G. (1982). A preliminary study on the law of scientific and technical literature reading. *Science and Technology Information Work*, 11, 12–16.
- Bar-Ilan, J. (2000). The web as an information source on informetrics? A content analysis. *Journal of the American Society for Information Science*, 51(5), 432–443.
- Bookstein, A. (1977). Patterns of scientific productivity and social change: A discussion of Lotka’s law and bibliometric symmetry. *Journal of the American Society for Information Science*, 28(4), 206.
- Bookstein, A. (1990). Informetric distributions, part II: Resilience to ambiguity. *Journal of the American Society for Information Science*, 41(5), 376.
- Braun, T. (1989). *Scientific metrology index*. Science Press.
- Brookes, B. C. (1977). Theory of the Bradford law. *Journal of Documentation*, 33(3), 180–209.
- Brookes, B. C. (1970). The growth, utility, and obsolescence of scientific periodical literature. *Journal of Documentation*, 26(4), 283–294.
- Brookes, B. C. (1970). The growth, utility, and obsolescence of scientific periodical literature. *Journal of Documentation*, 26(4), 283–294.
- Brookes, B. C. (1984). Towards informetrics: Haitun, Laplace, Zipf, Bradford and the Alvey programme. *Journal of Documentation*, 40(2), 120–143.
- Burrell, Q. L. (1985). Anto on ageing in a library circulation model. *Journal of Documentation*, 41, 100–115.
- Burrell, Q. L. (1987). A third note on ageing in a library circulation model: applications to future use and relegation. *Journal of Documentation*, 43(1), 24–45.
- Burrell, Q. L. (1990). Using the gamma-Poisson model to predict library circulations. *Journal of the American Society for Information Science*, 41(3), 164.
- Burton, R. E., & Kebler, R. W. (1960). The “half-life” of some scientific and technical literatures. *American Documentation*, 11(1), 18–22.
- Chai, X. (1997). The application of citation analysis: a quantitative study of international scientific communication. *Information Studies: Theory & Application* (2), 81–84.
- Cheng, H. (1993). Dynamic differential equation. *Journal of the China Society for Scientific and Technical Information*, 4, 284–290.
- Chemnitz, B., et al. (1998). The anatomy of large-scale hypertextual web search engine. In *Products of the Seventh International WWW Conference*.
- Chen, Y. S., & Leimkuhler, F. F. (1986). A relationship between Lotka’s law, Bradford’s law, and Zipf’s law. *Journal of the American Society for Information Science*, 37(5), 307.
- Cui, L., & Hu, H. (2000). Development of co-citation cluster analysis system. *Journal of the China Society for Scientific and Technical Information*, 19(4), 308–312.
- Clarke, S. J., & Willett, P. (1997, July). Estimating the recall performance of Web search engines. In *Aslib Proceedings* (Vol. 49, No. 7, pp. 184–189). MCB UP Ltd.

- Coughlin, J. P., & Baran, R. H. (1988). Stochastic models of information obsolescence. *Mathematical and Computer Modelling*, 11, 760–765.
- Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on web-based citation analysis. *Journal of Information science*, 27(1), 1–7.
- Dahal, T. M. (1999). Cybermetrics: The Use and Implication for Science to metrics and Bibliometrics. In *3rd Conference on Science and Technology. Royal Nepal Academy of Science and Technology*.
- Davis, H. T. (1941). *The analysis of economic time series* (pp. 45–50). Bloomington: Principia Press.
- De Bellis, N. (2009). Bibliometrics and citation analysis: from the science citation index to cybermetrics. Scarecrow Press.
- de Solla Price, D. J. (1986). Little science, big science... and beyond (p. 301). *New York: Columbia University Press*.
- de Solla Price, D., & Gürsey, S. (2014). Studies in Scientometrics I Transience and Continuance in Scientific Authorship.
- Deng, L. (1983). *Library and information science*. Northeast Normal University.
- Department of mathematics and mechanics, Zhongshan University. (1983). *Probability theory and mathematical statistics*. People's Education Press.
- Ding, X., & Wang, X. (1992). On the empirical formula of burton-kebler literature aging and Morbilevmodified form. *Journal of Peking University (Philosophy & Social Sciences)* (4), 107–113.
- Ding, X. (1993). *Basis of bibliometrics*. Peking University Press.
- Ding, Y., Rousseau, R., & Wolfram, D. (2014). Measuring Scholarly Impact. Springer, Cham.
- Egghe, L. (1988). On the classification of the classical bibliometric laws. *Journal of Documentation*, 44(1), 53–62.
- Egghe, L. (1990). The duality of informetric systems with applications to the empirical laws. *Journal of Information Science*, 16(1), 17–27.
- Egghe, L. (1986). On the 80/20 rule. *Scientometrics*, 10(1–2), 55–68.
- Egghe, L., & Rousseau, R. (1992). *An introduction to the Informetrics*. Scientific and Technical Documents Press.
- Egghe, L., Rao, I. R., & Rousseau, R. (1995). On the influence of production on utilization functions: Obsolescence or increased use? *Scientometrics*, 34(2), 285–315.
- Fairthorne, R. A. (1969). Progress in documentation. *Journal of Documentation*, 25, 325.
- Fairthorne, R. A. (1969). Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. *Journal of Documentation*, 25(4), 319–343.
- Fang, S. (1990). A fractal model for the distribution of the Bradford—Zipf—Lotka. *Information Science* (5), 25–30.
- Fedorowicz, J. (1982). The theoretical foundation of Zipf's law and its application to the bibliographic database environment. *Journal of the American Society for Information Science*, 33(5), 285–293.
- Fang, S., & Li, H. (1989). Lotka's law and fractal theory. *Library and Information Service*, 6, 9–14.
- Garfield, E. (1972, November). Citation analysis as a tool in journal evaluation. *American Association for the Advancement of Science*.
- Garfield, E., & Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities* (Vol. 8). New York: Wiley.
- Gosnell, C. F. (1943). *The rate of obsolescence in college library book collections as determined by an analysis of the three select lists of books for college libraries*.
- Gelman, E., & Sichel, H. S. (1987). Library book circulation and the beta-binomial distribution. *Journal of the American Society for Information Science*, 38(1), 4.
- Glänzel, W., & Schöpflin, U. (1994). A stochastic model for the ageing of scientific literature. *Scientometrics*, 30(1), 49–64.

- Gu, X., & Liu, S. (2001). Study on retrieval details of three SCI electronic versions. *New Technology of Library and Information Service*, 17(1), 41–44.
- Gupta, D. K. (1987). Lotka's law and productivity patterns of entomological research in Nigeria for the period, 1900–1973. *Scientometrics*, 12(1–2), 33–46.
- Gupta D. K. (1993). Groscientific Literature of Nigeria for the Period 1904-1979. *International Conference on Informetrics*.
- Ha Eino, J., & Xu, X. (translate). (1988). *Scientific communication and information science*. Scientific and Technical Documentation Press.
- Hitchcock, S., Carr, L., Harris, S., Hey, J. M. N., & Hall, W. (1997, July). Citation linking: improving access to online journals. In *Proceedings of the Second ACM International Conference on Digital Libraries* (pp. 115–122). ACM.
- Helen, A. (1999). The ISI Web of Science -Links and Electronic Journals, *D-Lib Magazine*, 5(9). Retrieved April 25, 2003, from <http://www.dlib.org/dlib/september99/atkins/09atkins.html>.
- Helen, A. (1999). The ISI Web of Science—Links and Electronic Journals. *D-Lib Magazine*, 5(9). Retrieved April 25, 2003, from <http://www.dlib.org/dlib/september99/atkins/09atkins.html>.
- Han, L., & Zheng, X. (1999). SCI network review. *New Technology of Library and Information Service*, 15(6), 47–48.
- Hu, W. (1997). Research on the model of citation attenuation. *Library and Information Service*, 10, 15–18.
- ISI Links. Retrieved April 25, 2003, from <http://www.isinet.com/isi/isilinks/>.
- Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2), 236–243.
- Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2), 236–243.
- Institute of Scientific and Technical Information of China. (1997). Statistics and analysis of scientific papers in China in 1996, (Annual Research Report). *Institute of Scientific and Technical Information of China*.
- Institute of Scientific and Technical Information of China. (1998). China Journal Citation Reports.
- Institute of Scientific and Technical Information of China. (2003). *Technology development forecast and review*. Beijing Institute of Technology Press.
- Jiang, L. (2001). The distinctive links of web of science. *New Technology of Library and Information Service*, 17(4), 44–45.
- Jhorne, F. C. (1997). The citation index:author case of spurious Validity. *Journal of Clinical Psychology*, 33, 1157–1161.
- Jing, P., & Wang, W. (1999). Mathematical model of aging of scientific and technical literature and calculation of aging rate. *Library and Information Service*, 6, 16–20.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25.
- Kendall, M. G. (1960). The bibliography of operational research. *OR*, 11(1/2), 31–36.
- Kendall, M. G. (1960). The bibliography of operational research. *OR*, 11(1/2), 31–36.
- King, D. W., McDonald, D. D., & Roderer, N. K. (1981). *Scientific journals in the United States*. Dowden, Hutchinson & Ross; distributed by Academic Press.
- Kyvik, S. (1989). Productivity differences fields of learning, and Lotka's law. *Scientometrics*, 15 (3–4), 205–214.
- Lai, M., & Xu, K. (1985). *Science and technology document retrieval*. Peking University Press.
- Lawrence, C. L. (1998). Giles. Search in the WWW. *Science*, 2(80), 98–100.
- Liang, L. (2000). Scientometrics and Informetrics: World and China-thinking after the seventh ISSI Conference. *Science Research Management*, 21(3), 95–101.
- Li, C. (2001). Research object and method of Webometric. *Information Science*, 1, 66–73.
- Liu, W. (1991). A new probe into the law of scientific literature aging. *Technology and Market* (5).
- Li, Z. (1990). Grey prediction model for the growth and aging of scientific documents (gm). *Journal of the China Society for Scientific and Technical Information*, 5, 342–352.

- Li, H., Yu, G., & Rong, Y. (2000). Mathematical identification model of the process of science and technology literature aging. *Journal of Library Science in China*, 26(3), 81-84.
- Line, M. B., & Sandison, A. (1974). progress in documentation: 'Obsolescence'and changes in the use of literature with time. *Journal of Documentation*, 30(3), 283-350.
- Liang, L., & Li, X. (2003). Word frequency analysis and quantitative study of SPRU's research themes. *Science Research Management*, 24(3), 97-108.
- Liang, L., & Xie, X. (2003). Investigation of China's nanotechnology study based on frequency analysis of key words. *Studies in Science of Science*, 21(2), 138-142.
- Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, 16(12), 317-323.
- Liu, D. (1981). A new field of Information Science: Informetrics. *Technology & Market*, 4, 50-53.
- Liu, R. (1991). The cooperation degree of the authors in the periodical papers. *Library and Information Service* (1), 24-26.
- Liu, Y. (1994). Research on several basic problems of Informetrics. *Information Science*, 1, 57-66.
- Liu, Y. (1994). The relationship and difference among Bibliometrics. *Scientometrics and Informetrics. Library and Information*, 1, 19-24.
- Liu, G., Luo, C., & Wu, P. (2005). Research status and development trend of Scientometrics. *Chinese Journal of Medical Science Research Management*, 18(3), 137-140.
- Lyman, P., & Varian, H. (2004). How much information 2003?. <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/>.
- Leavens, D. H. (1953). Letter to the editor. *Econometrics*, 21, 630-632.
- Ma, J. (2003). Research progress of Webometric. *Journal of Library Science in China*, 1, 78-81.
- Ma, F. (1983). On quantitative methods of Brooks's Information Science. *Information Science* (4), 1-9.
- Ma, Y. (1998). The application of fractals in information science. *Library and Information Service*, 5, 15-16.
- Ma, M., & Cao, X. (1986). Some applications of Mathematics in Psychology. *Chinese Journal of Nature* (10), 31-36.
- Meng, L. (1983). Chinese Science Citation Analysis. *Information Science*(1), 11-21.
- Merton, R. K. (1968). The Matthew effect in science. *Science*, 159(3810), 56-63.
- Murphy, L. G. (1977). Lotka frequency distribution of scientific production. *Journal of the ASIS*, 28(4), 366.
- McKieman, G. (2002). *CitedSites: Citation indexing of web resources*.
- Nicholls, P. T. (1986). Empirical validation of Lotka's law. *Information Processing & Management*, 22(5), 417-419.
- Nicholls, P. T. (1989). Bibliometric modeling processes and the empirical validity of Lotka's law. *Journal of the American Society for Information Science*, 40(6), 379.
- Oliver, M. R. (1971). The effect of growth on the Obsolescence of semionctuctor. *Physcs Literature. Journal of Documentation*, 27, 274-285.
- Oluic-Vukovic, V. (1992). Journal productivity distribution: Quantitative study of dynamic behavior. *Journal of the American Society for Information Science*, 43(6), 412.
- Pao, M. L. (1986). An empirical examination of Lotka's law. *Journal of the American Society for Information Science*, 37(1), 26-33.
- Pao, M. L. (1982). Lotka's test. *Collection Management*, 4(1-2), 111-124.
- Pao, M. L. (1985). Lotka's law: A testing procedure. *Information Processing & Management*, 21 (4), 305-320.
- Pang, J. (2002). *Methodology of scientometrics*. Scientific and Technical Documentation Press.
- Pratt, A. D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, 28(5), 285-292.
- Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5), 292-306.

- Qin, J. (1995). Collaboration and publication productivity: An experiment with a new variable in Lotka's law. In *International Society for Scientometrics and Informetrics. International conference* (pp. 445–454).
- Qiu, J. (1988). *Bibliometrics*. Scientific and Technical Documentation Press.
- Qiu, J. (1991). Quantitative analysis on the trend of Library Science and Information Science research in China. *Journal of Library Science in China*, 17(3), 3–11.
- Qiu, J. (1994). The progress and development of bibliometrics in China. *Journal of the China Society for Scientific and Technical Information*, 6, 454–463.
- Qiu, J. (1989). Temporal distribution and aging of scientific documents in China. *Library and Information Service* (4), 7–14.
- Qiu, J. (2000). Informerterics (1) Origin and development of informetrics. *Information Studies: Theory & Application*, 1, 75–80.
- Qiu, J. (2000). Informerterics (2) Literature information growth law and applications. *Information Studies: Theory & Application*, 23(2), 153–157.
- Qiu, J. (2000). Informerterics (3) Literature information obsolescence law and applications. *Information Studies: Theory & Application*, 3, 237–240.
- Qiu, J. (2000). Informerterics (4) Scattering distribution of literature information: Bradford's law. *Information Studies: Theory & Application*, 4, 315–320.
- Qiu, J. (2000). Informerterics (5) Word frequency distribution of literature information: Zipf's law. *Information Studies: Theory & Application*, 5, 77–81.
- Qiu, J. (2000). Informerterics (6) Author distribution of literature information: Lotka's law. *Information Studies: Theory & Application*, 6, 475–478.
- Qiu, J. (2000). Informerterics (7) Concentration and scattering distribution of literature information: Bradford-Zipf-Lotka. *Information Studies: Theory & Application*, 24(1), 77–80.
- Qiu, J. (2000). Informerterics (8) Statistical analysis method for literature information and applications. *Information Studies: Theory & Application*, 2, 156–159.
- Qiu, J. (2000). Informerterics (9) Literature information citation law and Citation analysis. *Information Studies: Theory & Application*, 24(3), 236–240.
- Qiu, J. (2000). Informerterics (10) Computer-aided informetrics analysis methods and instruments. *Information Studies: Theory & Application*, 24(4), 316–320.
- Qiu, J. (2000). Informerterics (11) Applications of informetrics to library and information science—take the research and determination of core journals for example. *Information Studies: Theory & Application*, 24(5), 396–400.
- Qiu, J. (2000). Informerterics (12) Applications of informetrics to the Science of Science and management of science and technology. *Information Studies: Theory & Application*, 6, 474–478.
- Qiu, J., & Chen, J. (2001). On webmetrics and its application. *Information Studies: Theory & Application*, 24(3), 161–163.
- Qiu, J., & Xia, L. (1997). The laws of papers on quantitative analysis of information science and the determination of core journals in China. *Document, Information & Knowledge*, 3, 17–21.
- Qiu, J., & Zhang, Y. (2005). Overview of webometrics. *Library Work in Colleges and Universities*, 25(1), 1–12.
- Qiu, J., Wang, H., & Feng, X. (2002). Journal of the China Society for Scientific and technical information and the development of information science in Our Country (I)—estimation analysis of the papers from *Journal of the China Society for Scientific and Technical Information* of Twenty Years since Its First Issue Published. *Journal of the China Society for Scientific and Technical Information*, 21(5), 514–523.
- Qiu, J., Wang, H., & Feng, X. (2002). Journal of the China Society for scientific and technical information and the development of information science in Our Country (II)—Estimation analysis of the quoted documents from *Journal of the China Society for Scientific and Technical Information* of Twenty Years since Its First Issue Published. *Journal of the China Society for Scientific and Technical Information*, 21(6), 514–523.

- Qiu, J., & Xie, X. (1987). The relation between Brad Ford's law and Zipf's law and Lotka's law. *Technology & Market* (5), 29–35.
- Qiu, J., & Xie, X. (1990). A new trend of bibliometrics research—Computer-aided Bibliometrics research. *Technology & Market*, 3, 180–184.
- Qiu, J., Zhao, R., & Hou, J. (2003). Developing trend of information management in China and over the World in 2002: A bibliometric analysis. *Journal of the China Society for Scientific and Technical Information*, 22(5), 515–519.
- Rao, R. (1980). Distribution of scientific productivity and social changes. *Journal of the ASIS*, 31 (2), 111–121.
- Rao, I. K. R. (1983). *Quantitative methods for library and information science*. Taylor & Francis.
- Rogge, A. E. (1976). A look at academic anthropology: through a graph darkly. *American Anthropologist*, 78(4), 829–843.
- Roland, W. D. (1995). Lotka's law. *Journal of Documentation*, 51(1).
- Rousseau, R. (1993). A table for estimating the exponent in Lotka's law.
- Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics*, 1(1). <http://www.cindoc.csic.es>.
- Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2(3), 1. (<http://www.cindoc.csic.es/cybermetrics/articles/v2lp1.html>).
- Rousseau, B., & Rousseau, R. (2000). LOTKA: A program to fit a power law distribution to observed frequency data. *Cybermetrics*, 4(1), 1–6.
- Schorr, A. E. (1974). Lotka's Law and Library Science. *RQ*: 32–33.
- Schreiber-Herbert, (1982). The static and dynamic ageing of scientific and technical literature. *Informatic*, 29(4), 26–29.
- Science Citation Index Expanded-Journal List. Retrieved May 3, 2003, from <http://www.isinet.com>.
- Science Citation Index Expanded. Retrieved May 3, 2003, from <http://www.isinet.com/isi>.
- Seglen, P., & Aksnes, D. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research. *Scientometrics*, 49(1), 125–143.
- Smith, A. G. (1999). ANZAC webometrics: Exploring Australasian Web structures. *Proceedings of Information Online and On Disc*, 99, 159–181.
- Stinson, E. R., & Lancaster, F. W. (1987). Synchronous versus diachronous methods in the measurement of obsolescence by citation studies. *Journal of Information Science*, 13(2), 65–74.
- Su, X. (2000). Chinese Social Science Citation Index (CSSCI). *Journal of the China Society for Scientific and Technical Information*, 19(4), 290–295.
- Sugimoto, C. R. (Ed.). (2016). Theories of informetrics and scholarly communication. Walter de Gruyter GmbH & Co KG.
- Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management*, 28(1), 1–3.
- Tague, J., & Ajiferuke, I. (1987). The Markov and the mixed-Poisson models of library circulation compared. *Journal of Documentation*, 43(3), 212–235.
- The mission of website.net: Building a Web citation index. <http://www.website.net.html>.
- Thomas, O., & Willett, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science*, 26(6), 421–428.
- Trueswell, R. W. (1969). Some behavioral [sic] patterns of library user: The 80/20 rule. *Wilson Library Bulletin*, 43(3), 458–459, 461.
- Vlachy, T. (1976). Time factor in Lotka's law. *Problems de Informare si Documentare*, 10, 44–87.
- Vlachy, T. (1976). Lotka's Law. *Problems de Informare si Documentare*(10): 44-87.
- Vlachy, J. (1978). Frequency distributions of scientific performance a bibliography of Lotka's law and related phenomena. *Scientometrics*, 1(1), 107–130.
- Voos, H. (1974). Lotka's law and information science. *Journal of the ASIS*, 25, 270–272.
- Wang, B. (2000). An comparison between two sci retrieval systems. *New Century Library* (4), 19–22.

- Wang, C. (1990). The research on price's law. *Journal of The China Society for Scientific and Technical Information* (3), 203–209.
- Wang, C. (1990). *The course of Bibliometrics*. Nankai University Press.
- Wang, C. (1992). *Principles of information science*. Agricultural Science Information Service Center.
- Wang, C. (1997). *An introduction to the science of literature metrology*. Guangxi Normal University Press.
- Wang, H. (1989). The GM model and its application in literature aging. *Technology and Market*, 6, 22–25.
- Wang, C. (1992). On Matthew effect in Information Science. *Journal of the China Society for Scientific and Technical Information* (2), 114–121.
- Wang, C. (1982). Research on the authors of scientific papers. *Journal of the China Society for Scientific and Technical Information* (2), 220–225.
- Wang, C. (1996). The trend to substitute informatics for bibliometrics. *Information Studies: Theory & Application*, 1, 5–8.
- Wang, H. (2000). The axiomatical structure of literature obsolescence and Its GM. *Journal of Intelligence*, 19(5), 9–10.
- Wang, H. (2002). *Study on Informetrics*. China National Photography Art Press.
- Wang, X. (2001). Designing Web Version of Chinese Social Science Citation Index(CSSCI). *New Technology of Library and Information Service*, 16(3), 46–47.
- Wang, Z. (1991). Suggestions for further modification of Price's law. *Information Science*, 4, 37–41.
- Wang, H., & Lei, Z. (1998). Some notes on the GM model of literature's obsolescence. *Library and Information Service*, 9, 58–59.
- Wang, H., Tao, Z., Yu, M., Wang, S., & Wang, H. (1999). The study and analysis on the same-time and along-time observation of literature obsolscence. *Journal of Xinyang Teachers College (Natural Science Edition)*, 12(4), 486–489.
- Wang, H., & Xie, H. (1996). Study on the model of literature aging and citation distribution. *Information Studies: Theory & Application*, 5, 20–22.
- Wang, H., Zhou, J., Li, X., & Zhang, Z. (1999). The negation of burton-kebler literature aging measurement method should be cautious—Discussion with Yu Peiguo. *Library and Information Service*, 1, 57–59.
- Wen, W. (1985). Several problems about the Lotka's law. *Information Science*, 6, 16–22.
- Wangner-Doelet, R., & Berg, J. (1993). *Mathematische Logick von 1874 bisur Gegenwart, Eine Bibliometrixche Untersuchung*. Berlin: de Gruyter.
- Weinstock, M. (1971). *Citation indexes*. Encyclopedia of Library and Information Science (Vol. 5). In A. Kent & H. Lancour (Eds.).
- William, P. (1990). The future of bibliometrics. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics*. [s.l.]: SAGE.
- Wang, H., & Qiu, J. (2000). The trend of bibliometrics development in the 21st century. *Library Work in Colleges and Universities*, 20(4), 9–16.
- Wangner-Doelet, R. (1995). Frequency distribution of scientific productivity. *Scientometrics*, 32, 123–132.
- Xu, X. (2003). On the ecological status of the monograph of communication in mainland China (1981—2001). *Media Observer*.<http://www.chuanmei.net>, 2003-3-22 13:20:22.
- Xu, Y. (1986). The trend of informetrics. *Information Science*, 6, 62–65.
- Xu, J. (1997). Brad Ford's law and the completeness of retrieval tools. *Library and Information* (1), 47–50.
- Xu, J., & Xu, L. (2002). Study of webmetrics. *Information Science*, 20(1), 62–65.
- Xu, X., Li, J., & Ge, C. (2001). Research on netmetrology: status. Problems and development. *Library Tribune*, 21(6), 44–47.
- Yu, P. (1993). A diachronic study on the aging of scientific literature. *Intelligence Service Research*, 2, 122–126.

- Yu, P. (1997). The method of Burton-Kebler literature aging measurement and the negation of measure result. *Library and Information Service*, 10, 23–27.
- You, T. (1996). The overall design of computer citation system. *New Technology of Library and Information Service*, 12(1), 20–23.
- You, T. (1996). *The methods of Bibliometrics and computer-aided Bibliometrics research*. [Dissertation]. Wuhan: Wuhan University.
- Yu, G., & Daren, Yu. (1998). The delay effect of literature citation and the revision of the literature aging model. *Journal of the China Society for Scientific and Technical Information*, 1, 74–78.
- Zhang, B. (1986). Advances in quantitative research of Information Science in China. *Journal of the China Society for Scientific and Technical Information* (5).
- Zhang, X. (2000). Comparison on fit method of the Lotka's law. *Journal of the China Society for Scientific and Technical Information*, 19(4), 15–20.
- Zhang, Y. (2000). Retrieval of SCI database and evaluation of webofscience using www. *Library Development*, 5, 83–84.
- Zhang, Z. (1988). The 80/20 Rule and the three laws of Bibliometrics. *Technology & Market* (4), 22–25.
- Zhang, X. (2001). Extension of the Egghe's Formula: Relation between θ and β based on the generalized Lotka's Law. *Journal of the China Society for Scientific and Technical Information*, 20(5), 625–631.
- Zhang, Y., & Zhou, L. (2005). On the interactive relationship between information management and research evaluation. *Document, Information & Knowledge*, 2, 5–8.
- Zhou, J., Su, X., & Yuan, P. (2002). Research on citation analysis system based on theory of data warehouse. *Journal of The China Society for Scientific and Technical Information*, 21(3), 290–294.
- Zhu, X. (1994). A comparative study of the law of science and technology literature and the diachronic method. *Information Science*, 4, 21–25.
- Zipf, G. K. (1949). Human behavior and the principle of least effort.
- Zou, F. (2001). Application of the webmetrics in the digital libraries. *Document, Information & Knowledge*, 1(1), 16–17.
- Zhu, X. (1994). Discussion on the mathematical model of Citation Distribution. *Journal of the China Society for Scientific and Technical Information*, 6, 421–429.
- Zou, Z. (1996). *Introduction to information science*. Nanjing University Press.