SUBJECT CODE : CS8080

Strictly as per Revised Syllabus of

ANNA UNIVERSITY

Choice Based Credit System (CBCS)

Semester - VIII (CSE / IT) Professional Elective-V

INFORMATION RETRIEVAL TECHNIQUES

Iresh A. Dhotre

M.E. (Information Technology)

Ex-Faculty, Sinhgad College of Engineering,

Pune.

TECHNICAL PUBLICATIONS

SINCE 1993

An Up-Thrust for Knowledge

Information Retrieval Techniques

Subject Code : CS8080

Semester - VIII (Computer Science and Engineering / Information Technology) Professional Elective-V

Published by :

TECHNICAL PUBLICATIONS
SINCE 1993 An Up-Thrust for Knowledge

Amit Residency, Office No.1, 412, Shaniwar Peth,
Pune - 411030, M.S. INDIA, Ph.: +91-020-24495496/97
Email : sales@technicalpublications.org Website : www.technicalpublications.org

Printer :

Yogiraj Printers & Binders
Sr.No. 10/1A,
Ghule Industrial Estate, Nanded Village Road,
Tal. - Haveli, Dist. - Pune - 411041.

ISBN 978-93-90450-97-8

AU 17

PREFACE

The importance of **Information Retrieval Techniques** is well known in various engineering fields. Overwhelming response to my books on various subjects inspired me to write this book. The book is structured to cover the key aspects of the subject **Information Retrieval Techniques.**

The book uses plain, lucid language to explain fundamentals of this subject. The book provides logical method of explaining various complicated concepts and stepwise methods to explain the important topics. Each chapter is well supported with necessary illustrations, practical examples and solved problems. All the chapters in the book are arranged in a proper sequence that permits each topic to build upon earlier studies. All care has been taken to make students comfortable in understanding the basic concepts of the subject.

Representative questions have been added at the end of each chapter to help the students in picking important points from that chapter.

The book not only covers the entire scope of the subject but explains the philosophy of the subject. This makes the understanding of this subject more clear and makes it more interesting. The book will be very useful not only to the students but also to the subject teachers. The students have to omit nothing and possibly have to cover nothing more.

I wish to express my profound thanks to all those who helped in making this book a reality. Much needed moral support and encouragement is provided on numerous occasions by my whole family. I wish to thank the **Publisher** and the entire team of **Technical Publications** who have taken immense pain to get this book in time with quality printing.

Any suggestion for the improvement of the book will be acknowledged and well appreciated.

Author
D. A. Dhotre

Dedicated to God

SYLLABUS

Information Retrieval Techniques - CS8080

UNIT I INTRODUCTION

Information Retrieval - Early Developments - The IR Problem - The User's Task - Information versus Data Retrieval - The IR System - The Software Architecture of the IR System - The Retrieval and Ranking Processes - The Web - The e-Publishing Era - How the web changed Search - Practical Issues on the Web - How People Search - Search Interfaces Today - Visualization in Search Interfaces. **(Chapter - 1)**

UNIT II MODELING AND RETRIEVAL EVALUATION

Basic IR Models - Boolean Model - TF-IDF (Term Frequency/Inverse Document Frequency) Weighting - Vector Model - Probabilistic Model - Latent Semantic Indexing Model - Neural Network Model - Retrieval Evaluation - Retrieval Metrics - Precision and Recall - Reference Collection - User-based Evaluation - Relevance Feedback and Query Expansion - Explicit Relevance Feedback. **(Chapter - 2)**

UNIT III TEXT CLASSIFICATION AND CLUSTERING

A Characterization of Text Classification - Unsupervised Algorithms : Clustering - Naïve Text Classification - Supervised Algorithms - Decision Tree - k-NN Classifier - SVM Classifier -Feature Selection or Dimensionality Reduction - Evaluation metrics - Accuracy and Error - Organizing the classes - Indexing and Searching - Inverted Indexes - Sequential Searching - Multi-dimensional Indexing. **(Chapter - 3)**

UNIT IV WEB RETRIEVAL AND WEB CRAWLING

The Web - Search Engine Architectures - Cluster based Architecture - Distributed Architectures - Search Engine Ranking - Link based Ranking - Simple Ranking Functions - Learning to Rank - Evaluations - Search Engine Ranking - Search Engine User Interaction - Browsing - Applications of a Web Crawler - Taxonomy - Architecture and Implementation - Scheduling Algorithms - Evaluation. **(Chapter - 4)**

UNIT V RECOMMENDER SYSTEM

Recommender Systems Functions - Data and Knowledge Sources - Recommendation Techniques - Basics of Content-based Recommender Systems - High Level Architecture - Advantages and Drawbacks of Content-based Filtering - Collaborative Filtering - Matrix factorization models - Neighborhood models. **(Chapter - 5)**

TABLE OF CONTENTS

UNIT - I

UNIT - II

UNIT - III

UNIT - IV

UNIT - V

UNIT - I

1 Introduction

Syllabus

Information Retrieval - Early Developments - The IR Problem - The User's Task - Information versus Data Retrieval - The IR System - The Software Architecture of the IR System - The Retrieval and Ranking Processes - The Web - The e-Publishing Era - How the web changed Search - Practical Issues on the Web - How People Search - Search Interfaces Today - Visualization in Search Interfaces.

Contents

1.1 Introduction of Information Retrieval

AU : May-17

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections.
- The role of an IR system is to retrieve all the documents, which are relevant to a query while retrieving as few non - relevant documents as possible. IR allows access to whole documents, whereas, search engines do not.
- There is a huge quantity of text, audio, video and other documents available on the Internet, on about any subject. Users need to be able to find relevant information to satisfy their particular information needs.
- There are two ways of searching for information : to use a search engine or to browse directories organized by categories. There is still a large part of the Internet that is not accessible (for example private databases and intranets).
- Information retrieval is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure.
- In IR systems, the information is not structured. It is contained in free form in text (webpages or other documents) or in multimedia content. The first IR systems implemented in 1970's were designed to work with small collections of text. Some of these techniques are now used in search engines.
- The information retrieval techniques focusing on the challenges faced by search engine. One particular challenge is the large scale, given by the huge number of webpages available on the Internet.
- Another challenge is inherent to any information retrieval system that deals with text : the ambiguity of the natural language (English or other languages) that makes it difficult to have perfect matches between documents and user queries.
- Information retrieval is never an easy task. The problem with IR is that document representation, either by index terms or texts cannot satisfy user need representation, which is dynamic and complicated.
- Moreover, traditional IR systems are designed to support only one type of information-seeking strategy that users engage in query formulation.

1.1.1 Early Developments

- Information Retrieval (IR) is about the process of providing answers to client's information needs. It is thus concerned with the collection, representation, storage, organization, accessing, manipulation and display, of the information items necessary to satisfying those needs.

- Definition : Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections.
- There is a huge quantity of text, audio, video and other documents available on the Internet, on about any subject. Users need to be able to find relevant information to satisfy their particular information needs.
- There are two ways of searching for information : To use a search engines or to browse directories organized by categories.
- IR is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure.
- In IR systems, the information is not structured. It is contained in free form in text (web pages or other documents) or in multimedia content.
- The first IR systems implemented in 1970's were designed to work with small collections of text. Some of these techniques are now used in search engines.
- The information retrieval techniques focusing on the challenges faced by search engine.
  1. One particular challenge is the large scale, given by the huge number of web-pages available on the Internet.
  2. The ambiguity of the natural language (English or other languages) that makes it difficult to have perfect matches between documents and user queries.
- Information retrieval is never an easy task. The problem with IR is that document representation, either by index terms or texts cannot satisfy user need representation, which is dynamic and complicated.
- Traditional IR systems are designed to support only one type of information-seeking strategy that users engage in : Query formulation.

University Question

1. Appraise the history of information retrieval. **AU : May-17, Marks 8**

1.2 The IR Problem

- Information retrieval is about retrieving information relevant to the user on the basis of a query. Early IR systems were boolean systems which allowed users to specify their information need using a complex combination of boolean ANDs, ORs and NOTs.

- In modern IR system, users need vast information for search engine. User looking for the link to the homepage of a government, company and colleges. They also looking for information required to execute tasks associated with their jobs or immediate needs.
- Sometime user types full description of query to IR system. To solve this query by search engine is not possible. Here user might want to first translate this information need into a query, to be posed to the system.
- Given the user query, the goal of the IR system is to retrieve information that is useful or relevant to the user.
- The key issues with IR models are selection of search vocabulary, search strategy formulations and information overload

1.2.1 The User's Task

- The user of a retrieval system has to translate his information need into a query in the language provided by the system. With an information retrieval system, this normally implies specifying a set of words which convey the semantics of the information need.
- With a data retrieval system, a query expression is used to convey the constraints that must be satisfied by objects in the answer set. In both cases, we say that the user searches for useful information executing a retrieval task. Fig. 1.2.1 shows Interaction of the user with the retrieval system.

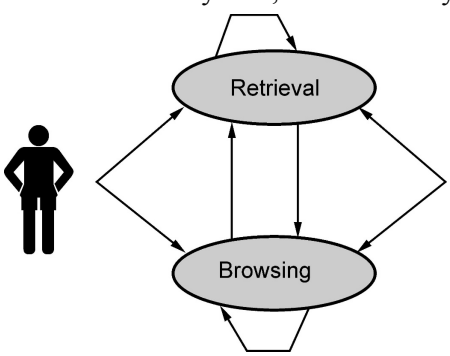


Fig. 1.2.1 : Interaction of the user with the retrieval system

- Suppose the user may be interested in web site about healthcare product. In this situation, the user might use an interactive interface to simply look around in the collection for documents related to healthcare product.
- User may be interested in new beauty product, weight loss or gain product. Here user is browsing the documents in the collection, not searching. It is still a process of retrieving information, but one whose main objectives are not clearly defined in the beginning and whose purpose might change during the interaction with the system.
- Pull technology : User requests information in an interactive manner. It perform three retrieval tasks, i.e. Browsing (hypertext), Retrieval (classical IR systems) and Browsing and retrieval (modern digital libraries and web systems).
- Push technology : Automatic and permanent pushing of information to user. It acts like a software agents.

1.3 Information versus Data Retrieval

- An information retrieval system is software that has the features and functions required to manipulate "information" items versus a DBMS that is optimized to handle "structured" data.
- Information retrieval and Data Retrieval (DR) are often viewed as two mutually exclusive means to perform different tasks, IR being used for finding relevant documents among a collection of unstructured/semi-structured documents.
- Data retrieval being used for finding exact matches using stringent queries on structured data, often in a Relational Database Management System (RDBMS).
- IR is used for assessing human interests, i.e., IR selects and ranks documents based on the likelihood of relevance to the user's needs. DR is different; answers to users' queries are exact matches which do not impose any ranking.
- Data retrieval involves the selection of a fixed set of data based on a well-defined query. Information retrieval involves the retrieval of documents of natural language.
- IR systems do not support transactional updates whereas database systems support structured data, with schemas that define the data organization. IR systems deal with some querying issues not generally addressed by database systems and approximate searching by keywords.

1.3.1 Difference between Data Retrieval and Information Retrieval

Parameters	Data retrieval	Information retrieval
Example	Data base query	WWW search
Matching	Exact	Partial match Best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive
Classification	Monotonic	Polytechnic

1.4 The IR System

AU : Dec.-16, 17

- An information retrieval system is an information system, which is used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations.

- Information retrieval is the process of searching some collection of documents, in order to identify those documents which deal with a particular subject. Any system that is designed to facilitate this literature searching may legitimately be called an information retrieval system.
- Conceptually, IR is the study of finding needed information. It helps users to find information that matches their information needs. Historically, IR is about document retrieval, emphasizing document as the basic unit.
- Information retrieval locates relevant documents, on the basis of user input such as keywords or example documents, for example : Find documents containing the words "database systems".
- Fig. 1.4.1 shows information retrieval system block diagram. It consists of three components : **Input, processor and output**.

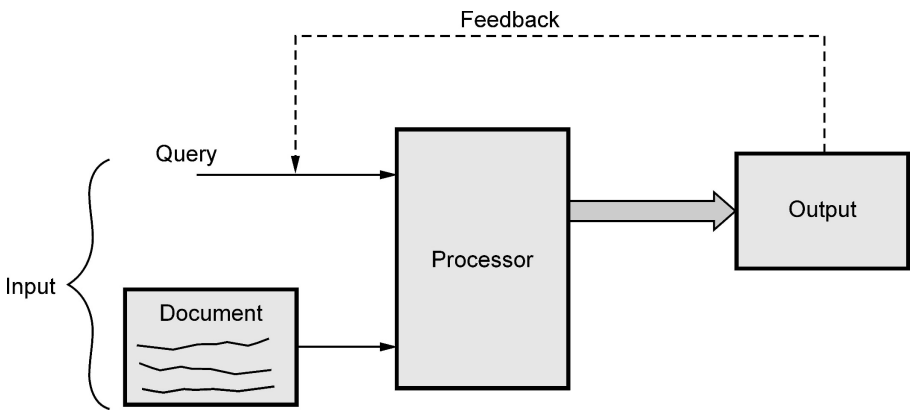


Fig. 1.4.1 : IR block diagram

a) **Input :** Store only a representation of the document or query which means that the text of a document is lost once it has been processed for the purpose of generating its representation.

b) A **document representative** could be a list of extracted words considered to be significant.

c) **Processor :** Involve in performing actual retrieval function, executing the search strategy in response to a query.

d) **Feedback :** Improving the subsequent run after sample retrieval.

e) **Output :** A set of document numbers.

- Information retrieval locates relevant documents, on the basis of user input such as keywords or example documents.

- The computer-based retrieval systems store only a representation of the document or query which means that the text of a document is lost once it has been processed for the purpose of generating its representation.
- The process may involve structuring the information, such as classifying it. It will also involve performing the actual retrieval function that is executing the search strategy in response to a query.
- Text document is the output of information retrieval system. Web search engines are the most familiar example of IR systems.

1.4.1 Process of Information Retrieval

- Information retrieval is often a continuous process during which you will consider, reconsider and refine your research problem, use various different information resources, information retrieval techniques and library services and evaluate the information you find.
- Fig. 1.4.2 shows that the stages follow each other during the process, but in reality they are often active simultaneously and you usually will repeat some stages during the same information retrieval process.

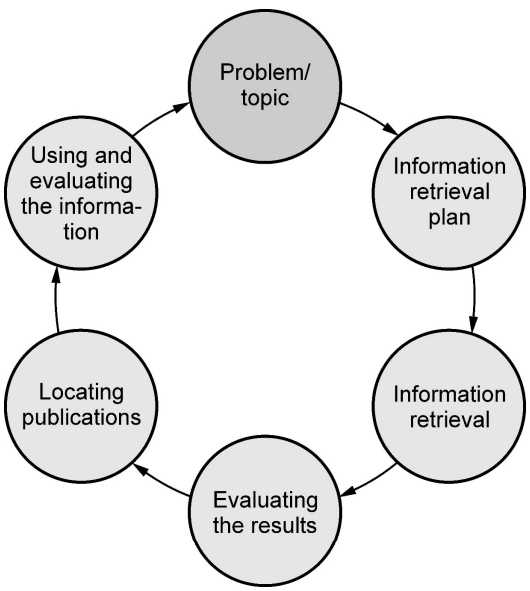


Fig. 1.4.2 Stages of IR process

- The different stages of the information retrieval process are :
  1. Problem / Topic : An information need occurs when more information is required to solve a problem
  2. Information retrieval plan : Define your information need and choose your information resources, retrieval techniques and search terms

3. Information retrieval : Perform your planned information retrieval (information retrieval techniques)
4. Evaluating the results : Evaluate the results of your information retrieval (number and relevance of search results)
5. Locating publications : Find out where and how the required publication, e.g. article, can be acquired
6. Using and evaluating the information : Evaluate the final results of the process (critical and ethical evaluation of the information and information resources)

1.4.2 The Software Architecture of the IR System

- Fig. 1.4.3 shows architecture of IR system.

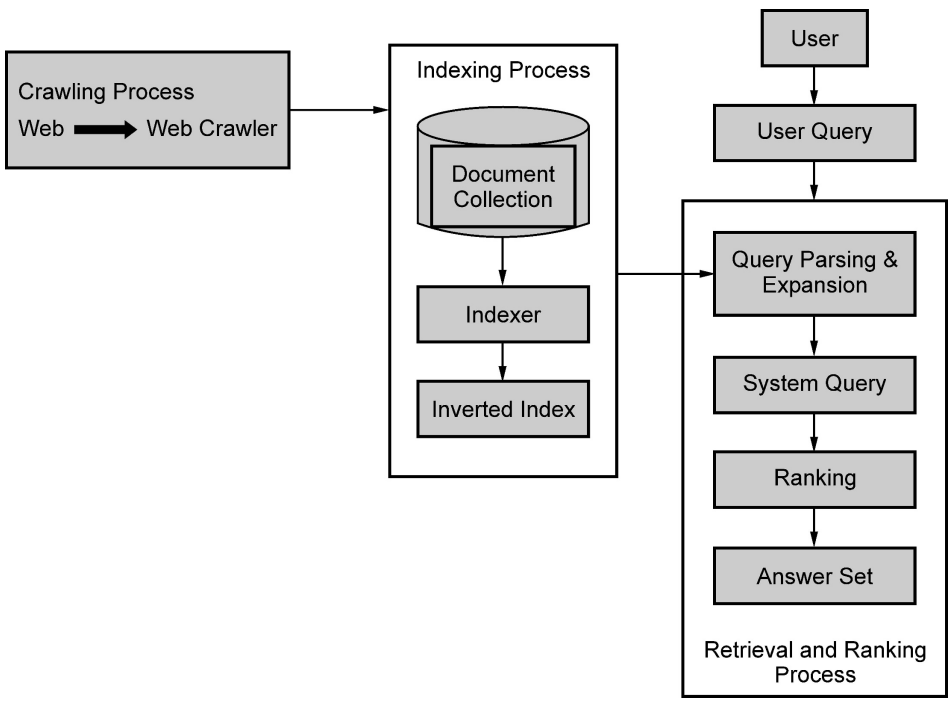


Fig. 1.4.3 : Architecture of IR system

- The user's query is processed by a search engine, which may be running on the user's local machine, on a large cluster of machines in a remote geographic location, or anywhere in between.
- A major task of a search engine is to maintain and manipulate an inverted index for a document collection. This index forms the principal data structure used by the engine for searching and relevance ranking.

- As its basic function, an inverted index provides a mapping between terms and the locations in the collection in which they occur.
- To support relevance ranking algorithms, the search engine maintains collection statistics associated with the index, such as the number of documents containing each term and the length of each document.
- In addition the search engine usually has access to the original content of the documents in order to report meaningful results back to the user.
- Using the inverted index, collection statistics, and other data, the search engine accepts queries from its users, processes these queries, and returns ranked lists of results.
- To perform relevance ranking, the search engine computes a score, sometimes called a Retrieval Status Value (RSV), for each document. After sorting documents according to their scores, the result list must be subjected to further processing, such as the removal of duplicate or redundant results.
- For example, a web search engine might report only one or results from a single host or domain, eliminating the others in favor of pages from different sources.

1.4.3 The Retrieval and Ranking Processes

- A good retrieval model will find documents that are likely to be considered relevant by the person who submitted the query. Some retrieval models focus on topical relevance, but a search engine deployed in a real environment must use ranking algorithms that incorporates user relevance.
- Relevancy ranking is the method that is used to order the results list in such a way that the records most likely to be of interest to a user will be at the front. This makes searching easier for users as they will not have to spend as much time looking through records for the information that interests them.
- Each relevancy ranking algorithm slightly biases one type of data over another. While most any of the relevancy ranking algorithms will make a large difference, it is sometimes worthwhile trying several of the ranking methods. This way, you will be able to find the algorithm which most closely reflects the needs of your application as well as you and your user's expectations.
- There are a number of ways of calculating how a given record ranks and the factors that are taken into consideration vary with each technique.

  a) The number of times the search term occurs within a given record.

  b) The number of times the search term occurs across the collection of records.

c) The number of words within a record.

d) The frequencies of words within a record.

e) The number of records in the index.

- Typically, relevancy ranking algorithms rank records in relation to each other. The weight assigned to a given record is a weight that reflects the weight of the record in relation to other records within the same database and for the same query.

University Questions

1. Explain in detail about the components of IR. **AU : Dec.-16, Marks 16**
2. Explain the issues in the process of information Retrieval. **AU : Dec.-17, Marks 8**
3. Explain in detail, the components of Information Retrieval and Search Engine. **AU : Dec.-17, Marks 16**

1.5 The Web

- World wide web is collection of millions of files stored on thousands of servers all over the world. These files represent documents, pictures, video, sounds, programs, interactive environments.
- A web page is an HTML document that is stored on a web server. A web site is a collection of web pages belonging to a particular organization.
- URL of these pages share a common prefix, which is the address of the home page of the size. Search engines are a bottom-up approach for finding your way around the web. Some search engines search only the titles of web pages. While other search every word. Keywords can be combined with Boolean operations, such as AND, OR and NOT, to produce rather complicated queries.

1.5.1 The e-Publishing Era

- E-publishing refers to a publishing process where the manuscript are submitted in E-format, edited, printed and even distributed to users in E-form by computer and communication technology, which may be online, CD-ROM, Networks etc. It involves the storage of information in electronic or digital form. It also refers to a type of publishing that does not include printed books.
- E-publishing has been defining as any non-print media material that is published in digitized form to an identifiable public. The media in electronic publishing can be text, numeric, graphic, still or motion pictures, video, sound or as infrequently the case a combination of any or all of these.

- There are four main reasons for the development of e-publishing,
  a) Rapid development and wide use of computer technology.
  b) The tremendous growth of computer networks.
  c) Merging of computer and telecommunication technology.
  d) Development of information industry.

1.5.2 How the Web Changed Search

- The web has introduced millions of people to search. The information retrieval community stands ready to suggest helpful strategies for finding information on the Web.
- Let us consider the impact of web on search engine:
  1. Characteristics of the document collection itself
  2. Size of the collection and volume of user queries
  3. Vast size of the document collection
  4. Web adverting
- Search has changed dramatically over the past year and semantic technology has been at the centre of it all. Consumers increasingly expect search engines to understand natural language and perceive the intent behind the words they type in, and search engine algorithms are rising to this challenge.

1.6 How People Search

1.6.1 Information Lookup Versus Exploratory Search

- Search activities are commonly divided into two broad categories: lookup and exploratory. Exploratory search is an increasingly important activity yet challenging for users.
- Lookup search is by far the better understood and assumed to have precise search goals. The predominant design goal in information retrieval systems has been fast and accurate completion of lookup searches.
- Exploratory search is presently thought to center around the acquisition of new knowledge and considered to be challenging for the user.
- Lookup is the most basic kind of search task and has been the focus of development for database management systems and much of what Web search engines support.
- Lookup tasks return discrete and well-structured objects such as numbers, names, short statements, or specific files of text or other media.

- Database management systems support fast and accurate data lookups in business and industry; in journalism, lookups are related to questions of who, when, and where as opposed to what, how, and why questions.
- In libraries, lookups have been called "known item" searches to distinguish them from subject or topical searches.
- A typical example would be a user wanting to make a reservation to a restaurant and looking for the phone number on the Web.
- On the other hand, exploratory search is described as open-ended, with an unclear information need, an ill-structured problem of search with multiple targets. This search activity is evolving and can occur over time.
- For example, a user wants to know more about Senegal, she doesn't really know what kind of information she wants or what she will discover in this searchsession; she only knows she wants to learn more about that topic.

1.7 Search Interfaces Today

- The job of the search user interface is to aid users in the expression of their information needs, in the formulation of their queries, in the understanding of their search results, and in keeping track of the progress of their information seeking efforts.
- The typical search interface today is of the form : type-keywords-in-entry-form, view-results-in-a-vertical-list.
- Some important reasons for the relative simplicity and unchanging nature of the standard Web search interface are :
  a) Search is a means towards some other end, rather than a goal in itself. When a person is looking for information, they are usually engaged in some larger task, and do not want their flow of thought interrupted by an intrusive interface.
  b) Search is a mentally intensive task. When a person reads text, they are focused on that task; it is not possible to read and to think about something else at the same time. Thus, the fewer distractions while reading, the more usable the interface.
  c) Since nearly everyone who uses the Web uses search, the interface design must be understandable and appealing to a wide variety of users of all ages, cultures and backgrounds, applied to an enormous variety of information needs.

1.7.1 Query Specification

- The query specification process is :
  1. The kind of information the searcher supplies. Query specification input spans a spectrum from full natural language sentences, to keywords and key phrases, to syntax-heavy command language-based queries.
  2. The interface mechanism the user interacts with to supply this information. These include command line interfaces, graphical entry form-based interfaces, and interfaces for navigating links.
- Queries over collections of textual information usually take on a textual form. Keyword queries consist of a list of one or more words or phrases -- rather than full natural language statements.
- Example : English keyword queries include flip cam, fresh chilli paste recipes, and video game addiction. Some keyword queries consist of lists of different words and phrases, which together suggest a topic.
- Many others are noun compounds and proper nouns. Less frequently, keyword queries contain syntactic fragments including prepositions and verbs and in some cases, full syntactic phrases.
- Dynamic query term suggestions can be provided as the user types in a term before they view the results or it can be presented following the result display stage.
- It is interesting to note that the performance of query term suggestions across the three search engines is varied in terms of the number of suggestions and how they handle single word and multi-word queries. Following table provides a comparative overview of the number of suggested query terms for TREC topics.

Search engine	Average number of words suggested	Median number of words suggested
Google	4	4
Yahoo!	6.46	10
Bing	6.18	8

- All the three search engines offer spelling error correction features and around 80% of the time they provide 4 or more dynamic query suggestions.

1.7.2 Retrieval Result Display

- When displaying search results, either the documents must be shown in full or else the searcher must be presented with some kind of representation of the content of those documents.
- The documents surrogate refers to the information that summarizes the document.
- The appearance of search engine results pages is constantly in flux due to experiments conducted by Google, Bing, and other search engine providers to offer their users a more intuitive, responsive experience.
- The quality of the surrogate can greatly effect the perceived relevance of the search results listing. In Web search, the page title is usually shown prominently along with the URL and sometimes other metadata.
- The user enters their search query, upon which the search engine presents them with a SERP. Every SERP is unique, even for search queries performed on the same search engine using the same keywords or search queries.
- This is because virtually all search engines customize the experience for their users by presenting results based on a wide range of factors beyond their search terms, such as the user's physical location, browsing history, and social settings. Two SERPs may appear identical, and contain many of the same results, but will often feature subtle differences.
- A deep link is a hypertext link to a page on a website other than its homepage. Deep links are often used to link directly to products of an online store to or appropriate content.
- Google itself uses deep links in the form of rich snippets or sitelinks. A hyperlink that points to a deeper level of a domain can also be useful for link hubs, lists of topics or in citations. Again, the user's interest is in the foreground.
- Price comparison portals also work with deep links. In this case, this type of link is necessary because the potential customer would want to find and buy the exact product he is comparing.

1.7.3 Query Reformulation

- After a query is specified and results have been produced, a number of tools exist to help the user reformulate their query.
- Query formulation is an essential part of successful information retrieval. The challenges in formulating effective queries are emphasized in web information search, because the web is used by a diverse population varying in their levels of expertise.

- Query formulation is the stage of the interactive information access process in which user translates an information need into a query and submits the query to an information access system such as a search engine.
- The system performs some computation to match the query with the documents most likely to be relevant to the query and returns a ranked list of relevant documents to the user.

1.8 Visualization in Search Interfaces

- Various method is used in search engine for visualization concept.
  1. Visualizing Boolean syntax
  2. Visualizing query terms within retrieval results
  3. Visualizing relationships among words and documents
  4. Visualization for text mining
- **Visualizing Boolean syntax :** Boolean query is rarely used in web search because of its difficult syntax**.** Venn diagram is better method than Boolean search for representing query. Problem with Boolean queries is that they can easily end up with empty results or too many results.
- **Visualizing query terms within retrieval results :** In standard search result listing, summary sentences are often selected that contains query terms and occurrence of these terms are highlighting or boldfaced where they appear in the title, summary and URL. Fig. 1.8.1 shows visualization in query.

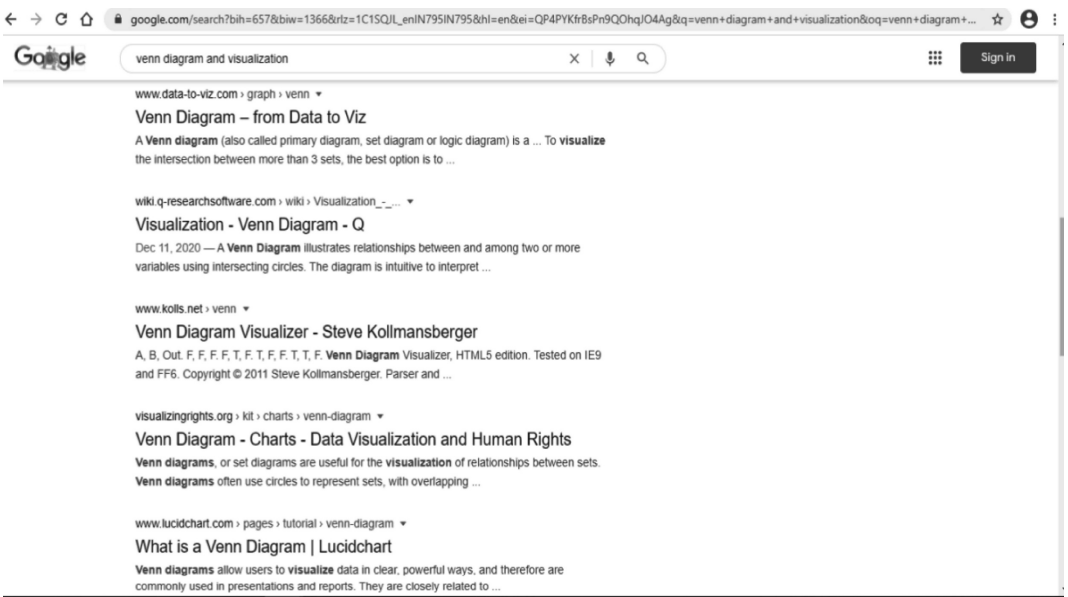


Fig. 1.8.1 : Visualization query

- **Visualizing relationships among words and documents :** Visualization developers suggest various idea of placing words and documents on a two-dimensional canvas, where proximity of glyphs represents semantic relationship among the terms or documents. Another method is to map documents or words from a very high-dimensional term space down into a two-dimensional plane and show where the documents or words fall within that plane using 2D or 3D.
- **Visualization for text mining :** Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories. Text mining can be visualized as consisting of two phases: Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.

1.9 Part A : Short Answered Questions [2 Marks Each]

Q.1 Define information retrieval. **AU : Dec.-16**

▣ **Ans. :** Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Q.2 Explain difference between data retrieval and information retrieval.

▣ **Ans. :**

Parameters	Data Retrieval	Information Retrieval
Example	Data Base Query	WWW Search
Matching	Exact	Partial Match, Best Match
Interference	Deduction	Induction
Model	Deterministic	Probabilistic

Q.3 List and explain components of IR block diagram.

▣ **Ans. :**

- **Input :** Store only a representation of the document or query which means that the text of a document is lost once it has been processed for the purpose of generating its representation.
- A **document representative** could be a list of extracted words considered to be significant.
- **Processor :** Involve in performing actual retrieval function, executing the search strategy in response to query.

- **Feedback :** Improving the subsequent run after sample retrieval.
- **Output :** A set of document numbers.

Q.4 What is objective term and nonobjective term ?

▣ **Ans. :**

- **Objective terms** are extrinsic to semantic content, and there is generally no disagreement about how to assign them. Examples include author name, document URL, and date of publication.
- **Nonobjective terms** are intended to reflect the information manifested in the document, and there is no agreement about the choice or degree of applicability of these terms. They are also known as content terms.

Q.5 Explain the type of natural language technology used in information retrieval.

▣ **Ans. :** Two types of natural language technology can be useful in information retrieval :

- Natural language interfaces make the task of communicating with the information source easier, allowing a system to respond to a range of inputs, possibly from inexperienced users, and to produce more customized output.
- Natural language text processing allows a system to scan the source texts, either to retrieve particular information or to derive knowledge structures that may be used in accessing information from the texts.

Q.6 What is search engine ?

▣ **Ans. :** A search engine is a document retrieval system designed to help find information stored in a computer system, such as on the WWW. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria.

Q.7 What is conflation ?

▣ **Ans. :** Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form. The process of stemming is often called **conflation**.

Q.8 What is an invisible web ?

▣ **Ans. :** Many dynamically generated sites are not indexable by search engines; this phenomenon is known as the invisible web.

Q.9 Define Zipf's law.

▣ **Ans. :** An empirical rule that describes the frequency of the text words. It states that the i^{th} most frequent word appears as many times as the most frequent one divided by i^{θ}, for some $\theta > 1$.

Q.10 What is supervised learning ?

▣ **Ans. :** In supervised learning, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network

Q.11 What is unsupervised learning ?

▣ **Ans. :** In an unsupervised learning, the network adapts purely in response to its inputs. Such networks can learn to pick out structure in their input.

Q.12 What is text mining ?

▣ **Ans. :** Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories. Text mining can be visualized as consisting of two phases : Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.

Q.13 Specify the role of an IR system. **AU : Dec.-16**

▣ **Ans. :** The role of an IR system is to retrieve all the documents, which are relevant to a query while retrieving as few non - relevant documents as possible. IR allows access to whole documents, whereas, search engines do not.

Q.14 Outline the impact of the web on information retrieval. **AU : May-17**

▣ **Ans. :** Web is a huge, widely-distributed, highly heterogeneous and semi-structured information. The rapid growth of the Internet, huge information is available on the Web and Web information retrieval presents additional technical challenges when compared to classic information retrieval due to the heterogeneity and size of the web.

- Web information retrieval is unique due to the dynamism, variety of languages used, duplication, high linkage, ill formed query and wide variance in the nature of users. IR helps users find information that matches their information needs expressed as queries. Historically, IR is about document retrieval, emphasizing document as the basic unit.

Q.15 Compare information retrieval and web search. **AU : May-17**

▣ **Ans. :** In information retrieval, databases usually cover only one language or indexing of documents written in different languages with the same vocabulary. In web search, documents are in many different languages. Usually search engines use full text indexing; no additional subject analysis.

1.10 Multiple Choice Questions

Q.1 _____ retrieval deals with the representation, storage, organization of and access to information items such as documents, web pages, online catalogs, structured and semi-structured records and multimedia objects.

[a] Data [b] Audio [c] Video [d] Information

Q.2 _____ of a retrieval system has to translate his information need into a query in the language provided by the system.

[a] The manager [b] The user

[c] The designer [d] The administrator

Q.3 _____ is usually provided by most modern information retrieval systems.

[a] Information and knowledge retrieval

[b] Information or knowledge retrieval

[c] Information and data retrieval

[d] Information or data retrieval

Q.4 _____ is an iterative process of formulating a conceptual from a large collection of information.

[a] Sense making [b] Data collection

[c] Information collection [d] All of these

Q.5 A web page is an _____ document that is stored on a web server.

[a] XML [b] HTML [c] XSL [d] Java

Q.6 A _____ is a hypertext link to a page on a website other than its homepage.

[a] hyper link [b] deep link [c] URL [d] HTML

Q.7 Which methods are used in search engine for visualization concept ?

[a] Visualizing Boolean syntax

[b] Visualizing query terms within retrieval results

[c] Visualizing relationships among words and documents

[d] All of these

Q.8 URL stands for _____.

[a] Uniform Ravar Location [b] Uniform Resource Locator

[c] Uni Resource Locate [d] Uniform Reverse Locator

Q.9 Which of the following is a search engine ?

[a] Google [b] Yahoo! [c] Bing [d] All of these

Q.10 In IR systems, the information is _____.

[a] Structured [b] semi-structured [c] not structured [d] None

Q.11 A search engine is a program to search _____.

[a] for information [b] web pages for information using specified search terms

[c] web pages [d] web pages for specified index terms

Q.12 Early IR systems were _____ systems which allowed users to specify their information need using a complex combination of Boolean ANDs, ORs and NOTs.

[a] Boolean [b] vector [c] logical [d] All of these

Q.13 Information Retrieval Systems is characterized by _____ data format.

[(a)] structured [b] semi-structured [c] unstructured [d] all of these

Q.14 _____ are a set of electronic resources and associated technical capabilities for creating, searching, and using information.

[a] Analog libraries [b] Digital libraries

[c] Digital information [d] Digital data

Q.15 Which of the following is NOT components of IR block diagram ?

[a] Input [b] Processor [c] Feedback [d] Information

Q.16 Web browser is a software program that interprets and displays the contents of _____ web pages.

[a] XML [b] HTML [c] static [d] dynamic

Q.17 _____ diagram is a diagram that visually displays all the possible logical relationships between collections of sets.

[a] Text [b] Information [c] Binary [d] Venn

➤ Answer Keys for Multiple Choice Questions

Q.1	d	**Q.2**	b	**Q.3**	c	**Q.4**	a
Q.5	b	**Q.6**	b	**Q.7**	d	**Q.8**	b
Q.9	d	**Q.10**	c	**Q.11**	b	**Q.12**	a
Q.13	c	**Q.14**	b	**Q.15**	d	**Q.16**	b
Q.17	d						

❑❑❑

Notes

UNIT - II

2 Modeling and Retrieval Evaluation

Syllabus

Basic IR Models - Boolean Model - TF-IDF (Term Frequency/Inverse Document Frequency) Weighting - Vector Model - Probabilistic Model - Latent Semantic Indexing Model - Neural Network Model - Retrieval Evaluation - Retrieval Metrics - Precision and Recall - Reference Collection - User-based Evaluation - Relevance Feedback and Query Expansion - Explicit Relevance Feedback.

Contents

2.1 Basic IR Models

AU : May-17

- An information retrieval system is a software program that stores and man-ages information on documents, often textual documents but possibly multimedia.
- Reason behind use of models in information retrieval is that models guide research and provide the means for academic discussion. Models can also serve as a blueprint to implement an actual retrieval system.
- A model of information retrieval predicts and explains what a user will find relevant given the user query. The correctness of the model's predictions can be tested in a controlled experiment.

2.1.1 Basic Concept

- Each document represented by a set of representative keywords or index terms. An index term is a document word useful for remembering the document main themes.
- Usually, index terms are nouns because nouns have meaning by themselves. However, search engines assume that all words are index terms.
- Not all terms are equally useful for representing the document contents: less frequent terms allow identifying a narrower set of documents. The importance of the index terms is represented by weights associated to them.
- Model is an idealization or abstraction of an actual process. Mathematical models are used to study the properties of the process, draw conclusions, and make predictions. Conclusions derived from a model depend on whether the model is a good approximation of the actual situation. Statistical models represent repetitive processes; make predictions about frequencies of interesting events.

2.1.2 Boolean Model

- The Boolean model is the first model of information retrieval and probably also the most criticized model. It is based on set theory and Boolean algebra.
- It is based on a binary decision criterion without any notion of a grading scale. Boolean expressions have precise semantics. It is not simple to translate an information need into a Boolean expression. It can be represented as a disjunction of conjunction vectors(in disjunctive normal form-DNF).

D : Set of words (indexing terms) present in a document. Each term is either present (1) or absent (0).

Q : A Boolean expression. The terms are index terms and operators are AND, OR, and NOT.

F : Boolean algebra over sets of terms and sets of documents.

R : A document is predicted as relevant to a query expression if and only if it satisfies the query expression

((text ∨ information) ∧ retrieval ∧¬ theory)

- Each query term specifies a set of documents containing the term :

  a. AND (∧) : The intersection of two sets b. OR (∨) : The union of two sets

  c. NOT (¬) : Set inverse, or really set difference.

➤ **Boolean Relevance example :**

((text ∨ information) ∧ retrieval ∧¬ theory)

It gives following list :

"Information Retrieval"

"Information Theory"

"Modern Information Retrieval: Theory and Practice"

"Text Compression"

➤ **Implementing the Boolean Model :**

- First, consider purely conjunctive queries ($t_a \wedge t_b \wedge t_c$).
- It only satisfied by a document containing all three terms.
- If $D(t_a) = \{d \mid t_a \in d\}$, then the maximum possible size of the retrieved set is the size of the smallest $D(t_a)$.
- $|D(t_a)|$ is the length of the inverted list for t_a.
- For instance, the query **social AND economic** will produce the set of documents that are indexed both with the term social and the term economic, i.e. the intersection of both sets.
- Combining terms with the OR operator will define a document set that is bigger than or equal to the document sets of any of the single terms.
- So, the query social OR political will produce the set of documents that are indexed with either the term social or the term political or both, i.e. the union of both sets.
- This is visualized in the Venn diagrams of Fig. 2.1.1 in which each set of documents is visualized by a disc.

- The intersections of these discs and their complements divide the document collection into 8 non-overlapping regions, the unions of which give 256 different Boolean combinations of "social, political and economic documents". In Fig. 2.1.1 the retrieved sets are visualized by the shaded areas.

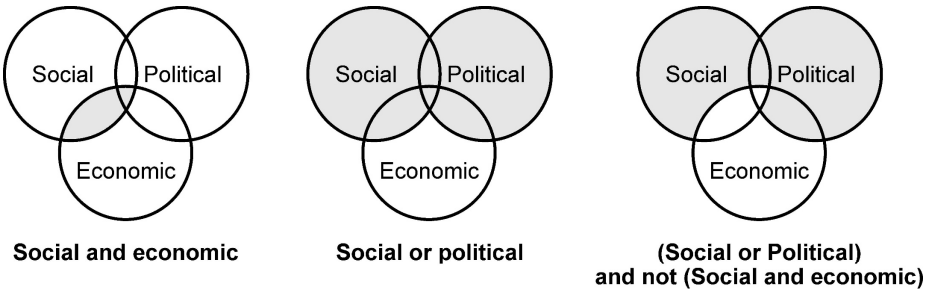


Fig. 2.1.1 : Boolean combinations of sets visualized as Venn diagrams

➤ **Advantages :**

a. Clean formalism

b. Simplicity

c. It is very precise in nature. The user exactly gets what is specified.

d. Boolean model is still widely used in small scale searches like searching emails, files from local hard drives or in a mid-sized library.

➤ **Disadvantages :**

a. It is not simple to translate an information need into a Boolean expression

b. Exact matching may lead to retrieval of too few or too many documents.

c. The retrieved documents are not ranked.

d. The model does not use term weights.

➥ 2.1.3 Vector Model

- Assign non-binary weights to index terms in queries and in documents. Compute the similarity between documents and query. The index terms in the query are also weighted

$$\bar{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

$$\bar{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

- Term weights are used to compute the degree of similarity between documents and the user query. Then, retrieved documents are sorted in decreasing order.

- They considered the index representations and the query as vectors embedded in a high dimensional Euclidean space, where each term is assigned a separate dimension. The similarity measure is usually the cosine of the angle that separates the two vectors $\bar{d}$ and $\bar{q}$.
- The cosine of an angle is 0 if the vectors are orthogonal in the multidimensional space and 1 if the angle is 0 degrees. The cosine formula is given by :

$$\text{score}\left(\bar{d}, \bar{q}\right) = \frac{\sum_{k=1}^{m} d_k \cdot q_k}{\sqrt{\sum_{k=1}^{m} (d_k)^2} \cdot \sqrt{\sum_{k=1}^{m} (q_k)^2}}$$

- The metaphor of angles between vectors in a multidimensional space makes it easy to explain the implications of the model to non-experts. Up to three dimensions, one can easily visualise the document and query vectors. Fig. 2.1.2 shows a query and document representation in the vector space model.

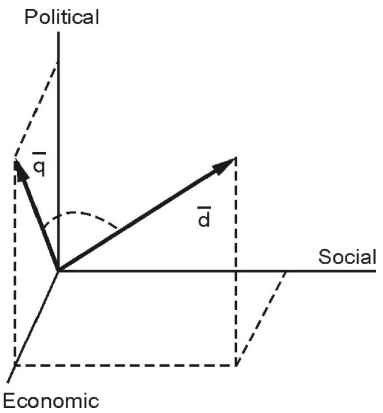


Fig. 2.1.2 : Query and document representation in the vector space model

- Measuring the cosine of the angle between vectors is equivalent with normalizing the vectors to unit length and taking the vector inner product. If index representations and queries are properly normalised, then the vector product measure of equation 1 does have a strong theoretical motivation. The formula then becomes :

$$\text{score}\left(\bar{d}, \bar{q}\right) = \sum_{k=1}^{m} n(d_k) \cdot n(q_k)$$

$$\text{where } n(v_k) = \frac{v_k}{\sqrt{\sum_{k=1}^{m} (v_k)^2}}$$

- We think of the documents as a collection C of objects and think of the user query as a specification of a set A of objects. In this scenario, the IR problem can be reduced to the

problem of determine which documents are in the set A and which ones are not (i.e. the IR problem can be viewed as a clustering problem).

1. **Intra-cluster :** One needs to determine what are the features which better describe the objects in the set A.
2. **Inter-cluster :** One needs to determine what are the features which better distinguish the objects in the set A.

- **t_f :** Inter-clustering similarity is quantified by measuring the raw frequency of a term k_i inside a document d_j , such term frequency is usually referred to as the t_f factor and provides one measure of how well that term describes the document contents.
- **idf :** Inter-clustering similarity is quantified by measuring the inverse of the frequency of a term k_i among the documents in the collection. This frequency is often referred to as the **inverse document frequency**.

➤ **Advantages :**

1. Its term-weighting scheme improves retrieval performance.
2. Its partial matching strategy allows retrieval of documents that approximate the query conditions.
3. Its cosine ranking formula sorts the documents according to their degree of similarity to the query.

➤ **Disadvantages :**

1. The assumption of mutual independence between index terms.
2. It cannot denote the "clear logic view" like Boolean model.

University Question

1. Explain vector space retrieval model with an example.	**AU : May-17, Marks 10**

2.2 Term Weighting

AU : Dec.-16

- A computerized information retrieval system can actually operate to retrieve some information, that information must have already been stored inside the computer. Originally it will usually have been in the form of documents.
- The computer, however, is not likely to have stored the complete text of each document in the natural language in which it was written. It will have, instead, a document representative which may have been produced from the documents either manually or automatically.

- Term weighting is an important aspect of modern text retrieval systems. Terms are words, phrases, or any other indexing units used to identify the contents of a text. Since different terms have different importance in a text, an important indicator - the term weight - is associated with every term.
- The retrieval performance of the information retrieval systems is largely dependent on similarity measures. Furthermore, a term weighting scheme plays an important role for the similarity measure.
- There are three components in a weighting scheme :

$$a_{ij} = g_i * t_{ij} * d_j$$

Where g_i is the global weight of the i^{th} term,

t_{ij} is the local weight of the i^{th} term in the j^{th} document,

d_j is the normalization factor for the j^{th} document.

2.2.1 TF-IDF Weighting

- Term Frequency (TF) : Frequency of occurrence of query keyword in document.
- Inverse Document Frequency (IDF) : How many documents the query keyword occurs in.
- Inverse Document Frequency (IDF) is a popular measure of a word's importance. It's defined as the logarithm of the ratio of number of documents in a collection to the number of documents containing the given word. This means rare words have high IDF and common words have low IDF.
- Term frequency is a measure of the importance of terms i in document j.
- Inverse document frequency is a measure of the general importance of the term.
- High term frequency for "apple" means that apple is an important word in a specific document. But high document frequency (low inverse document frequency) for "apple", given a particular set of documents, means that apple is not all that important overall, since it is in all of the documents.
- The weight increases as the number of documents in which the term appears decreases. High value indicates that the word occurs more often in this document than average.
- The term frequency $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.
- A document with $t_f = 10$ occurrences of the term is more relevant than a document with $t_f = 1$ occurrence of the term. But not 10 times more relevant. Relevance does not increase proportionally with term frequency.

- The document frequency is the number of documents in the collection that the term occurs in. We define the idf weight of term t as follows :

$$\text{idf weight } (idf_t) = \log 10 \frac{N}{df_t}$$

here N is the number of documents in the collection

- The tf-idf weight of a term is the product of its tf weight and its idf weight

$$W_{t,d} = (1 + \log t f_{t,d}) \log \frac{N}{df_t}$$

➤ Stop lists and Stemming :

- **Stoplists :** This is a list of words that we should ignore when processing documents, since they give no useful information about content.
- **Stemming :** This is the process of treating a set of words like "fights, fighting, fighter, …" as all instances of the same term - in this case the stem is "fight".

➥ 2.2.2 Luhn's Ideas

- Luhn's determined significant index words by their frequency counts in the document text, as representation or characterization of a document. The result, a derived list of index terms, can be called keywords or terms for each document.
- Luhn's idea was that words with high and low frequency are too common and too rate to contribute significantly to the content of a document, and that only words with medium frequency are significant.
- The idea of using frequency to measure word significance is based on the fact that a writer normally repeats certain words when writing on a subject.
- The more a specific word is found in a document, the more significance may be assigned to these words. He also takes common words in consideration. Such words must be present to tie the words together, but the type of significance does not reside in such words.
- These common words must be segregated and excluded from being considered as significant words by some method. As frequency has been proposed as a criterion, this step can be seen as an upper boundary of significant words.
- A lower boundary is needed too. This is to bracket the most useful range of words. Fig. 2.2.1 shows Luhn's idea.

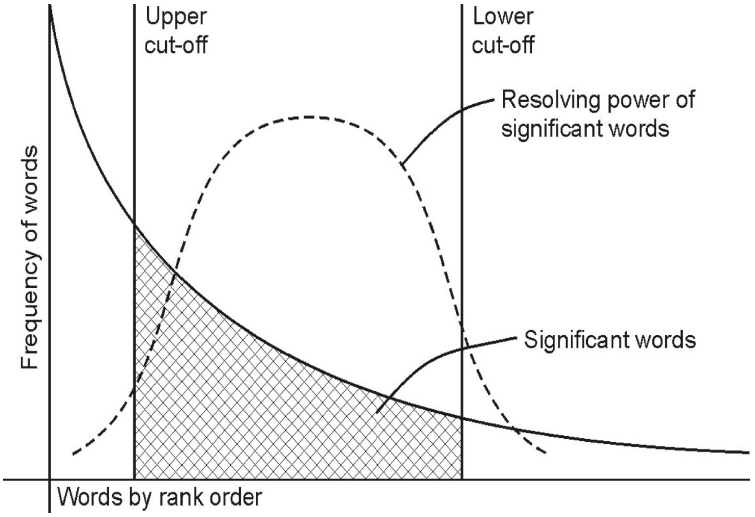


Fig. 2.2.1 : Luhn's idea

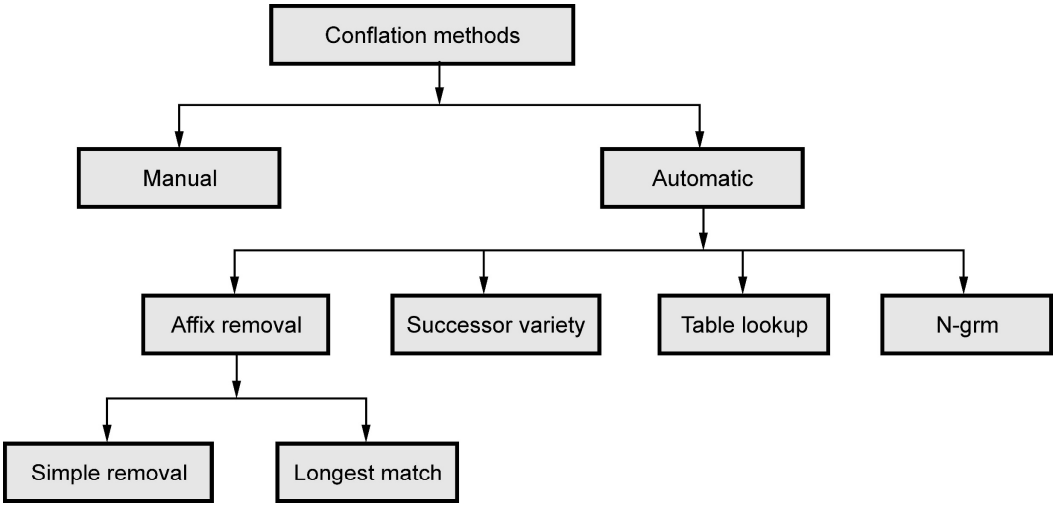
- The frequency curve actually demonstrates Zipf's law. With this law Zipf stated that the product of the frequency of occurrences of words and the rank order is approximately constant.
- Luhn's method for each document :
  1. Filter terms in the document using a list of stopwords.
  2. Normalize terms by stemming like differentiate, different, differently.
  3. Calculate frequencies of normalized terms.
  4. Remove non-frequent terms.
- Problem with Luhn's model is that the cut off points to be determined empirically, but trial and error.

➥ 2.2.3 Conflation Algorithm

- Information retrieval is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information. The user's information need is represented by a query or profile, and contains one or more search terms, plus some additional information such importance weights.
- The retrieval decision is made by comparing the terms of the query with the index terms (important words or phrases) appearing in the document itself.

- The decision may be binary (retrieve/reject), or it may involve estimating the degree of relevance that the document has to the query.
- For example morphological variants (e.g., COMPUTATIONAL, COMPUTER, COMPUTERS, COMPUTING etc.) are generally the most common, with other sources including valid alternative spellings, mis-spellings, and variants arising from transliteration and abbreviation.
- A number of stemming algorithms, or stemmers, have been developed, which attempt to reduce a word to its stem or root form. Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form. The process of stemming is often called conflation. These programs are commonly referred to as stemming algorithms or stemmers.
- Word conflation is the process by which a group of words that have the same meaning are reduced to a single term. For example : COLLECT, COLLECTED, COLLECTING, COLLECTION, COLLECTIONS , can all be conflated to the common term COLLECT.
- The key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form. It also reduces the dictionary size. The number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time.
- Conflation algorithms are classified into two main classes : Stemming algorithms which are language dependent and string-similarity algorithms which are language independent.
- Stemming algorithms typically try to reduce related words to the same base (or stem).System will usually consist of three parts :

  1. Removal of high frequency words,
  2. Suffix stripping,
  3. Detecting equivalent stems.



- The removal of high frequency words, 'stop' words or 'fluff' words is one way of implementing Luhn's upper cut-off. This is normally done by comparing the input text with a 'stop list' of words which are to be removed.
- The advantages of the process are not only that non-significant words are removed and will therefore not interfere during retrieval, but also that the size of the total document file can be reduced by between 30 % to 50 %.
- The second stage, suffix stripping, is more complicated. A standard approach is to have a complete list of suffixes and to remove the longest possible one. For example, we may well want UAL removed from FACTUAL but not from EQUAL.
- Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" is stored which provide a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include :
  1. If the word ends in 'ed', remove the 'ed'
  2. If the word ends in 'ing', remove the 'ing'
  3. If the word ends in 'ly', remove the 'ly'
- Following are the simple steps of conflation algorithm :
  1. Open and read each input file and create a single index file.
  2. Remove or filer out all stop words.
  3. Remove all suffixes / affixes from each word if present.
  4. Count frequencies of occurrences for each root word from 3.
  5. Apply porters rules / algorithm for each root word from 3 and store in index file.

➤ **Affix removal stemmers :**

a) If a word ends in "ies" but not "eies" or "aies" then replace "ies" with "y"

b) If a word ends in "es" but not "aes", "ees", or "oes" then replace "es" with "e"

c) If a word ends in "s", but not "us" or "ss" then replace "s" with NULL.

➥ 2.2.4 Cosine Similarity

- This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities.
- As a result, the likelihood that two documents do not share the majority is very high and does not create a satisfactory metric for determining similarities.
- In this similarity metric, the attributes (or words, in the case of the documents) is used as a vector to find the normalized dot product of the two documents. By determining the cosine similarity, the user is effectively trying to find cosine of the angle between the two objects.
- For cosine similarities resulting in a value of 0, the documents do not share any attributes (or words) because the angle between the objects is 90 degrees.
- Cosine similarity is expressed as a mathematical equation :

$$\text{similarity}(x, y) = \cos(\theta) = \frac{x \cdot y}{\| x \| * \| y \|}$$

- Cosine is a normalized dot product. Documents are ranked by decreasing cosine value.
- A term that appears in many documents should not be regarded as more important than one that appears in few documents.
- A document with many occurrences of a term should not be regarded as less important than a document with few occurrences of the term

University Question

1. Briefly explain weighting and cosine similarity. **AU : Dec.-16, Marks 16**

➠ 2.3 Probabilistic Model **AU : Dec.-17**

- This model is introduced by Roberston and Sparck Jones in 1976. It is also called Binary Independence Retrieval (BIR) model.
- Idea : Given a user query q, and the ideal answer set R of the relevant documents, the problem is to specify the properties for this set.

- Assumption (probabilistic principle) : *The probability of relevance depends on the query and document representations only; ideal answer set R should maximize the overall probability of relevance*.
- The probabilistic model tries to estimate the probability that the user will find the document d_j relevant with ratio P(d_j relevant to q) / P(d_j nonrelevant to q).

➤ **Definition**

- All index term weights are all binary i.e. $w_{i,j} \in \{0,1\}$
- Let R be the set of documents known to be relevant to query q
- Let $\bar{R}$ be the complement of R.
- Let P(R|d_j) be the probability that the document d_j is relevant to the query q
- Let P(R|d_j) be the probability that the document d_j is nonelevant to query q
- The similarity sim(d_j, q) of the document d_j to the query q is defined as the ratio

$$\text{sim}(\bar{d}_j, q) = \frac{\Pr(R \mid \bar{d}_j)}{\Pr(\bar{R} \mid \bar{d}_j)}$$

- Using Bayes' rule

$$\text{sim}(\bar{d}_j, q) = \frac{\Pr(\bar{d}_j \mid R) \times \Pr(R)}{\Pr(\bar{d}_j \mid \bar{R}) \times \Pr(\bar{R})}$$

where :

a. P(R) stands for the probability that a document randomly selected from the entire collection is relevant.

b. P(d_j | R) stands for the probability of randomly selecting the document dj from the set R of relevant documents.

$$\text{sim}(\bar{d}_j, q) = \frac{\Pr(\bar{d}_j \mid R)}{\Pr(\bar{d}_j \mid \bar{R})} + \log \frac{\Pr(R)}{\Pr(\bar{R})}$$

- Assuming independence of index terms and given q = (d_1, d_2, …, d_t),

$$\Pr(\bar{d}_j|R) = \prod_{i=1}^{t} \text{pr}(k_i = d_i|R)$$

$$\Pr(\bar{d}_j|\bar{R}) = \prod_{i=1}^{t} \text{pr}(k_i = d_i|\bar{R})$$

$$\text{sim}(\bar{d}_j, q) = \frac{\prod_{i=1}^{t} \text{pr}(k_i = d_i|R)}{\prod_{i=1}^{t} \text{pr}(k_i = d_i|\bar{R})}$$

- $\Pr(k_i|R)$ stands for the probability that the index term k_i is present in a document randomly selected from the set R.
- $\Pr(\bar{k}_i|R)$ stands for the probability that the index term k_i is not present in a document randomly selected from the set R.

$$\text{sim}(\bar{d}_j, q) = \frac{\prod g_i(d_j = 1) \Pr(k_i|R) \prod g_i(d_j = 1) \Pr(\bar{k}_i|R)}{\prod g_i(d_j = 1) \Pr(k_i|\bar{R}) \prod g_i(d_j = 1) \Pr(\bar{k}_i|\bar{R})}$$

$$\therefore \quad \Pr(\bar{d}_j|R) + \Pr(\bar{d}_j|\bar{R}) = 1$$

$$\text{sim}(\bar{d}_j, q) = \sum_{i=1}^{t} \left(\log \frac{p(k_i|R)}{1 - p(k_i|R)} + \log \frac{1 - p(k_i|\bar{R})}{p(k_i|\bar{R})} \right)$$

$$\text{sim}(\bar{d}_j, q) = \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \left(\log \frac{p(k_i|R)}{1 - p(k_i|R)} + \log \frac{1 - p(k_i|\bar{R})}{p(k_i|\bar{R})} \right)$$

➤ **Advantage :**

Documents are ranked in decreasing order of their probability of being relevant.

➤ **Disadvantages :**

1. The need to guess the initial relevant and non-relevant sets.
2. Term frequency is not considered.
3. Independence assumption for index terms.

University Question

1. Explain in detail about binary independance model for Probability Ranking Principle (PRP). **AU : Dec.-17, Marks 10**

2.4 Latent Semantic Indexing Model

AU : May-17

- Information is retrieved by literally matching terms in documents with those of a query. However, lexical matching methods can be inaccurate when they are used to match a user's query.
- Latent Semantic Indexing (LSI) tries to overcome the problems of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval. LSI assumes that there is some underlying or latent structure in word usage that is partially obscured by variability in word choice.
- LSI is an indexing and retrieval method that uses a mathematical technique called Singular Value Decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text.
- Latent semantic indexing is a technique that projects queries and documents into a space with "latent" semantic dimensions.
- It is statistical method for automatic indexing and retrieval that attempts to solve the major problems of the current technology. It is intended to uncover latent semantic structure in the data that is hidden. It creates a semantic space wherein terms and documents that are associated are placed near one another.

➤ General idea

1. Map documents (and terms) to a low-dimensional representation.
2. Design a mapping such that the low-dimensional space reflects semantic associations.
3. Compute document similarity based on the inner product in the latent semantic space.

➤ Goals

1. Similar terms map to similar location in low dimensional space.
2. Noise reduction by dimension reduction.

- Latent semantic indexing is the application of a particular mathematical technique, called Singular Value Decomposition or SVD, to a word-by-document matrix. SVD is a least-squares method.

Example 2.4.1 : Following are the technical memo titles and construct document matrix

1. c1 : Human machine interface for ABC computer applications
2. c2 : A survey of user opinion of computer system response time
3. c3 : The EPS user interface management system

4. c4 : System and human system engineering testing of EPS
5. c5 : Relation of user perceived response time to error measurement
6. m1 : The generation of random, binary, ordered trees
7. m2 : The intersection graph of paths in trees
8. m3 : Graph minors IV : Widths of trees and well-quasi-ordering
9. m4 : Graph minors : A survey

▣ Solution :

First we construct the document matrix

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

(**Note :** 1 represents that word is present in that technical memo titles)

University Question

1. Give an example for latent semantic indexing and explain the same.

AU : May-17, Marks 8

2.5 Neural Network Model

- Neural networks are known to be good pattern matchers.
- Fig. 2.5.1 shows neural network model for information retrieval. It consists of three layers: query, document term and document nodes.

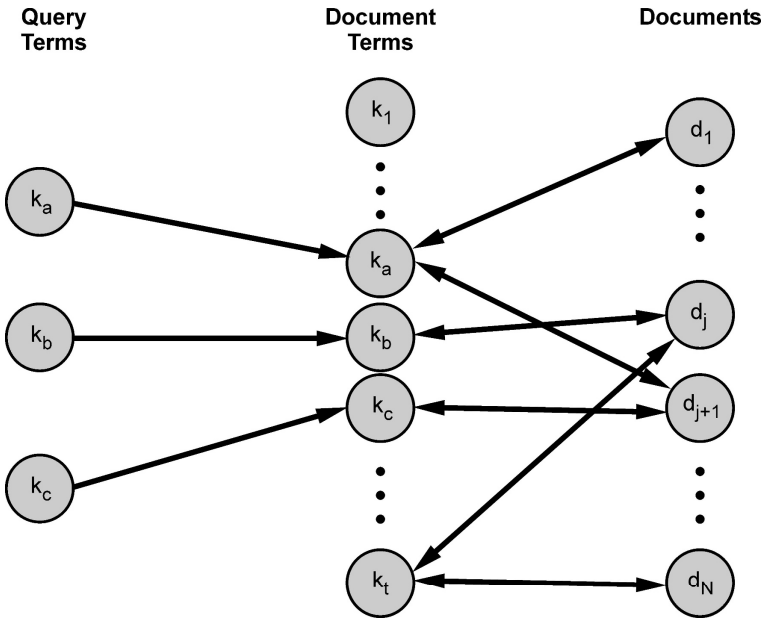


Fig. 2.5.1 : Neural network model for information retrieval

- In neural network, nodes are processing unit and edges play the role of synaptic connections. Weights are assigned to each edge of neural network.
- The strength of a propagating signal is modelled by a weight assigned to each edge. The state of a node is defined by its *activation level. D*epending on its activation level, a node might issue an output signal.
- First level of propagation :
  a) Query terms issue the first signals.
  b) These signals propagate accross the network to reach the document nodes.
- Second level of propagation :
  a) Document nodes might themselves generate new signals which affect the document term nodes.
  b) Document term nodes might respond with new signals of their own.

2.6 Relevance Feedback and Query Expansion

AU : Dec.-16, 17, May-17

- Users have no detailed knowledge of collection makeup and the retrieval environment. Most users often need to reformulate their queries to obtain the results of their interest.

- Thus, the first query formulation should be treated as an initial attempt to retrieve relevant information. Documents initially retrieved could be analyzed for relevance and used to improve the initial query.
- Fig. 2.6.1 shows relevance feedback on initial query.

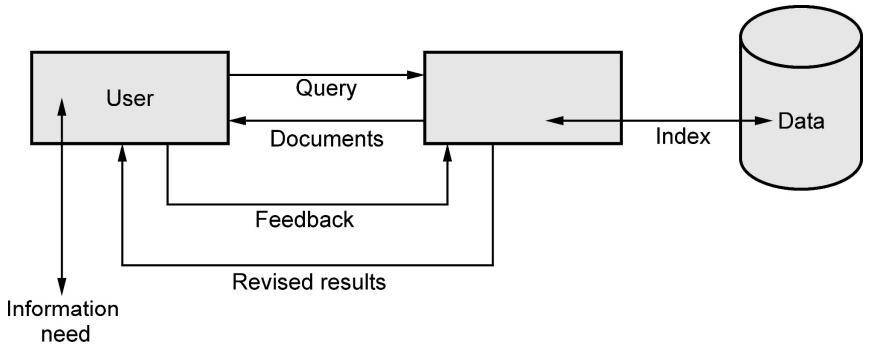


Fig. 2.6.1 (a) : Relevance feedback

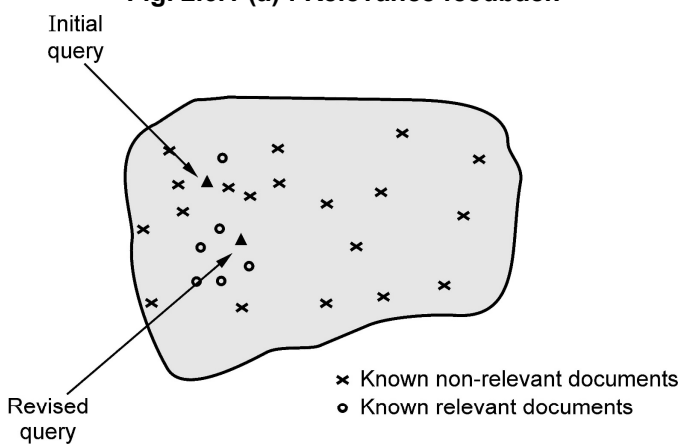


Fig. 2.6.1 (b) : Relevance feedback on initial query

- The process of query modification is commonly referred as
  1. Relevance feedback, when the user provides information on relevant documents to a query.
  2. Query expansion, when information related to the query is used to expand it
- Here we refer to both of them as feedback methods.
- Two basic approaches of feedback methods :
  1. **Explicit feedback :** The information for query reformulation is provided directly by the users. However, collecting feedback information is expensive and time consuming.
  2. **Implicit feedback :** The information for query reformulation is implicitly derived by the system.

- There are two basic approaches for compiling implicit feedback information :
  1. Local analysis, which derives the feedback information from the top ranked documents in the result set.
  2. Global analysis, which derives the feedback information from external sources such as a thesaurus.
- **Goal of relevance feedback :**
  1. Add query terms and adjust term weights.
  2. Improve ranks of known relevant documents.
  3. Other relevant docs will also be ranked higher.

➤ **Relevance feedback : Basic idea**

- The user issues a short and simple query. The search engine returns a set of documents. User marks some docs as relevant, some as non-relevant.
- Search engine computes a new representation of the information need. Hope that it is better than the initial query. Search engine runs new query and returns new results. New results have better recall. Fig. 2.6.2 shows relevance and click feedback.

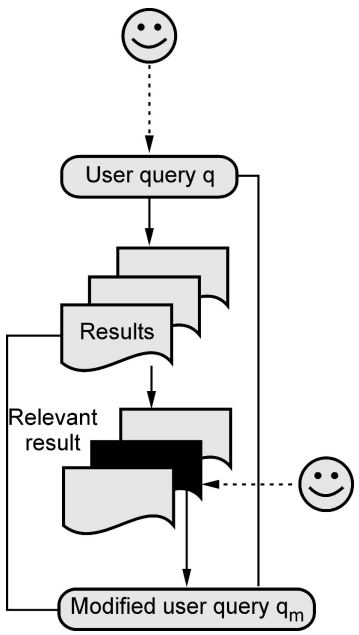


Fig. 2.6.2 (a) : Relevance feedback

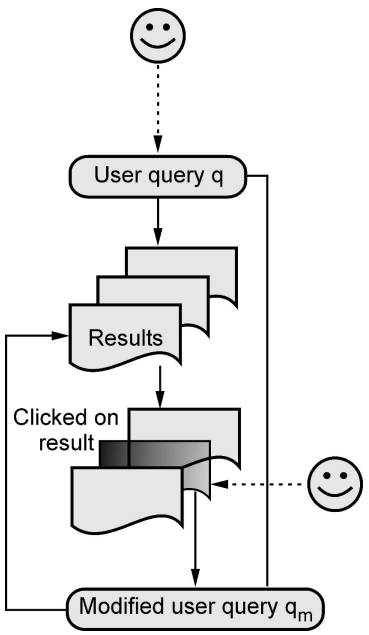


Fig. 2.6.2 (b) : Click feedback

- Characteristics of relevance feedback :
  1. It shields the user from the details of the query reformulation process.
  2. It breaks down the whole searching task into a sequence of small steps which are easier to grasp.
  3. Provide a controlled process designed to emphasize some terms (relevant ones) and de-emphasize others (non-relevant ones).
- Issues with relevance feedback
  1. The user must have sufficient knowledge to form the initial query.
  2. This does not work too well in cases like : Misspellings, CLIR and mismatch in user's and document's vocabulary (Burma vs. Myanmar).
  3. Relevant documents have to be similar to each other while similarity between relevant and non-relevant document should be small.
  4. Long queries generated may cause long response time.
  5. Users are often reluctant to participate in explicit feedback.

➤ **Advantages of relevance feedback**

1. Relevance feedback usually improves average precision by increasing the number of good terms in the query.
2. It breaks down the search operation into a sequence of small search steps, designed to approach the wanted subject area gradually.
3. It provides a controlled query alteration process designed to emphasize some terms and to deemphasize others, as required in particular search environments.

➤ **Disadvantages of relevance feedback**

1. More computational work
2. Easy to decrease precision

➥ 2.6.1 Rocchio Method

- Rocchio's model is a classic framework for implementing relevance feedback via improving the query representation.
- Rocchio's algorithm is an iterative method which can generate successive queries based on an initial query and a set of relevant documents selected among the top k ranked documents.
- This algorithm uses vector space model to pick a relevance feedback theory.

➤ **Rocchio Properties :**

1. Does not guarantee a consistent hypothesis.
2. Forms a simple generalization of the examples in each class (a prototype).
3. Prototype vector does not need to be averaged or otherwise normalized for length since cosine similarity is insensitive to vector length.
4. Classification is based on similarity to class prototypes.

- Rocchio method is represented by following formula :

$$Q_1 = Q_0 + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} + \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2}$$

where Q_0 = The vector for the initial query

R_i = The vector for the relevant document i

S_i = The vector for the non-relevant document i

n_1 = The number of relevant documents chosen

n_2 = The number of non-relevant documents chosen.

β and γ tune the importance of relevant and non relevant terms (in some studies best to set β to 0.75 and γ to 0.25).

- Rocchio's clustering algorithm was developed on the SMART project. It operates in three stages.

1. In the first stage it selects a number of objects as cluster centers. The remaining objects are then assigned to the centers or to a 'rag-bag' cluster. On the basis of the initial assignment the cluster representatives are computed and all objects are once more assigned to the clusters. The assignment rules are explicitly defined in terms of thresholds on a matching function. The final clusters may overlap.
2. The second stage is essentially an iterative step to allow the various input parameters to be adjusted so that the resulting classification meets the prior specification of such things as cluster size, etc. more nearly.
3. The third stage is for 'tidying up'. Unassigned objects are forcibly assigned and overlap between clusters is reduced.

Example 2.6.1 : An IR system returns 8 relevant documents and 10 non-relevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search and what is its recall ?

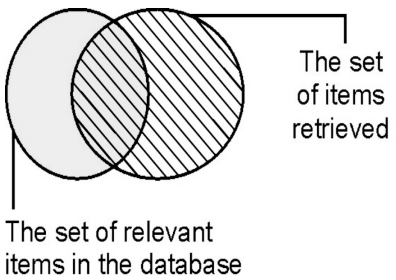
▣ Solution :

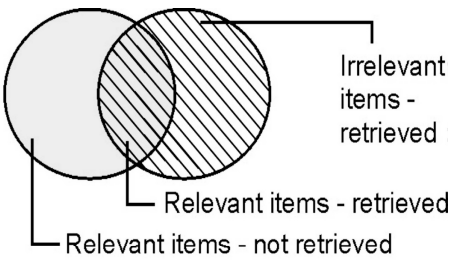
Precision = 8 / 18 = 0.44

Recall = 8 / 20 = 0.40

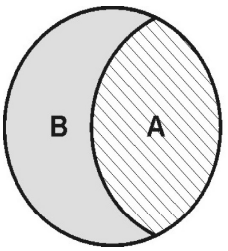
➥ 2.6.2 Precision and Recall

- **Relevance :** Relevance is a subjective notion. Different users may differ about the relevance or non-relevance of particular documents to given questions.
- In response to a query, an IR system searches its document collection and returns a ordered list of responses. It is called the retrieved set or ranked list. The system employs a search strategy or algorithm and measure the quality of a ranked list.
- A better search strategy yields a better ranked list and better ranked lists help the user fill their information need.
- Precision and recall are the basic measures used in evaluating search strategies. As shown in the first two figures, these measures assume :
  1. There is a set of records in the database which is relevant to the search topic
  2. Records are assumed to be either relevant or irrelevant.
  3. The actual retrieval set may not perfectly match the set of relevant records.





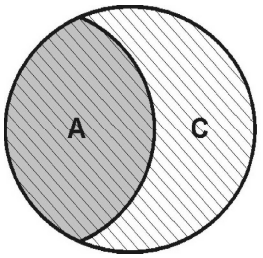
- **Recall** is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.



A = Number of relevant records retrieved.
B = Number of relevant records not retrieved.

$$\text{Recall} = \frac{A}{A+B} \times 100\ \%$$

- **Precision** is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.



A = Number of relevant records retrieved.
C = Number of irrelevant records retrieved

$$\text{Precison} = \frac{A}{A+C} \times 100\ \%$$

- As recall increases, the precision decreases and recall decreases the precision increases.

Example 2.6.2 : Assume the following :
A database contains 80 records on a particular topic
A search was conducted on that topic and 60 records were retrieved.
Of the 60 records retrieved, 45 were relevant.
Calculate the precision and recall scores for the search.

▣ **Solution :** Using the designations above :

A = The number of relevant records retrieved,

B = The number of relevant records not retrieved, and

C = The number of irrelevant records retrieved.

In this example A = 45, B = 35 (80 – 45) and C = 15 (60 – 45).

$$\text{Recall} = \frac{45}{45+35} \times 100\ \%$$

$$\text{Recall} = \frac{45}{80} \times 100\ \%$$

$$\textbf{Recall} = \textbf{56.25 \%}$$

$$\text{Precision} = \frac{A}{A+C} \times 100\ \%$$

$$\text{Precision} = \frac{45}{45+15} \times 100\ \% = \frac{45}{60} \times 100$$

$$\textbf{Precision} = \textbf{75 \%}$$

Example 2.6.3 : 20 found documents, 18 relevant, 3 relevant documents are not found, 27 irrelevant are as well not found. Calculate the precision and recall and fallout scores for the search.

▣ **Solution :**

Precision : 18/20 = 90 %

Recall : 18/21 = 85.7 %

Fall-out : 2/29 = 6.9 %

- Recall is a non-decreasing function of the number of docs retrieved. In a good system, precision decreases as either the number of docs retrieved or recall increases. This is not a theorem, but a result with strong empirical confirmation.
- The set of ordered pairs makes up the precision-recall graph. Geometrically when the points have been joined up in some way they make up the precision-recall curve. The performance of each request is usually given by a precision-recall curve. To measure the overall performance of a system, the set of curves, one for each request, is combined in some way to produce an average curve.
- Assume that set R_q containing the relevant document for q has been defined. Without loss of generality, assume further that the set R_q is composed of the following documents :

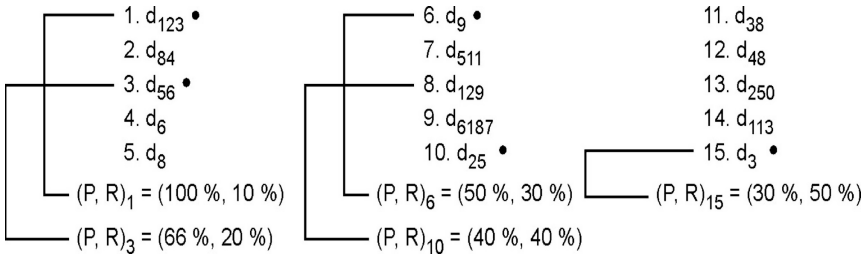
  $$R_q = \{ d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123} \}$$

  There are ten documents which are relevant to the query q.
- For the query q, a ranking of the documents in the answer set as follows.

  Ranking for query q :

1. d_{123} *	6. d_9 *	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{59} *	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25} *	15. d_3 *

- The documents that are relevant to the query q are marked with star after the document number. Ten relevant documents, five included in Top 15.



- Fig 2.6.3 shows the curve of precision versus recall. By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a precision-recall curve.
- The precision versus recall curve is usually plotted based on 11 standard recall level : 0 %,10 %,….,100 %.

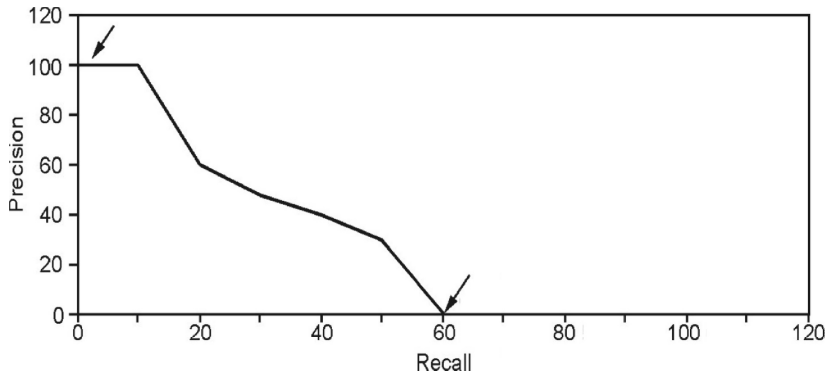


Fig. 2.6.3 : Precision versus recall curve

- In this example : The precisions for recall levels higher than 50 % drop to 0 because no relevant documents were retrieved. There was an interpolation for the recall level 0 %.
- Since the recall levels for each query might be distinct from the 11 standard recall levels.

2.6.2.1 Interpolated Recall-Precision

- Idea : If locally precision increases with increasing recall, then you should get to count that. So you take the max of precisions to right of value.
- Consider again the set of 15 ranked documents. Assume that the set of relevant documents for the query q has changed and is now given by $R_q = \{ d_3, d_{56}, d_{129} \}$

- The three relevant document :

1. d_{123}
2. d_{84}
3. d_{56} •
4. d_6
5. d_8

$(P, R)_3 = (33.3\ \%, 33.3\ \%)$

6. d_9
7. d_{511}
8. d_{129} •
9. d_{187}
10. d_{25}

$(P, R)_8 = (25\ \%, 66.6\ \%)$

11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 •

$(P, R)_{15} = (20\ \%, 100\ \%)$

- Interpolated precisions at standard recall levels :

$$\bar{P}(r_j) = \max_{rj \le r \le rj+1} P(r)$$

The j^{th} standard recall level.

- Which means that the interpolated precision at the j^{th} standard recall level is the maximum known precision at any recall level between the j^{th} recall level and the $(j + 1)^{th}$ recall level.
- At recall levels 0 %, 10 %, 20 % and 30 %, the interpolated precision is equal to 33.3 %.
- At recall levels 40 %, 50 %, 60 % the interpolated precision is equal to 25 %.

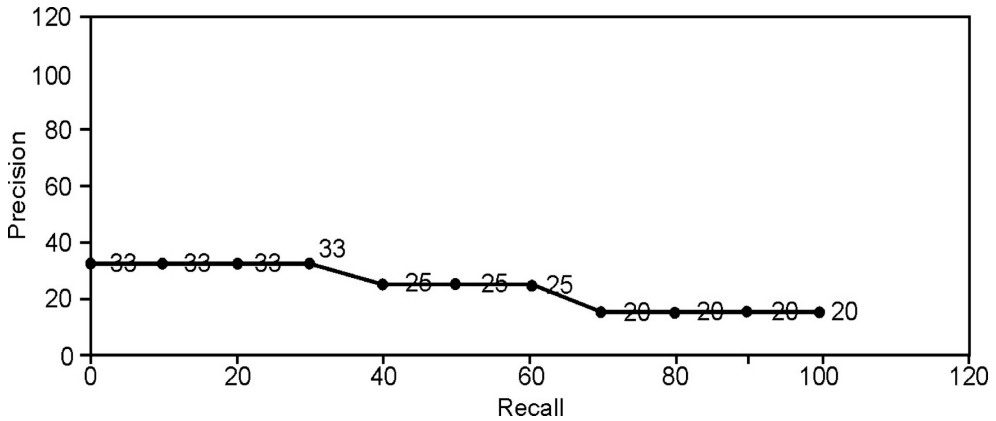


Fig. 2.6.4 : Curve for interpolated precision and recall

- At recall levels 70 %, 80 %, 90 % and 100 %, the interpolated precision is equal to 20 %.
- Fig. 2.6.4 shows the curve for interpolated precision and recall.
- Following Fig. 2.6.5 shows the comparison between precision-recall curve and interpolated precision.

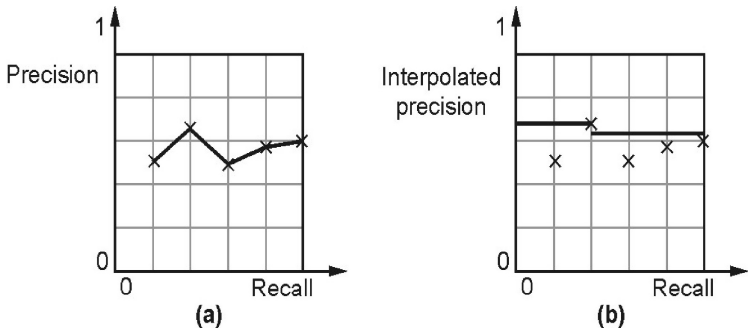


Fig. 2.6.5

➤ **Advantages of interpolated recall-precision**

1. Simple, intuitive, and combined in single curve.
2. Provide quantitative evaluation of the answer set and comparison among retrieval algorithms.
3. A standard evaluation strategy for IR systems.

➤ **Disadvantages of interpolated recall-precision**

1. Can not know true recall value except in small document collections.
2. Assume a strict document rank ordering.

- It is an experimental fact that average precision-recall graphs are monotonically decreasing. Consistent with this, a linear interpolation estimates the best possible performance between any two adjacent observed points. To avoid inflating the experimental results it is probably better to perform a more conservative interpolation.

2.6.2.2 Mean Average Precision (MAP)

- Also called average precision at seen relevant documents. It determine precision at each point when a new relevant document gets retrieved.
- Average of the precision value obtained for the top k documents, each time a relevant doc is retrieved.
- Avoids interpolation, use of fixed recall levels. MAP for query collection is arithmetic averaging.
- Average precision-recall curves are normally used to compare the performance of distinct IR algorithms.

- Use P = 0 for each relevant document that was not retrieved. Determine average for each query, then average over queries :

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j} P(doc_i)$$

where

Q_j = Number of relevant document for query j.

N = Number of queries.

$P(doc_i)$ = Precision at i^{th} relevant document

Example 2.6.4 :

Query 1		
Rank	**Relev.**	**P(doc$_i$)**
1	X	1.00
2		
3	X	0.67
4		
5		
6	X	0.50
7		
8		
9		
10	X	0.40
11		
12		
13		
14		
15		
16		
17		
18		
19		
20	X	0.50
AVG :		0.564

Query 2		
Rank	**Relev.**	**P(doc)**
1	X	1.00
2		
3	X	0.67
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15	X	0.2
AVG :		0.623

MAP favors systems which return relevant documents fast.

▣ **Solution :**

$$\text{MAP} = \frac{0.564 + 0.623}{2}$$

$$\text{MAP} = 0.594$$

- A necessary consequence of its monotonicity is that the average P-R curve will also be monotonically decreasing. It is possible to define the set of observed points in such a way that the interpolate function is not monotonically decreasing. In practice, even for this case, we have that the average precision-recall curve is monotonically decreasing.

➤ **Precision-recall appropriateness**

- Precision and recall have been extensively used to evaluate the retrieval performance of IR algorithms. However, a more careful reflection reveals problems with these two measures :
- First, the proper estimation of maximum recall for a query requires detailed knowledge of all the documents in the collection.
- Second, in many situations the use of a single measure could be more appropriate.
- Third, recall and precision measure the effectiveness over a set of queries processed in batch mode.
- Fourth, for systems which require a weak ordering though, recall and precision might be inadequate.

➤ **Single value summaries**

- Average precision-recall curves constitute standard evaluation metrics for information retrieval systems. However, there are situations in which we would like to evaluate retrieval performance over individual queries. The reasons are twofold :
  1. First, averaging precision over many queries might disguise important anomalies in the retrieval algorithms under study.
  2. Second, we might be interested in investigating whether a algorithm outperforms the other for each query.
- In these situations, a single precision value can be used.

➤ **R-Precision**

- If we have a known set of relevant documents of size Rel, then calculate precision of the top Rel docs returned.

- Let R be the total number of relevant docs for a given query. The idea here is to compute the precision at the Rth position in the ranking.
- For the query q1, the R value is 10 and there are 4 relevant among the top 10 documents in the ranking. Thus, the R-Precision value for this query is 0.4.
- The R-precision measure is a useful for observing the behavior of an algorithm for individual queries. Additionally, one can also compute an average R-precision figure over a set of queries.
- However, using a single number to evaluate a algorithm over several queries might be quite imprecise.

➤ **Precision histograms**

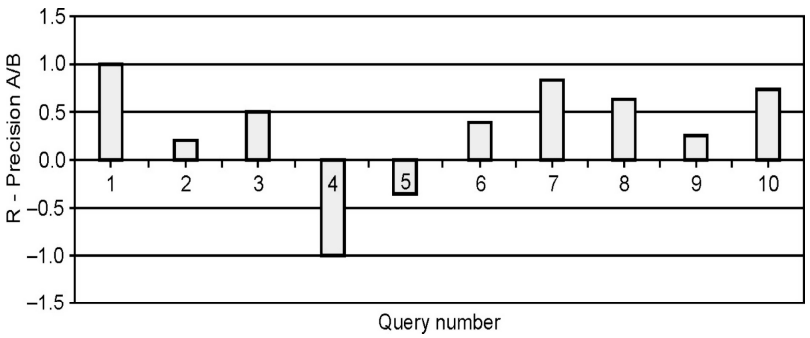


Fig. 2.6.6 : Precision histograms

- The R-precision computed for several queries can be used to compare two algorithms as follows :
- Let,

  $RP_A(i)$: R-precision for algorithm A for the i-th query

  $RP_B(i)$: R-precision for algorithm B for the i-th query

- Define, for instance, the difference :

$$RP_{A/B}(i) = RP_A(i) - RP_B(i)$$

- A positive value of $RP_{A/B}(i)$ indicates a better retrieval performance by algorithm A while a negative value indicates a better retrieval performance by algorithm B. Fig. 2.6.6 shows the $RP_{A/B}(i)$ values for two retrieval algorithms over 10 example queries.
- The algorithm A performs better for 8 of the queries, while the algorithm B performs better for the other 2 queries.

2.6.3 Probability Relevance Feedback

- Guess a preliminary probabilistic description of R and use it to retrieve a first set of documents V. Interact with the user to refine the description: partition V into relevant and non-relevant VNR.
- If user has told us some relevant and some irrelevant documents, then we can proceed to build a probabilistic classifier, such as a Naive Bayes model :

$$P(t_k|R) = |D_{rk}| / |D_r|$$

$$P(t_k|NR) = |D_{nrk}| / |D_{nr}|$$

Where t_k is a term; D_r is the set of known relevant documents; D_{rk} is the subset that contain t_k ; D_{nr} is the set of known irrelevant documents; D_{nrk} is the subset that contain t_k.

- Even though the set of known relevant documents is a perhaps small subset of the true set of relevant documents, if we assume that the set of relevant documents is a small subset of the set of all documents then the estimates given above will be reasonable. This gives a basis for another way of changing the query term weights.
- Reestimate p_i and r_i on the basis of these

$$p_i^{(2)} = \frac{|V_i| + kp_i^{(1)}}{|V| + k}$$

- Repeat, thus generating a succession of approximations to R.

2.6.4 Pseudo Relevance Feedback

- Pseudo relevance feedback , also known as blind relevance feedback , provides a method for automatic local analysis. It automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction.
- The method is to do normal retrieval to find an initial set of most relevant documents, to then assume that the top k ranked documents are relevant, and finally to do relevance feedback as before under this assumption.
- Relevance feedback is considered as pseudo (or blind) relevance feedback when there is an assumption that the top documents retrieved have a higher precision and that their terms represent the subject expected to be retrieved.
- In other words, it is assumed that the documents on the top of the retrieval list are relevant to the query, and information from these documents is extracted to generate a new retrieval set.

- Blind relevance feedback is typically performed in two stages :
  1. An initial search using the original topic statement is performed, after which a number of terms are selected from some number of the top-ranked documents. The selected terms are then weighted and then merged with the initial query to formulate a new query.
  2. Finally the reweighted and expanded query is submitted against the same collection to produce a final ranked list of documents. Obviously there are important choices to be made regarding the number of top-ranked documents to consider, and the number of terms to extract from those documents.
- Following table shows this contingency table

	Relevant	**Not Relevant**	
In doc	R_t	$N_t - R_t$	N_t
Not in doc	$R - R_t$	$N - N_t - R + R_t$	$N - N_t$
	R	$N - R$	N

- The relevance weight is calculated using the assumption that the first 10 documents are relevant and all others are not. For each term in these documents the following weight is calculated :

$$w_t = \log \frac{\frac{R_t}{R - R_t}}{\frac{N_t - R_t}{N - N_t - R + R_t}}$$

➥ 2.6.5 Indirect Relevance Feedback

- We can also use indirect sources of evidence rather than explicit feedback on relevance as the basis for relevance feedback. This is often called implicit (relevance) feedback.
- Implicit feedback is less reliable than explicit feedback, but is more useful than pseudo relevance feedback, which contains no evidence of user judgments.
- Moreover, while users are often reluctant to provide explicit feedback, it is easy to collect implicit feedback in large quantities for a high volume system, such as a web search engine.
- DirectHit ranked documents higher that users look at more often
- On the web, DirectHit introduced the idea of ranking more highly documents that users chose to look at more often. In other words, clicks on links were assumed to indicate that the page was likely relevant to the query.

- This approach makes various assumptions, such as that the document summaries displayed in results lists (on whose basis users choose which documents to click on) are indicative of the relevance of these documents.
- In the original DirectHit search engine, the data about the click rates on pages was gathered globally, rather than being user or query specific. This is one form of the general area of clickstream mining. Today, a closely related approach is used in ranking the advertisements that match a web search query.

University Questions

1. Write about relevance feedback and query expansion. **AU : Dec.-16, Marks 16**
2. Write short notes on query expansion. **AU : May-17, Marks 6**
3. What is relevance feedback ? Explain with an example an algorithm for relevance feedback. **AU : May-17, Marks 8**
4. Write short notes on the following :

   i) Probabilistic relevance feedback ii) Pseudo relevance feedback

   iii) Indirect relevance feedback **AU : Dec.-17, Marks 16**

2.7 Reference Collection

2.7.1 TREC Collection

- TREC is a workshop series that provides the infrastructure for large-scale testing of (text) retrieval technology.
- The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program.
- Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.
- In particular, the TREC workshop series has the following goals :
  1. To encourage research in information retrieval based on large test collections;
  2. To increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
  3. To speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and

4. To increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

- First TREC conference was held at NIST in November 1992, while the second TREC conference occurred in August 1993. Starting in TREC-3, a variety of other "tracks," were introduced.
- The tracks invigorate TREC by focusing research on new areas or particular aspects of text retrieval. The tracks represent the majority of the work done in TREC.
- Set of tracks in a particular TREC depends on :
  1. Interest of participants
  2. Appropriateness of task to TREC
  3. Needs of sponsors
  4. Resource constraints

➤ **Data collection**

- The TREC collection has been growing steadily over the year. At TREC-3, the collection size was roughly 2 gigabytes while at TREC-6 it had gone up to roughly 5.8 gigabytes.
- The TREC collection is distributed in six CD-ROM disks of roughly 1 gigabyte of compressed text each. The documents come from the following sources :

WSJ	Wall Street Journal
AP	Associated Press
ZIFF	Computer Selects, Ziff-Davis
FR	Federal Register
DOE	US DOE Publications
SJMN	San Jose Mercury News
PAT	US Patents
FT	Financial TImes
CR	Congressional Record
FBIS	Foregin Broadcast Information service
LAT	LA Times

➤ Example information request

- The TREC collection includes a set of example information requests which can be used for testing a new ranking algorithm. Each request is a description of an information need in natural language. In the TREC nomenclature, each test information request is referred to as a topic.
- Following Fig. 2.7.1 shows the sample topic of TREC.

<top>

<num> Number: 351

<title> Falkland petroleum exploration

<desc> Description:

What information is available on petroleum exploration in

the South Atlantic near the Falkland Islands?

<narr> Narrative:

Any document discussing petroleum exploration in the

South Atlantic near the Falkland Islands is considered

relevant. Documents discussing petroleum exploration in

continental South America are not relevant.

</top>

Fig. 2.7.1 Sample topic of TREC

- The task of converting an information request into a system query must be done by the system itself and is considered to be an integral part of the evaluation performance.
- The topic numbered 1 to 150 was prepared for use with the TREC-1 and TREC-2 conference. The topics numbered 151 to 200 were prepared for use with the TREC-3 conference. It includes only three fields : Title, Description and Narrative.
- The topic numbered 201 to 250 was prepared for use with the TREC-4 conference. TREC-5 includes topic 251 to 300 and TREC-6 conference includes topic number 301 to 350.

➤ **Relevant documents for each example information request**

- At the TREC conferences, the set of relevant documents for each example information request is obtained from a pool of possible relevant documents. This pool is created by taking the top K documents (usually, K=100) in the ranking generated by the various participating retrieval systems.
- The documents in the pool are then shown to human assessors who ultimately decide on the relevance of each document. This technique for accessing relevance is called the pooling method and based on two assumptions.
  1. The vast majority of the relevant documents is collected in the assembled pool.
  2. The documents which are not in the pool can be considered to be not relevant.

➤ **Benchmark tasks at the TREC conferences**

- The TREC conferences include two main information retrieval tasks.

➤ **1. Ad hoc task :**

- A set of new requests are run against a fixed document database. Example for this type is library. In library, where user is asking new queries against a set of static documents.
- The participant systems receive the test information requests and execute on a pre-specified document collection.

➤ **2. Routing task :**

- A set of fixed requests are run against a database whose documents are continually changing. This is like filtering task in which the same questions are always being asked.
- The participant systems receive the test information requests and two distinct document collections.
- The first collection is used for training and allows the tuning of the retrieval algorithm. The second collection is used for testing the tuned retrieval algorithm.

➤ **TREC-6 secondary task**

- Secondary task are as follows :
  1. Chinese : Ad hoc task in which both the documents and the topics are in Chinese.
  2. Filtering : Routing task in which the retrieval algorithm has only to decide whether a new incoming document is relevant or not. No ranking of the documents taken needs to be provided. The test data is processed in time stamp order.

3. Interactive : Task in which a human searcher interacts with the retrieval system to determine the relevant documents. Documents are ruled relevant or not relevant.
4. NLP : Task aimed at verifying whether retrieval algorithms based on natural language processing offer advantages when compared to the more traditional retrieval algorithms based on the index terms.
5. Cross languages : Ad hoc task in which the documents are in one language but the topics are in a different language.
6. High precision : Task in which the user of a retrieval system is asked to retrieve ten documents that answer a given information request within five minutes.
7. Spoken document retrieval : Task in which the documents are written transcripts of radio broadcast news shows.
8. Very large corpus : Ad hoc task in which the retrieval systems have to deal with collections of size 20 gigabytes.

- In TREC-7, the NLP and Chinese secondary tasks were discontinued. TREC-7 also included a new task called Query Task in which several distinct query versions were created for each example information request.

➤ Evaluation measures at the TREC conferences

TREC conferences uses four types of evaluations :

1. Summary table statistics
2. Recall-precision average
3. Document level averages
4. Average precision histograms

➤ Summary table statistics

- Single value measures can also be stored in a table to provide a statistical summary regarding the set of all the queries in a retrieval task.
- For instance, these summary table statistics could include :

1. The number of queries.
2. Total number of documents retrieved by all queries.
3. Total number of relevant documents which were effectively retrieved when all queries are considered.
4. Total number of relevant documents which could have been retrieved by all queries.

➤ **Recall-precision average**

- It consists of a table or graph with average precision at 11 standard recall levels.
- Since the recall levels of the individual queries are seldom equal to the standard recall levels, interpolation is used to define the precision at the standard recall levels.

➤ **Document level averages**

- Average precision is computed at specified document cutoff values.
- The average precision might be computed when 5, 10, 20, 100 relevant documents have been seen.
- The average R-precision value might also be provided.

➤ **Average precision histogram**

- It consists of a graph which includes a single measure for each separate topic.
- It is difference between the R-precision for a target retrieval algorithm and the average R-precision computed from the results of all participating retrieval system.

➥ 2.7.2 The CACM and ISI Collection

- The TREC collection is a large collection which requires time consuming preparation before experiments can be carried out effectively at a local site. A small collection might include features which are not present in the larger TREC collection.
- For the experimental studies in five different test collections were developed: ADI (documents on information science), CACM, INSPEC (abstracts on electronics, computer, and physics), ISI, and Medlars (medical articles).

➤ **CACM collections**

- It is small collections about computer science literature. It is text of 3,204 documents. The documents in the CACM test collection consist of all articles published in the communication of the ACM.
- CACM collection also includes information on structured subfields as follows :
  1. Word stems from the title and abstract sections.
  2. Categories.
  3. Direct references between articles.
  4. Bibliographic coupling connections.
  5. Number of co-citations for each pair of articles.
  6. Author names.
  7. Date information.

- A unique environment for testing retrieval algorithms which are based on information derived from cross-citing patterns. The CACM collection also includes a set of 52 test information requests.
- Example : "What articles exist which deal with TSS (Time Sharing System), an operating system for IBM computers?"
- Also includes two Boolean query formulations and a set of relevant documents. Since the requests are fairly specific, the average number of relevant documents for each request is small (around 15). Precision and recall tend to be low.

➤ The ISI collection

- The 1,460 documents in the ISI test collection were selected from a previous collection assembled by Small at ISI (Institute of Scientific Information).
- The documents selected were those most cited in a cross-citation study done by Small. The main purpose is to support investigation of similarities based on terms and on cross-citation patterns.
- The documents in the ISI collection include three types of subfields as follows.
  1. Author names.
  2. Word stems from the title and abstract sections.
  3. Number of co-citations for each pair of articles.
- The ISI collection includes a total of 35 test information requests for which there are Boolean query formulations. It also includes 41 additional test information requests for which there is no Boolean query formulation.

➤ Statistics for the CACM and ISI collections

1. Document statistics for the CACM and ISI Collections

Collection	Number of documents	Number of terms	Terms/Documents
CACM	3204	10446	40.1
ISI	1460	7392	104.9

2. Query statistics for the CACM and ISI Collections

Collection	Number of queries	Terms per query	Relevant per query	Relevant in top 10
CACM	52	11.4	15.3	1.9
ISI	35 and 76	8.1	49.8	1.7

➥ 2.7.3 Benefits of TREC

➤ Benefits :

1. Made research systems scale to large collections (pre-WWW)
2. Allows for somewhat controlled comparisons.

➤ Drawbacks :

1. Emphasis on high recall, which may be unrealistic for what most users want.
2. Very long queries, also unrealistic
3. Comparisons still difficult to make, because systems are quite different on many dimensions.
4. Focus on batch ranking rather than interaction.

➠ 2.8 Part A : Short Answered Questions [2 Marks Each]

Q.1 What do you mean Information Retrieval Models ?

▣ **Ans. :** A retrieval model can be a description of either the computational process or the human process of retrieval : The process of choosing documents for retrieval; the process by which information needs are first articulated and then refined.

Q.2 What is cosine similarity ?

▣ **Ans. :** This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities.

Q.3 What is language model based IR ?

▣ **Ans. :** A language model is a probabilistic mechanism for generating text. Language models estimate the probability distribution of various natural language phenomena

Q.4 Define unigram language.

▣ **Ans. :** A unigram (1-gram) language model makes the strong independence assumption that words are generated independently from a multinomial distribution .

Q.5 What are the characteristics of relevance feedback ?

▣ **Ans. :** Characteristics of relevance feedback :

1. It shields the user from the details of the query reformulation process.
2. It breaks down the whole searching task into a sequence of small steps which are easier to grasp.

3. Provide a controlled process designed to emphasize some terms (relevant ones) and de-emphasize others (non-relevant ones)

Q.6 What are the assumptions of vector space model ?

▣ **Ans. :** Assumption of vector space model :

1. The degree of matching can be used to rank-order documents;
2. This rank-ordering corresponds to how well a document satisfying a user's information needs.

Q.7 What are the disadvantages of Boolean model ?

▣ **Ans. :** Disadvantages :

a. It is not simple to translate an information need into a Boolean expression

b. Exact matching may lead to retrieval of too few or too many documents.

c. The retrieved documents are not ranked.

d. The model does not use term weights.

Q.8 Define term frequency.

▣ **Ans. :** Term frequency (TF) : Frequency of occurrence of query keyword in document

Q.9 Explain Luhn's Ideas.

▣ **Ans. :** Luhn's basic idea to use various properties of texts, including statistical ones, was critical in opening handling of input by computers for IR. Automatic input joined the already automated output.

Q.10 What is a stemming ? Give example. **AU : May-17**

▣ **Ans. :** Conflation algorithms are used in information retrieval systems for matching the morphological variants of terms for efficient indexing and faster retrieval operations. The conflation process can be done either manually or automatically. The automatic conflation operation is also called stemming.

Q.11 What is Recall ?

▣ **Ans. :** Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection.

Q.12 What is precision ?

▣ **Ans. :** Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved.

Q.13 Explain Latent Semantic Indexing.

▣ **Ans. :** Latent Semantic Indexing is a technique that projects queries and documents into a space with "latent" semantic dimensions. It is statistical method for automatic indexing and retrieval that attempts to solve the major problems of the current technology. It is intended to uncover latent semantic structure in the data that is hidden. It creates a semantic space wherein terms and documents that are associated are placed near one another.

Q.14 List the retrieval model. **AU : Dec-16**

▣ **Ans. :** Retrieval models are Boolean model and vector model. Boolean model based on set theory and Boolean algebra. Vector model is used in information filtering, information retrieval, indexing and relevancy rankings.

Q.15 Define document preprocessing. **AU : Dec-16**

▣ **Ans. :** Document pre-processing is the process of incorporating a new document into an information retrieval system. It is a complex process that leads to the representation of each document by a select set of index terms.

Q.16 Define an inverted index. **AU : May-17**

▣ **Ans. :** An inverted index is an index into a set of documents of the words in the documents. The index is accessed by some search method. Each index entry gives the word and a list of documents, possibly with locations within the documents, where the word occurs. The inverted index data structure is a central component of a typical search engine indexing algorithm

Q.17 What is Zone index ? **AU : Dec.-17**

▣ **Ans. :** A zone is a region of the document that can contain an arbitrary amount of text, e.g., Title, Abstract and References. Build inverted indexes on zones as well to permit querying. Zones are similar to fields, except the contents of a zone can be arbitrary free text.

Q.18 State Bayes rule. **AU : Dec.-17**

▣ **Ans. :** Bayes' theorem is a method to revise the probability of an event given additional information. Bayes's theorem calculates a conditional probability called a posterior or revised probability. Bayes' theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, P(A|B) denotes the conditional probability of A occurring, given that B occurs. The two conditional probabilities P(A|B) and P(B|A) are in general different.

2.9 Multiple Choice Questions

Q.1 List the classic information retrieval model.

[a] Boolean [b] Vector [c] Probabilistic [d] All of these

Q.2 In the vector model, document's and queries are represented as vectors in a t-dimensional space.

[a] Boolean [b] Vector [c] Probabilistic [d] All of these

Q.3 _____ is a list of words that we should ignore when processing documents, since they give no useful information about content.

[a] Verb [b] Startlist [c] Stopworks [d] All of these

Q.4 The automatic conflation operation is also called _____.

[a] stemming [b] recall [c] precision [d] all of these

Q.5 _____ is the ratio of the number of relevant documents retrieved to the total number of documents retrieve(d).

[a] Recall [b] Precision [c] Stemming [d] None

Q.6 Recall is the ratio of the number of _____ retrieved to the total number of relevant documents in the collection.

[a] relevant document [b] irrelevant document

[c] structured document [d] unstructured document

Q.7 _____ model is introduced by Roberston and Sparck Jones in 1976. It is also called Binary Independence Retrieval model.

[a] Latent semantic indexing [b] Boolean

[c] Vector [d] Probabilistic

Q.8 LSI is an indexing and retrieval method that uses a mathematical technique called _____ to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text.

[a] singular value decomposition [b] multiple value decomposition

[c] singular value document [d] multiple value document

Q.9 Rocchio's model is a classic framework for implementing _____ via improving the query representation.

[a] precision and recall [b] relevance Feedback

[c] query formation [d] none

Q.10 Pseudo relevance feedback, also known as _____ relevance feedback, provides a method for automatic local analysis.

[a] probability [b] blind

[c] mean average precision [d] all of these

Q.11 Which of the following are the TREC conferences uses for evaluations ?

[a] Summary table statistics [b] Recall-precision average

[c] Document level averages [d] All of these

Q.12 The representation of a set of documents as vectors in a common vector space is known as the _____ model.

[a] Boolean [b] document selection

[c] probabilistic [d] vector

Q.13 _____ ranking methods use the query to rank all documents in the order of relevance.

[a] Text [b] Word [c] document [d] information

➤ **Answer Keys for Multiple Choice Questions**

Q.1	d	**Q.2**	b	**Q.3**	c	**Q.4**	a
Q.5	b	**Q.6**	a	**Q.7**	d	**Q.8**	a
Q.9	b	**Q.10**	b	**Q.11**	d	**Q.12**	d
Q.13	c						

❑❑❑

UNIT - III

3 Text Classification and Clustering

Syllabus

A Characterization of Text Classification - Unsupervised Algorithms: Clustering - Naïve Text Classification - Supervised Algorithms - Decision Tree - k-NN Classifier - SVM Classifier - Feature Selection or Dimensionality Reduction - Evaluation metrics - Accuracy and Error - Organizing the classes - Indexing and Searching - Inverted Indexes - Sequential Searching - Multi-dimensional Indexing.

Contents

3.1 Characterization of Text Classification

Text classification is the process of assigning tags or categories to text according to its content.

3.1.1 Machine Learning

- A (machine learning) problem is well-posed if a solution to it exists, if that solution is unique, and if that solution depends on the data / experience but it is not sensitive to (reasonably small) changes in the data / experience.
- The main goal of machine learning is to study, engineer, and improve mathematical models which can be trained with context-related data to infer the future and to make decisions without complete knowledge of all influencing elements.
- We discuss basic design issues and approaches to machine learning.
- Goal : Design a system to learn how to play checkers and enter it into the world checkers tournament.
  1) Choose the training experience
  2) Choose the target function
  3) Choose a representation for the target function
  4) Choose a function approximation algorithm
- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Another goal is to develop computational models of human learning process and perform computer simulations.
- The goal of machine learning is to build computer systems that can adapt and learn from their experience.
- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instruction. It should carry out to transform the input to output.
- For example, for addition of four numbers is carried out by giving four number as input to the algorithm and output is sum of all four numbers.
- For the same task, there may be various algorithms. It is interested to find the most efficient one, requiring the least number of instructions or memory or both.

➤ Why is Machine Learning Important ?

- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.

- Machine learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.

➤ **Supervised Learning :**

- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples.
- The task of the supervised learner is to predict the output behavior of a system for any set of input values, after an initial training phase.
- Supervised learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs are usually provided by an external teacher.
- A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier or a regression function.

➤ **Un-Supervised Learning :**

- The model is not provided with the correct results during the training. It can be used to cluster the input data in classes on the basis of their statistical properties only. Cluster significance and labelling.
- The labeling can be carried out even if the labels are only available for a small number of objects representative of the desired classes. All similar inputs patterns are grouped together as clusters.
- If matching pattern is not found, a new cluster is formed. There is no error feedback.
- They are called unsupervised because they do not need a teacher or super-visor to label a set of training examples. Only the original data is required to start the analysis.

➤ **Semi-Supervised Learning**

- Semi-supervised learning uses both labeled and unlabeled data to improve supervised learning. The goal is to learn a predictor that predicts future test data better than the predictor learned from the labeled training data alone.
- Semi-supervised learning is motivated by its practical value in learning faster, better, and cheaper. In many real-world applications, it is relatively easy to acquire a large amount of unlabeled data x.
- For example, documents can be crawled from the Web, images can be obtained from surveillance cameras, and speech can be collected from broadcast.
- Semi-supervised Learning makes use of both labeled and unlabeled data for training, typically a small amount of labeleddata with a large amount of unlabeled data.

3.1.2 Text Classification Problem

- In text classification, given a description d ∈ X of a document, where x is the *document space* ; and a fixed set of *classes* C = { c_1, c_2, c_3, ….. c_j}. Classes are also called *categories* or *labels*.
- Typically, the document space X is some type of high-dimensional space, and the classes are human defined for the needs of an application, as in the examples India and documents that talk about multicore computer chips.
- For given a *training set* D of labeled documents < d, c > where < d, c > ∈ X × C.
- For example : < d, c > = < Mumbai joins the World Trade Organization, India > for the one-sentence document Mumbai joins the World Trade Organization and the class (or label) India.
- If no restrictions are posed to the classifier, two or more classes might be assigned to a single document.

3.1.3 Text Classification Algorithm

- Text classification algorithms are classified as supervised or unsupervised algorithm. Unsupervised algorithm are used for large collections and without training data. Supervised algorithm requires training data.
- Supervised text classification algorithms are decision tree, nearest neighbours, relevance feedback, naïve Bayes, support vector machine, ensemble etc.
- Unsupervised algorithms are clustering and Naïve text classification.

3.2 Unsupervised Algorithms

3.2.1 Clustering

- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.
- Conceptual clustering : Two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.
- Clustering methods are unsupervised learning techniques. cluster distance is a distance of two closest members in each class. Clustering methods are usually categorized according to the type of cluster they produce. The clustering methods are categorized as :
  1. Hierarchical methods : These types cluster produces the output list of cluster. Small clusters of highly similar documents nested within larger clusters of less similar documents.
  2. Non-hierarchical methods : This method produced unordered lists.

- Other clustering methods are exclusive cluster and overlapping cluster. In the first case (exclusive cluster) data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster.
- The overlapping clustering uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.
- Cluster analysis involves applying one or more clustering algorithms with the goal of finding hidden patterns or groupings in a dataset. Clustering algorithms form groupings or clusters in such a way that data within a cluster have a higher measure of similarity than data in any other cluster. The measure of similarity on which the clusters are modeled can be defined by Euclidean distance, probabilistic distance, or another metric.
- Cluster analysis is an unsupervised learning method and an important task in exploratory data analysis. Popular clustering algorithms include :
    1. Hierarchical clustering : Builds a multilevel hierarchy of clusters by creating a cluster tree.
    2. k-Means clustering : Partitions data into k distinct clusters based on distance to the centroid of a cluster.
    3. Gaussian mixture models : Models clusters as a mixture of multivariate normal density components.
    4. Self-organizing maps : Uses neural networks that learn the topology and distribution of the data.

➤ **Desirable Properties of a Clustering Algorithm**

1. Scalability (in terms of both time and space)
2. Ability to deal with different data types
3. Minimal requirements for domain knowledge to determine input parameters
4. Interpretability and usability

➤ **Distance between Clusters :**

1. Single Link : smallest distance between points
2. Complete Link : largest distance between points
3. Average Link : average distance between points
4. Centroid : distance between centroids

➤ Supervised learning after clustering

- Supervised clustering is the task of automatically adapting a clustering algorithm with the aid of a training set consisting of item sets and complete partitioning of these item sets.
- Dimensionality reduction methods find correlations between features and group features. Clustering methods find similarities between instances and group instances.
- It allows knowledge extraction through number of clusters, prior probabilities and cluster parameters, i.e., center, range of features.
- Example : CRM, customer segmentation
- One advantage of preceding a supervised learner with unsupervised clustering or dimensionality reduction is that the latter does not need labeled data. Labeling the data is costly. We can use a large amount of unlabeled data for learning the cluster parameters and then use a smaller labeled data to learn the second stage of classification or regression.

➤ Examples of Clustering Applications

- **Marketing :** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- **Land use :** Identification of areas of similar land use in an earth observation database.
- **Insurance :** Identifying groups of motor insurance policy holders with a high average claim cost.
- **City-planning :** Identifying groups of houses according to their house type, value, and geographical location.
- **Earthquake studies :** Observed earthquake epicenters should be clustered along continent faults.

➥ 3.2.2 K-Mean Clustering

AU : May-17

- K-Means clustering is heuristic method. Here each cluster is represented by the center of the cluster. "K" stands for number of clusters, it is typically a user input to the algorithm; some criteria can be used to automatically estimate K.
- This method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart.
- Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

- Given K, the K-means algorithm consists of four steps :
  1. Select initial centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.
- The x_1, ..., x_N are data points or vectors of observations. Each observation (vector x_i) will be assigned to one and only one cluster. The C(i) denotes cluster number for the i^{th} observation. K-means minimizes within-cluster point scatter :

$$W(C) = \frac{1}{2} \sum_{K=1}^{K} \sum_{C(i)=K} \sum_{C(j)=K} = \| x_i - x_j \|^2 = \sum_{K=1}^{K} N_k \sum_{C(i)=K}^{K} \| x_i - m_K \|^2$$

Where m_K is the mean vector of the K^{th} cluster.

N_K is the number of observations in K^{th} cluster.

➤ K-Means Algorithm Properties

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

➤ The K-Means Algorithm Process

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point.
   a. Calculate the distance from the data point to each cluster.
   b. If the data point is closest to its own cluster, leave it where it is.
   c. If the data point is not closest to its own cluster, move it into the closest cluster.
3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

- K-means algorithm is iterative in nature. It converges, however only a local minimum is obtained. It works only for numerical data. This method easy to implement.

➤ **Advantages of K-Means Algorithm :**

1. Efficient in computation
2. Easy to implement

➤ **Weaknesses**

1. Applicable only when mean is defined.
2. Need to specify K, the number of clusters, in advance.
3. Trouble with noisy data and outliers.
4. Not suitable to discover clusters with non-convex shapes.

➥ 3.2.3 Agglomerative Hierarchical Clustering

- In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.
- Again distance between the data point is recalculated but which distance to consider when the groups has been formed ? For this there are many available methods.
- Some of them are :
  1) Single-nearest distance or single linkage.
  2) Complete-farthest distance or complete linkage.
  3) Average-average distance or average linkage.
  4) Centroid distance.
- Algorithmic steps for Agglomerative Hierarchical clustering :
  Let X = { x_1, x_2, x_3, …, x_n } be the set of data points.
  1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
  2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r), (s)] = min d[(i), (j)] where the minimum is over al pairs of clusters in the current clustering.
  3) Increment the sequence number : m = m + 1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r), (s)].
  4) Update the distance matrix, D by deleting the rows and columns correspondng to the newly formed cluster. The distance between the new cluster, denoted (r, s) and old cluster (k) is defined in this way : d[(k), (r,s)] = min (d[(k), (r)], d[(k), (s)]).
  5) If all the data points are in one cluster then stop, else repeat from step 2.

- In the agglomerative hierarchical approach, we start by defining each data point to be a cluster and combine existing clusters at each step.
- Here ar four different methods for doing this :

➤ **1. Single linkage**

- Smallest pairwise distance between elements from each cluster. Also referred to as nearest neighbour or minimum method.
- This measure defines the distance between two clustes as the minimum distance found between one case from the first cluster and one case from the second cluster.
- For example, if cluster 1 contains cases a and b, and cluster 2 contains cases c, d and e, then the distance between cluster 1 and cluster 2 would be the smallest distance found between the following pairs of cases : (a, c), (a, d), (a, e), (b, c), (b, d), and (b, e).

➤ **2. Complete linkage**

- Largest distance between elements from each cluster.
- Also referred to as furthest neighbour or maximum method. This measure is similar to the single linkage measure described above, but instead of searching for the minimum distance between pairs of cases, it considers the furthest distance between pairs of cases.
- Although this solves the problem of chaining, it creates another problem.
- Imagine that in the above example cases a, b, c, and d are within close proximity to one another based upon the pre-established set of variables; however, if case e differs considerably from the rest, then cluster 1 and cluster 2 may no longer be joined together because of the differnce in scores between (a, e) and (b, e).
- In complete linkage, outlying cases prevent close clusters to merge together because the measure of the furthest neighbour exacerbates the effects of outlying data.

➤ **3. Average linkage :**

- The average distance between elemetns from each cluster.
- Also referred to as the unweighted pair-group method using arithmetic averages.
- To overcome the limitations of single and complete linkage, Sokal and Michener proposed taking an average of the distance values between pair of cases.
- This method is supposed to represent a natural compromise between the linkage measures to provide a more acurate evaluation of the distance between clusters.
- For average linkage, the distances between each case in the first cluster and evey case in the second cluster are calculated and then averaged.

3.2.4 Naïve Text Classification

- Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. It is highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.
- A Naive Bayes classifier is a program which predicts a class value given a set of attributes.
- For each known class value,
  1. Calculate probabilities for each attribute, conditional probability for the attributes.
  2. Use the product rule to obtain a joint conditional probability for the attributes.
  3. Use Bayes rule to derive conditional probabilities for the class variable.
- Once this has been done for all class values, output the class with the highest probability.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.
- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

University Question

1. Explain k - means clustering algorithm with an example. **AU : May-17, Marks 8**

3.3 Supervised Algorithms

- In training data, data are assign the labels. In test data, data labels are unknown but not given. The training data consist of a set of training examples.
- The real aim of supervised learning is to do well on test data that is not known during learning. Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy.
- The training error is the mean error over the training sample. The test error is the expected prediction error over an independent test sample.
- Problem is that training error is not a good estimator for test error. Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to over fitting and poor generalization.
- **Training set :** A set of examples used for learning, where the target value is known.
- **Test set :** It is used only to assess the performances of a classifier. It is never used during the training process so that the error on the test set provides an unbiased estimate of the generalization error.
- Training data is the knowledge about the data source which we use to construct the classifier.

3.3.1 Decision Tree

- A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning.
- A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.
- A decision tree is a simple representation for classifying examples. A decision tree or a classification tree is a tree in which each internal node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.
- In this method a set of training examples is broken down into smaller and smaller subsets while at the same time an associated decision tree get incrementally developed. At the end of the learning process, a decision tree covering the training set is returned.
- The key idea is to use a decision tree to partition the data space into cluster (or dense) regions and empty (or sparse) regions.
- Decision tree consists of
  1. Nodes : test for the value of a certain attribute
  2. Edges : correspond to the outcome of a test and connect to the next node or leaf
  3. Leaves : terminal nodes that predict the outcome

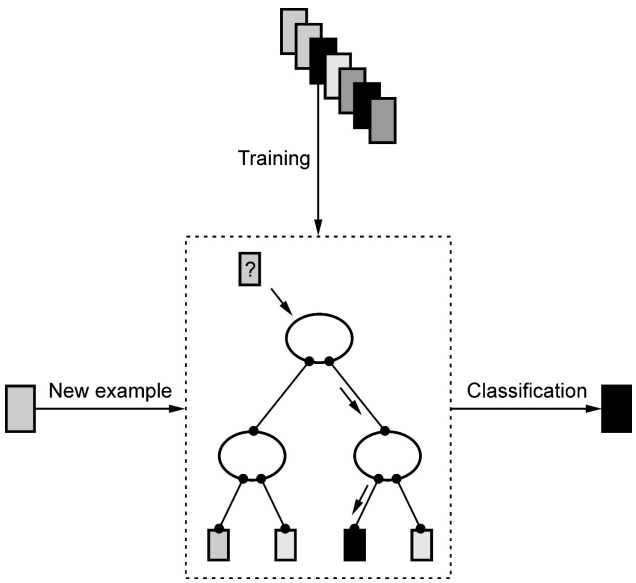


Fig. 3.3.1

- In decision tree learning, a new example is classified by submitting it to a series of tests that determine the class label of the example. These tests are organized in a hierarchical structure called a decision tree.
- Decision tree has three other names :
  1. Classification tree analysis is a term used when the predicted outcome is the class to which the data belongs.
  2. Regression tree analysis is a term used when the predicted outcome can be considered a real number (e.g. the price of a house,or a patient's length of stay in a hospital).
- Internal node denotes a test on an attribute. Branch represents an outcome of the test. Leaf nodes represent class labels or class distribution.
- A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules.

➤ **Decision Tree Algorithm**

- To generate decision tree from the training tuples of data partition D.

➤ **Input :**

1. Data partition (D)
2. Attribute list
3. Attribute selection method

➤ **Algorithm :**

1. Create a node (N)
2. If tuples in D are all of the same class then
3. Return node (N) as a leaf node labeled with the class C.
4. If attribute list is empty then return N as a leaf node labeled with the majority class in D
5. Apply Attribute selection method(D, attribute list) to find the "best" splitting criterion;
6. Label node N with splitting criterion;
7. If splitting attribute is discrete-valued and multiway splits allowed
8. Then attribute list -> attribute list -> splitting attribute
9. For (each outcome j of splitting criterion)
10. Let D_j be the set of data tuples in D satisfying outcome j;

11. If D_j is empty then attach a leaf labeled with the majority class in D to node N;
12. Else attach the node returned by Generate decision tree(Dj , attribute list) to node N;
13. End of for loop
14. Return N;

- CART analysis is a term used to refer to both of the above procedures. The name CART is an acronym from the words Classification And Regression Trees, and was first introduced by Breiman et al.
- **Learn trees in a Top-Down fashion :**

1. Divide the problem in subproblems 2. Solve each problem

➤ **Basic Divide-And-Conquer Algorithm :**

1. Select a test for root node. Create branch for each possible outcome of the test.
2. Split instances into subsets. One for each branch extending from the node.
3. Repeat recursively for each branch, using only instances that reach the branch.
4. Stop recursion for a branch if all its instances have the same class.

- Goal : Build a decision tree for classifying examples as positive or negative instances of a concept.
- Supervised learning, batch processing of training examples, using a preference bias.
- Decision tree is a tree where
  a. Each non-leaf node has associated with it an attribute (feature).
  b. Each leaf node has associated with it a classification (+ or –).
  c. Each arc has associated with it one of the possible values of the attribute at the node from which the arc is directed.
- For example :

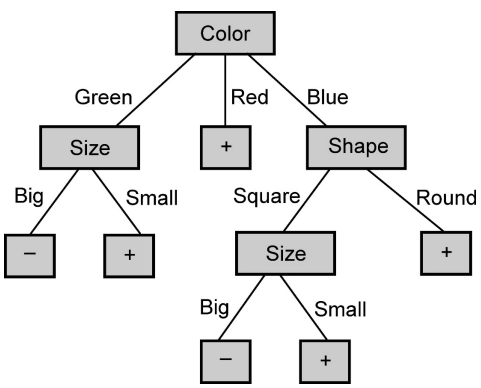


Fig. 3.3.2

- Decision tree generation consists of two phases : Tree construction and pruning.
- In tree construction phase, all the training examples are at the root. Partition examples recursively based on selected attributes.
- In tree pruning phase, the identification and removal of branches that reflect noise or outliers.
- There are various paradigms that are used for learning binary classifiers which include :
  1. Decision Trees
  2. Neural Networks
  3. Bayesian Classification
  4. Support Vector Machines

➤ Inductive Learning and Bias

- Suppose that we want to learn a function f(x) a y and we are given some sample (x,y) pairs, as in Fig. 3.3.3 (a). There are several hypotheses we could make about this function, e.g : Fig. 3.3.3 (b), (c) and (d).

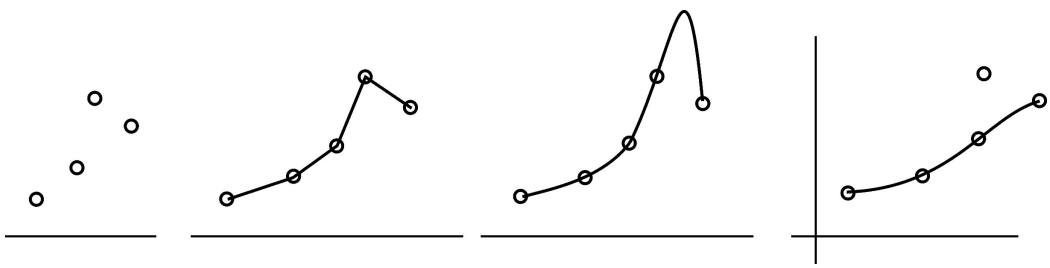


Fig. 3.3.3

- A preference for one over the others reveals the bias of our learning technique, e.g. :
  i) Prefer piece-wise functions
  ii) Prefer a smooth function
  iii) Prefer a simple function and treat outliers as noise
- **Preference Bias :** The simplest explanation that is consistent with all observations is the best. Here, that means the smallest decision tree that correctly classifies all of the training examples is best.
- Finding the provably smallest decision tree is an NP-Hard problem, so instead of constructing the absolute smallest tree that is consistent with all of the training examples, construct one that is pretty small.

- **Training Set Error :** For each record, follow the decision tree to see what it would predict. For what number of records does the decision tree's prediction disagree with the true value in the database ? This quantity is called the training set error. The smaller the better.
- **Test Set Error :** We hide some data away when we learn the decision tree. But once learned, we see how well the tree predicts that data. This is a good simulation of what happens when we try to predict future data. It is called Test Set Error.

➥ 3.3.2 Advantages and Disadvantages of Decision Trees

➤ Advantages :

1. Decision trees can handle both nominal and numeric input attributes.
2. Decision tree representation is rich enough to represent any discrete value classifier.
3. Decision trees are capable of handling datasets that may have errors.
4. Decision trees are capable of handling datasets that may have missing values.
5. It is self-explanatory and when compacted they are also easy to follow.

➤ Disadvantages

1. Most of the algorithms require that the target at-tribute will have only discrete values.
2. Most decision-tree algorithms only examine a single field at a time.
3. Decision trees are prone to errors in classification problems with many class.
4. As decision trees use the "divide and conquer" method, they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present.

Example 3.3.1 : Consider the following six training examples, where each example has three attributes : color, shape and size. Color has three possible values : red, green and blue. Shape has two possible values : square and round. Size has two possible values : big and small.

Example	Color	Shape	Size	Class
1	red	square	big	+
2	blue	square	big	+
3	red	round	small	–
4	green	square	small	–
5	red	round	big	+
6	green	square	big	–

Which is best attribute for the root node of decision tree ?

▣ **Solution :**

$$H(\text{class}) = H(3/6, 3/6) = 1$$

$$H(\text{class} \mid \text{color}) = 3/6 * H(2/3, 1/3) + 1/6\ H(1/1, 0/1) + 2/6\ H(0/2, 2/2)$$

			1 of 6 is blue	2 of 6 are green
		1 of the red is negative		
	2 of the red are positive			

3 out of 6 are red

$$= 1/2 * (-2/3 \log_2 2/3 - 1/3 \log_2 1/3) + 1/6 * (-1 \log_2 1 - 0 \log_2 0) + 2/6 * (-0 \log_2 0 - 1 \log_2 1)$$

$$= 1/2 * (-2/3(\log_2 2 - \log_2 3) - 1/3(\log_2 1 - \log_2 3)) + 1/6 * 0 + 2/6 * 0$$

$$= 1/2 * (-2/3(1 - 1.58) - 1/3(0 - 1.58))$$

$$= 1/2 * 0.914$$

$$= 0.457$$

$$I(\text{class; color}) = H(\text{class}) - H(\text{class} \mid \text{color}) = 1.0 - 0.457$$

$$= 0.543$$

$$H(\text{class} \mid \text{shape}) = 4/6\ I(2/4, 2/4) + 2/6\ I(1/2, 1/2) = 4/6 * 1.0 + 2/6 * 1.0$$

$$= 1.0$$

$$I(\text{class; shape}) = H(\text{class}) - H(\text{class} \mid \text{shape}) = 1.0 - 1.0$$

$$= 0.0$$

$$H(\text{class} \mid \text{size}) = 4/6\ I(3/4, 1/4) + 2/6\ I(0/2, 2/2)$$

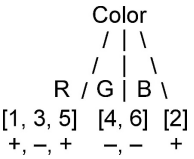
$$= 0.541$$

$$I(\text{class; size}) = H(\text{class}) - H(\text{class} \mid \text{size}) = 1.0 - 0.541$$

$$= 0.459$$

Max(0.543, 0.0, 0.459) = 0.543, so color is best. Make the root node's attribute color and partition the examples for the resulting children nodes as shown :

The children associated with values green and blue are uniform, containing only – and + examples, respectively. So make these children leaves with classifications – and +, respectively.



➥ 3.3.3 K-NN Classifier

- The K-nearest neighbor (KNN) is a classical classification method and requires no training effort, critically depends on the quality of the distance measures among examples.
- The KNN classifier uses Mahalanobis distance function. A sample is classified according to the majority vote of its nearest K training samples in the feature space. Distance of a sample to its neighbors is defined using a distance function.
- For all points x, y and z, a distance function F(., .), must satisfy the following conditions :

1.	Nonnegativity	$F(x, y) \geq 0$
2.	Reflexivity	$F(x, y) = 0$ if and only if $x = y$
3.	Symmetry	$F(x, y) = F(y, x)$
4.	Triangle inequality	$F(x, y) + F(y, x) \geq F(x, y)$

➤ The Euclidean distance

- The Euclidean distance is the most common distance metric used in low dimensional data sets. It is also known as the **L_2 norm**. The Euclidean distance is the usual manner in which distance is measured in real world.

$$d_{euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

where x and y are m-dimensional vectors and denoted by $x = \{x_1, x_2, x_3, \ldots, x_m\}$ and $y = \{y_1, y_2, y_3, \ldots, y_m\}$ represent the m attribute values of two records.

- While Euclidean metric is useful in low dimensions, it doesn't work well in high dimensions and for categorical variables. The drawback of Euclidean distance is that it ignores the similarity between attributes. Each attribute is treated as totally different from all of the attributes.

➤ Mahalanobis distance

- Mahalanobis distance is also called quadratic distance.
- Mahalanobis distance is a distance measure between two points in the space defined by two or more correlated variables. Mahalanobis distance takes the correlations within a data set between the variableinto considerations.
- If there are two non-correlated variables, the Mahalanobis distance between the points of the variable in a 2D scatter plot is same as Euclidean distance.
- The Mahalanobis distance is the distance between an observation and the center for each group in m-dimensional space defined by m variables and their covariance. Thus, a small value of Mahalanobis distance increases the chance of an observation to be closer to the group's center and the more likely it is to be assigned to that group.
- Mahalanobis distance between two samples (x, y) of a random variable is defined as

$$d_{\text{Mahalanobis}}(x, y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$$

- The Mahalanobis metric is defined in independence of the data matrix.
- No pre-processing of labeled data samples is needed before using KNN algorithm. A dominated class label in K-nearest neighbors of a data point is assigned as class label to that data point. A tie occurs when neighborhood has same amount of labels from multiple classes.
- To break the tie, the distances of neighbors can be summed up in each class that is tied and vector f is assigned to the class with minimal distance. Or, the class can be chosen with the nearest neighbor. Clearly, tie is still possible here, in which case an arbitrary assignment is taken.
- Three distance functions that can be used in KNN classifier are :

1.	L_p norm	$L_p(x, y) = \left(\sum_{i=1}^{d} \lvert x_i - y_i \rvert^p \right)^{1/p}$
2.	L_2 norm (Euclidean distance)	$L_2(x, y) = \left(\sum_{i=1}^{d} \lvert x_i - y_i \rvert^2 \right)^{1/2}$
3.	L_1 norm (Manhattan distance)	$L_1(x, y) = \sum_{i=1}^{d} \lvert x_i - y_i \rvert$

- Mahalanobis distance that takes into account the correlation S of the dataset :

$$L_m(x, y) = \sqrt{(x-y)\, S^{-1} (x-y)}$$

- The steps that need to be carried out during the KNN algorithm are as follows :
  a. Divide the data into training and test data.
  b. Select a value K.
  c. Determine which distance function is to be used.
  d. Choose a sample from the test data that needs to be classified and compute the distance to its n training samples.
  e. Sort the distances obtained and take the k-nearest data samples.
  f. Assign the test class to the class based on the majority vote of its K neighbors.
- Fig. 3.3.4 shows geometrical representation of K-nearest neighbor algorithm.

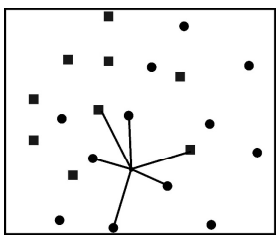


Fig. 3.3.4

- The performance of the KNN algorithm is influenced by three main factors :
  1. The distance function or distance metric used to determine the nearest neighbors.
  2. The decision rule used to derive a classification from the K-nearest neighbors.
  3. The number of neighbors used to classify the new example.

➤ **Advantages**

1. The KNN algorithm is very easy to implement.
2. Nearly optimal in the large sample limit.
3. Uses local information, which can yield highly adaptive behavior.
4. Lends itself very easily to parallel implementations.

➤ **Disadvantages**

1. Large storage requirements.
2. Computationally intensive recall.
3. Highly susceptible to the curse of dimensionality.

3.3.4 SVM Classifier

- Support Vector Machines (SVMs) are a set of supervised learning methods which learn from the dataset and used for classification. SVM is a classifier derived from statistical learning theory by Vapnik and Chervonenkis.
- An SVM is a kind of large-margin classifier : it is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data.
- Given a set of training examples, each marked as belonging to one of two classes, an SVM algorithm builds a model that predicts whether a new example falls into one class or the other. Simply speaking, we can think of an SVM model as representing the examples of the separate classes are divided by a gap that is as wide as possible.
- New examples are then mapped into the same space and classified to belong to the class based on which side of the gap they fall on.
- SVM are primarily two-class classifiers with the distinct characteristic that they aim to find the optimal hyperplane such that the expected generalization error is minimized. Instead of directly minimizing the empirical risk calculated from the training data, SVMs perform structural risk minimization to achieve good generalization.
- The empirical risk is the average loss of an estimator for a finite set of data drawn from P. The idea of risk minimization is not only measure the performance of an estimator by its risk, but to actually search for the estimator that minimizes risk over distribution P. Because we don't know distribution P we instead minimize empirical risk over a training dataset drawn from P. This general learning technique is called **empirical risk minimization**.
- Fig. 3.3.5 shows empirical risk.

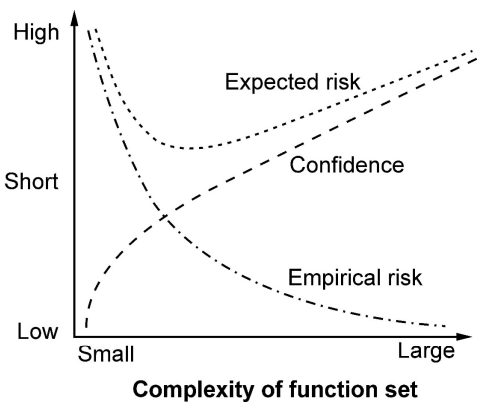


Fig. 3.3.5 : Empirical risk

➤ Key Properties of Support Vector Machines

1. Use a single hyperplane which subdivides the space into two half-spaces, one which is occupied by Class 1 and the other by Class 2.
2. They maximize the margin of the decision boundary using quadratic optimization techniques which find the optimal hyperplane.
3. Ability to handle large feature spaces.
4. Overfitting can be controlled by soft margin approach
5. When used in practice, SVM approaches frequently map the examples to a higher dimensional space and find margin maximal hyperplanes in the mapped space, obtaining decision boundaries which are not hyperplanes in the original space.
6. The most popular versions of SVMs use non-linear kernel functions and map the attribute space into a higher dimensional space to facilitate finding "good" linear decision boundaries in the modified space.

➤ Soft Margin SVM

- For the very high dimensional problems common in text classification, sometimes the data are linearly separable. But in the general case they are not, and even if they are, we might prefer a solution that better separates the bulk of the data while ignoring a few weird noise documents.
- What if the training set is not linearly separable ? *Slack variables* can be added to allow misclassification of difficult or noisy examples, resulting margin called *soft*.
- A *soft-margin* allows a few variables to cross into the margin or over the hyperplane, allowing misclassification.
- We penalize the crossover by looking at the number and distance of the misclassifications. This is a trade-off between the hyperplane violations and the margin size. The slack variables are bounded by some set cost. The farther theyare from the soft margin, the less influence they have on the prediction.
- All observations have an associated slack variable.
  1. Slack variable = 0 then all points on the margin.
  2. Slack variable > 0 then a point in the margin or on the wrong side of the hyperplane.
  3. C is the tradeoff between the slack variable penalty and the margin.

➤ Limitations of SVM

1. It is sensitive to noise.
2. The biggest limitation of SVM lies in the choice of the kernel.
3. Another limitation is speed and size.
4. The optimal design for multiclass SVM classifiers is also a research area.

University Question

1. Explain the process of choosing K in K - nearest neighbour clustering.

AU : Dec.-17, Marks 8

3.4 Feature Selection or Dimensionality Reduction

- In machine learning, "dimensionality" simply refers to the number of features (i.e. input variables) in your database.
- When the number of features is very large relative to the number of observations in your dataset, certain algorithms struggle to train effective models. This is called the "Curse of Dimensionality," and its especially relevant for clustering algorithms that rely on distance calculations.
- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into features selection and feature extraction.
- Classification problem example : We have an input data $\{ X_1, X_2, X_3, ..., X_N \}$ such that $x_i = \left(x_i^1, x_i^2, ..., x_i^d \right)$ and a set of corresponding output labels. Assume the dimension d of the data point x is very large and we want to classify x.
- Problem with high dimensional input vectors are large number of parameters to learn, if a dataset is small, this can result in overfit and large variance of estimates.

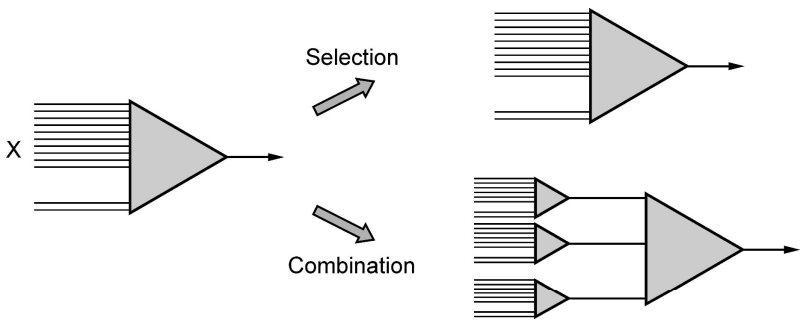


Fig. 3.4.1 : Dimensionality reduction

- Solution to this problem is as follows :
  1. Selection of a smaller subset of inputs from a large set of inputs; train classifier on the reduced input set.
  2. Combination of high dimensional inputs to a smaller set of features $\phi_k(x)$; train classifier on new features.

- There are two components of dimensionality reduction :
  1. **Feature selection :** User try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways : Filter, Wrapper and Embedded.
  2. **Feature extraction :** This reduces the data in a high dimensional space to a lower dimensions space i.e. a space with lesser number of dimensions.
- There are many methods to perform dimension reduction.
  1. **Missing Values :** While exploring data, if we encounter missing values, what we do ? Our first step should be to identify the reason then impute missing values / drop variables using appropriate methods. But, what if we have too many missing values ? Should we impute missing values or drop the variables ?
  2. **Low Variance :** Let's think of a scenario where we have a constant variable in our data set.
  3. **Decision Trees :** It can be used as a ultimate solution to tackle multiple challenges like missing values, outliers and identifying significant variables.
  4. **Random Forest :** Similar to decision tree is Random Forest.
  5. **High Correlation :** Dimensions exhibiting higher correlation can lower down the performance of model. Moreover, it is not good to have multiple variables of similar information or variation also known as "Multicollinearity".
  6. **Backward Feature Elimination :** In this method, we start with all n dimensions. Compute the sum of square of error (SSR) after eliminating each variable (n times). Then, identifying variables whose removal has produced the smallest increase in the SSR and removing it finally, leaving us with n – 1 input features.

➤ Advantages of Dimensionality Reduction

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.

➤ Disadvantages of Dimensionality Reduction

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.
- We may not know how many principal components to keep in practice, some thumb rules are applied.

➥ 3.4.1 TF-IDF Weighting

- **Term Frequency (TF) :** Frequency of occurrence of query keyword in document.
- **Inverse Document Frequency (IDF) :** How many documents the query keyword occurs in.
- Inverse Document Frequency (IDF) is a popular measure of a word's importance. It's defined as the logarithm of the ratio of number of documents in a collection to the number of documents containing the given word. This means rare words have high IDF and common words have low IDF.
- Term frequency is a measure of the importance of terms i in document j.
- Inverse document frequency is a measure of the general importance of the term.
- High term frequency for "apple" means that apple is an important word in a specific document. But high document frequency (low inverse document frequency) for "apple", given a particular set of documents, means that apple is not all that important overall, since it is in all of the documents.
- The weight increases as the number of documents in which the term appears decreases. High value indicates that the word occurs more often in this document than average.
- The term frequency $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.
- A document with $tf_{t,d} = 10$ occurrences of the term is more relevant than a document with $tf_{t,d} = 1$ occurrence of the term. But not 10 times more relevant. Relevance does not increase proportionally with term frequency.
- The document frequency is the number of documents in the collection that the term occurs in. We define the idf weight of term t as follows :

$$\text{idf weight } (idf_t) = \log 10 \frac{N}{df_t}$$

  here N is the number of documents in the collection.

- The tf-idf weight of a term is the product of its tf weight and its idf weight

$$W_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

➤ Stop lists and Stemming :

- **Stoplists :** This is a list of words that we should ignore when processing documents, since they give no useful information about content.
- **Stemming :** This is the process of treating a set of words like "fights, fighting, fighter, …" as all instances of the same term - in this case the stem is "fight".

3.4.2 Information Gain

- Entropy measures the impurity of a collection. Information gain is defined in terms of entropy.
- Information gain tells us how important a given attribute of the feature vectors is.
- Information gain of attribute A is the reduction in entropy caused by partitioning the set of examples S.

$$\text{Gain (S, A)} = \text{Entropy (S)} - \sum_{v \in \text{Values}} \frac{|S_v|}{S} \text{Entropy } (S_v)$$

where Values (A) is the set of all possible values for attribute A and S_v is the subset of S for which attribute A has value v.

Pruning by Information Gain :

- The simplest technique is to prune out portions of the tree that result in the least information gain.
- This procedure does not require any additional data, and only bases the pruning on the information that is already computed when the tree is being built from training data.
- The process of information gain based pruning requires us to identify "twigs", nodes whose children are all leaves.
- "Pruning" a twig removes all of the leaves which are the children of the twig, and makes the twig a leaf. The Fig. 3.4.2 illustrates this.

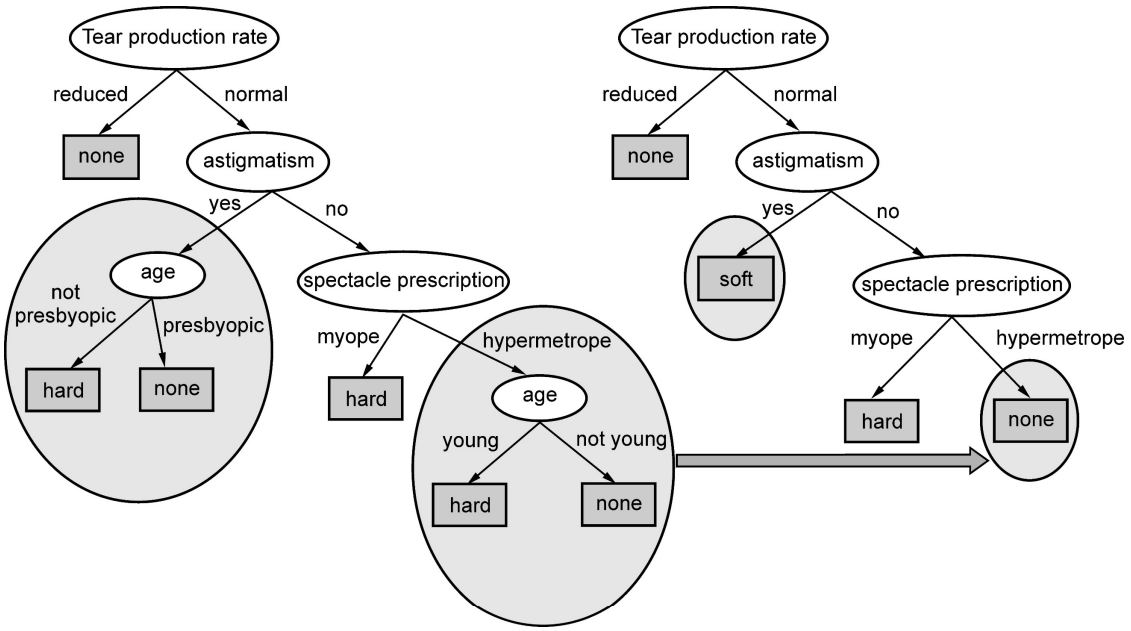


Fig. 3.4.2

- The algorithm for pruning is as follows :
  1. Catalog all twigs in the tree.
  2. Count the total number of leaves in the tree.
  3. While the number of leaves in the tree exceeds the desired number :
     a. Find the twig with the least Information Gain.
     b. Remove all child nodes of the twig.
     c. Relabel twig as a leaf.
     d. Update the leaf count.

3.5 Evaluation Metrics

3.5.1 Contingency Table

- A binary classification rule is a method that assigns a class to an object, on the basis of its description.
- The performance of a binary classifier can be assessed by tabulating its predictions on a test set with known labels in a contingency table or confusion matrix, with actual classes in rows and predicted classes in columns.
- Measures of performance need to satisfy several criteria :
  1. They must coherently capture the aspect of performance of interest;
  2. They must be intuitive enough to become widely used, so that the same measures are consistently reported by researchers, enabling community-wide conclusions to be drawn;
  3. They must be computationally tractable, to match the rapid growth in the scale of modern data collection.
  4. They must be simple to report as a single number for each method-dataset combination.
- Performance metrics for binary classification are designed to capture trade-offs between four fundamental population quantities : true positives, false positives, true negatives and false negatives.
- The evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, named confusion matrix.
- Confusion matrix is also called a contingency table.
  1. **False positives :** Examples predicted as positive, which are from the negative class.
  2. **False negatives :** Examples predicted as negative, whose true class is positive.

3. **True positive :** Examples correctly predicted as pertraining to the positive class.
4. **True negatives :** Examples correctly predicted as belonging to the negative class.

3.5.2 Accuracy and Error

- **Accuracy :** Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
- The evaluation measure most used in practice is the accuracy rate. It evaluates the effectiveness of the classifier by its percentage of correct predictions.

$$\text{Accuracy rate} = \frac{|\text{True negatives}| + |\text{True positives}|}{|\text{False negatives}| + |\text{False positives}| + |\text{True negatives}| + |\text{True positives}|}$$

- The complement of accuracy rate is the error rate, which evaluates a classifier by its percentage of incorrect predictions.

$$\text{Error rate} = \frac{|\text{False negatives}| + |\text{False positives}|}{|\text{False negatives}| + |\text{False positives}| + |\text{True negatives}| + |\text{True positives}|}$$

$$\text{Error rate} = 1 - (\text{Accuracy rate})$$

- The recall and specificity measures evaluate the effectiveness of a classifier for each class in the binary problem. The recall is also known as sensitivity or true positive rate. **Recall** is the proportion of examples belonging to the positive class which were correctly predicted as positive.
- The **specificity** is a statistical measure of how well a binary classification test correctly identifies the negative cases.

$$\text{Recall (R)} = \frac{|\text{True positive}|}{|\text{True positive}| + |\text{False negative}|}$$

$$\text{Specificity} = \frac{|\text{True negatives}|}{|\text{False positives}| + |\text{True negatives}|}$$

- True Positive Rate (TPR) is also called sensitivity, hit rate, and recall.

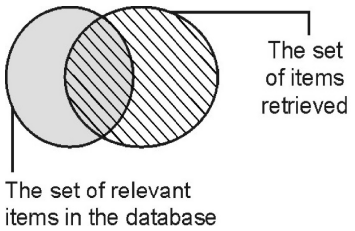
$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negative}}$$

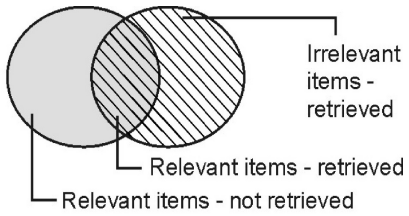
- A statistical measure of how well a binary classification test correctly identifies a condition. Probability of correctly labelling members of the target class.
- No single measure tells the whole story. A classifier with 90 % accuracy can be useless if 90 percent of the population does not have cancer and the 10 % that do are misclassified by the classifier. Use of multiple recommended.

- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predictive model.

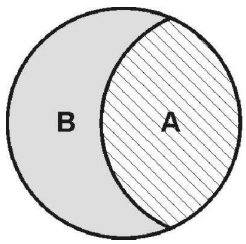
➥ 3.5.3 Precision and Recall

- **Relevance :** Relevance is a subjective notion. Different users may differ about the relevance or non-relevance of particular documents to given questions.
- In response to a query, an IR system searches its document collection and returns a ordered list of responses. It is called the retrieved set or ranked list. The system employs a search strategy or algorithm and measure the quality of a ranked list.
- A better search strategy yields a better ranked list and better ranked lists help the user fill their information need.
- Precision and recall are the basic measures used in evaluating search strategies. As shown in the first two figures, these measures assume :
  1. There is a set of records in the database which is relevant to the search topic.
  2. Records are assumed to be either relevant or irrelevant.
  3. The actual retrieval set may not perfectly match the set of relevant records.





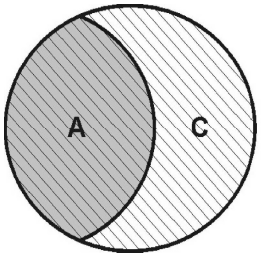
- **Recall** is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.



A = Number of relevant records retrieved.
B = Number of relevant records not retrieved.

$$\text{Recall} = \frac{A}{A+B} \times 100\ \%$$

- **Precision** is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.



A = Number of relevant records retrieved.
C = Number of irrelevant records retrieved

$$\text{Precison} = \frac{A}{A+C} \times 100\ \%$$

- As recall increases, the precision decreases and recall decreases the precision increases.

Example 3.5.1 : Assume the following :

A database contains 80 records on a particular topic. A search was conducted on that topic and 60 records were retrieved of the 60 records retrieved, 45 were relevant. Calculate the precision and recall scores for the search.

▣ **Solution :** Using the designations above :

A = The number of relevant records retrieved,

B = The number of relevant records not retrieved, and

C = The number of irrelevant records retrieved.

In this example A = 45, B = 35 (80 – 45) and C = 15 (60 – 45).

$$\text{Recall} = \frac{45}{45+35} \times 100\ \%$$

$$\text{Recall} = \frac{45}{80} \times 100\ \%$$

$$\textbf{Recall} = \textbf{56.25 \%}$$

$$\text{Precision} = \frac{A}{A+C} \times 100\ \%$$

$$\text{Precision} = \frac{45}{45+15} \times 100\ \% = \frac{45}{60} \times 100$$

$$\textbf{Precision} = \textbf{75 \%}$$

Example 3.5.2 : 20 found documents, 18 relevant, 3 relevant documents are not found, 27 irrelevant are as well not found. Calculate the precision and recall and fallout scores for the search.

▣ Solution :

Precision : 18/20 = 90 %

Recall : 18/21 = 85.7 %

Fall-out : 2/29 = 6.9 %

- Recall is a non-decreasing function of the number of docs retrieved. In a good system, precision decreases as either the number of docs retrieved or recall increases. This is not a theorem, but a result with strong empirical confirmation.
- The set of ordered pairs makes up the precision-recall graph. Geometrically when the points have been joined up in some way they make up the precision-recall curve. The performance of each request is usually given by a precision-recall curve. To measure the overall performance of a system, the set of curves, one for each request, is combined in some way to produce an average curve.
- Assume that set R_q containing the relevant document for q has been defined. Without loss of generality, assume further that the set R_q is composed of the following documents :

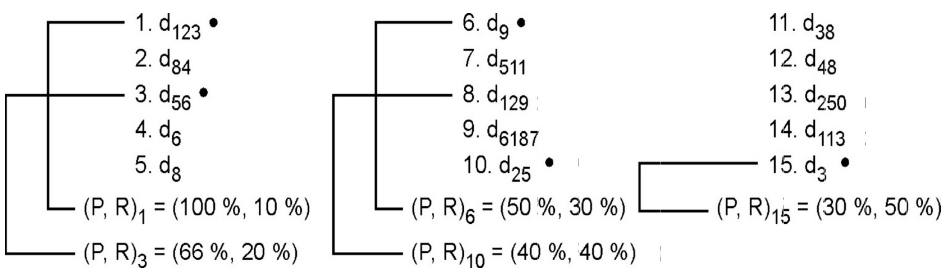
  $R_q = \{ d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123} \}$

  There are ten documents which are relevant to the query q.
- For the query q, a ranking of the documents in the answer set as follows.

  Ranking for query q :

1. d_{123} *	6. d_9 *	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{59} *	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25} *	15. d_3 *

- The documents that are relevant to the query q are marked with star after the document number. Ten relevant documents, five included in Top 15.



- Fig. 3.5.1 shows the curve of precision versus recall. By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a precision-recall curve.
- The precision versus recall curve is usually plotted based on 11 standard recall level : 0 %, 10 %,…., 100 %.

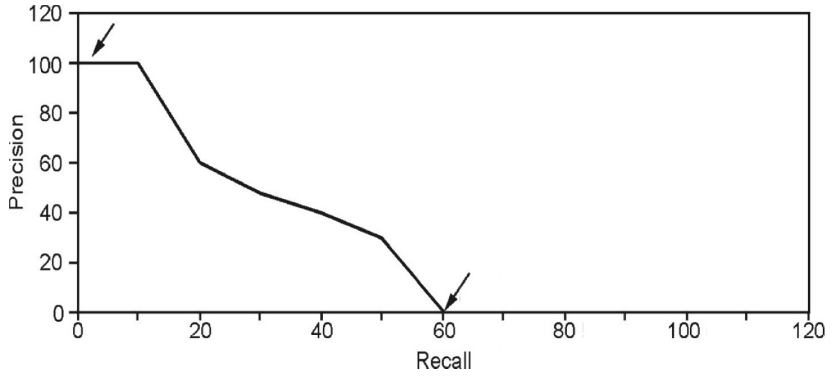
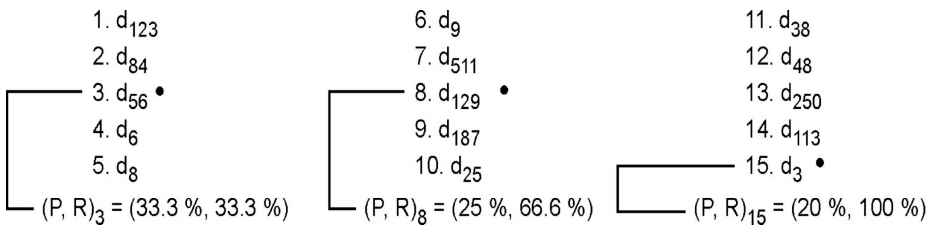


Fig. 3.5.1 : Precision versus recall curve

- In this example : The precisions for recall levels higher than 50 % drop to 0 because no relevant documents were retrieved. There was an interpolation for the recall level 0 %.
- Since the recall levels for each query might be distinct from the 11 standard recall levels.

3.5.3.1 Interpolated Recall-Precision

- Idea : If locally precision increases with increasing recall, then you should get to count that. So you take the max of precisions to right of value.
- Consider again the set of 15 ranked documents. Assume that the set of relevant documents for the query q has changed and is now given by $R_q = \{ d_3, d_{56}, d_{129} \}$.
- The three relevant document :



- Interpolated precisions at standard recall levels :

$$\bar{P}(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

The j^{th} standard recall level.

- Which means that the interpolated precision at the j^{th} standard recall level is the maximum known precision at any recall level between the j^{th} recall level and the $(j + 1)^{th}$ recall level.
- At recall levels 0 %, 10 %, 20 % and 30 %, the interpolated precision is equal to 33.3 %.
- At recall levels 40 %, 50 %, 60 % the interpolated precision is equal to 25 %.

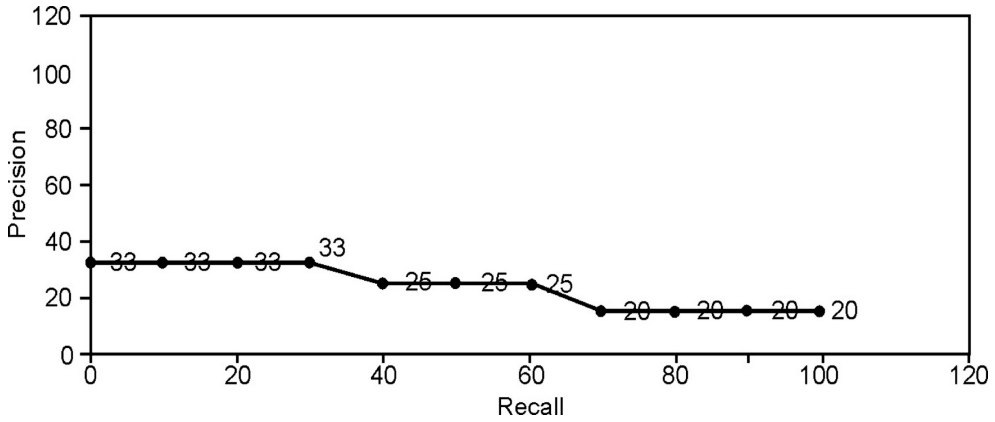


Fig. 3.5.2 : Curve for interpolated precision and recall

- At recall levels 70 %, 80 %, 90 % and 100 %, the interpolated precision is equal to 20 %.
- Fig. 3.5.2 shows the curve for interpolated precision and recall.
- Following Fig. 3.5.3 shows the comparison between precision-recall curve and interpolated precision.

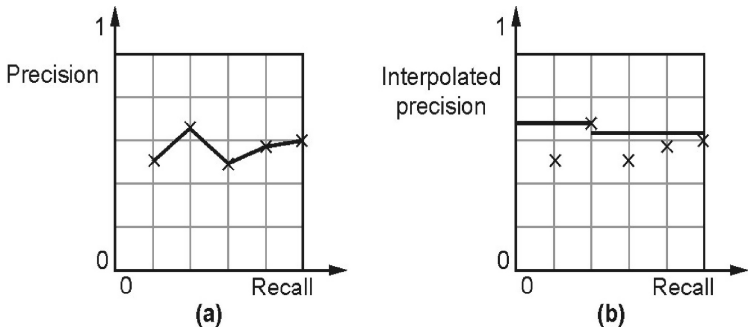


Fig. 3.5.3

➤ **Advantages of interpolated recall-precision**

1. Simple, intuitive, and combined in single curve.
2. Provide quantitative evaluation of the answer set and comparison among retrieval algorithms.
3. A standard evaluation strategy for IR systems.

➤ **Disadvantages of interpolated recall-precision**

1. Can not know true recall value except in small document collections.
2. Assume a strict document rank ordering.

- It is an experimental fact that average precision-recall graphs are monotonically decreasing. Consistent with this, a linear interpolation estimates the best possible performance between any two adjacent observed points. To avoid inflating the experimental results it is probably better to perform a more conservative interpolation.

➥ **3.5.3.2 Mean Average Precision (MAP)**

- Also called average precision at seen relevant documents. It determine precision at each point when a new relevant document gets retrieved.
- Average of the precision value obtained for the top k documents, each time a relevant doc is retrieved.
- Avoids interpolation, use of fixed recall levels. MAP for query collection is arithmetic averaging.
- Average precision-recall curves are normally used to compare the performance of distinct IR algorithms.
- Use P = 0 for each relevant document that was not retrieved. Determine average for each query, then average over queries :

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j} P(doc_i)$$

where

Q_j = Number of relevant document for query j.

N = Number of queries.

$P(doc_i)$ = Precision at i^{th} relevant document

Example 3.5.3 :

Query 1		
Rank	**Relev.**	**P(doc$_i$)**
1	X	1.00
2		
3	X	0.67
4		
5		
6	X	0.50
7		
8		
9		
10	X	0.40
11		
12		
13		
14		
15		
16		
17		
18		
19		
20	X	0.50
AVG :		0.564

Query 2		
Rank	**Relev.**	**P(doc)**
1	X	1.00
2		
3	X	0.67
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15	X	0.2
AVG :		0.623

MAP favors systems which return relevant documents fast.

▣ **Solution :**

$$MAP = \frac{0.564 + 0.623}{2}$$

$$MAP = 0.594$$

- A necessary consequence of its monotonicity is that the average P-R curve will also be monotonically decreasing. It is possible to define the set of observed points in such a way that the interpolate function is not monotonically decreasing. In practice, even for this case, we have that the average precision-recall curve is monotonically decreasing.

➤ Precision-recall appropriateness

- Precision and recall have been extensively used to evaluate the retrieval performance of IR algorithms. However, a more careful reflection reveals problems with these two measures :
- First, the proper estimation of maximum recall for a query requires detailed knowledge of all the documents in the collection.
- Second, in many situations the use of a single measure could be more appropriate.
- Third, recall and precision measure the effectiveness over a set of queries processed in batch mode.
- Fourth, for systems which require a weak ordering though, recall and precision might be inadequate.

➤ Single value summaries

- Average precision-recall curves constitute standard evaluation metrics for information retrieval systems. However, there are situations in which we would like to evaluate retrieval performance over individual queries. The reasons are twofold :
  1. First, averaging precision over many queries might disguise important anomalies in the retrieval algorithms under study.
  2. Second, we might be interested in investigating whether a algorithm outperforms the other for each query.
- In these situations, a single precision value can be used.

➤ R-Precision

- If we have a known set of relevant documents of size Rel, then calculate precision of the top Rel docs returned.
- Let R be the total number of relevant docs for a given query. The idea here is to compute the precision at the Rth position in the ranking.
- For the query q_1, the R value is 10 and there are 4 relevant among the top 10 documents in the ranking. Thus, the R-Precision value for this query is 0.4.
- The R-precision measure is a useful for observing the behavior of an algorithm for individual queries. Additionally, one can also compute an average R-precision figure over a set of queries.

- However, using a single number to evaluate a algorithm over several queries might be quite imprecise.

➤ **Precision histograms**

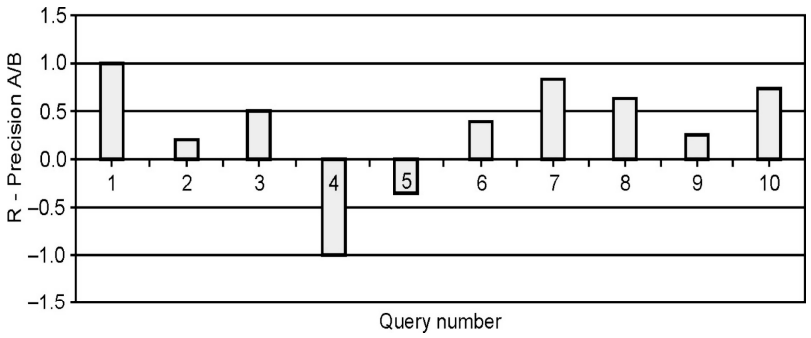


Fig. 3.5.4 : Precision histograms

- The R-precision computed for several queries can be used to compare two algorithms as follows :
- Let,

  $RP_A(i)$: R-precision for algorithm A for the i-th query.

  $RP_B(i)$: R-precision for algorithm B for the i-th query.

- Define, for instance, the difference :

  $$RP_{A/B}(i) = RP_A(i) - RP_B(i)$$

- A positive value of $RP_{A/B}(i)$ indicates a better retrieval performance by algorithm A while a negative value indicates a better retrieval performance by algorithm B. Fig. 3.5.4 shows the $RP_{A/B}(i)$ values for two retrieval algorithms over 10 example queries.
- The algorithm A performs better for 8 of the queries, while the algorithm B performs better for the other 2 queries.

3.6 Organizing the Classes

- A taxonomy is a "subject map" to an organization's content. A taxonomy reflects the organization's purpose or industry, the functions and responsibilities of the persons or groups who need to access the content, and the purposes/reasons for accessing the content.
- Fig. 3.6.1 shows taxonomy process.

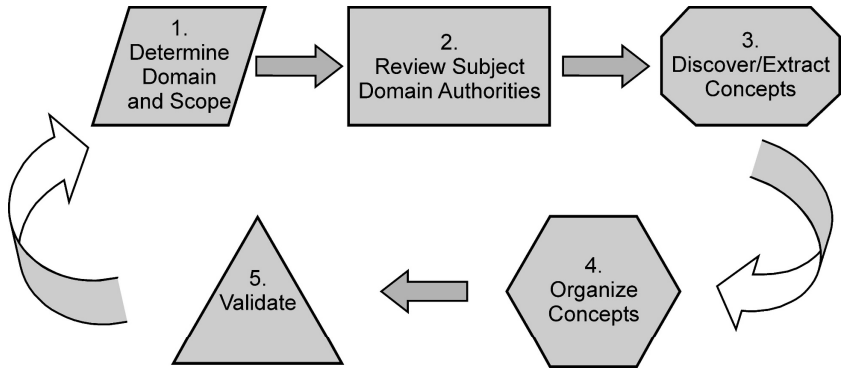


Fig. 3.6.1 : Taxonomy process

- Taxonomy is the technical term for the guiding principles behind the organisation of information a key concern for web developers.
- Business context is the business environment for the taxonomy efforts in terms of business objectives, Web applications where taxonomy will be used, corporate culture, past or current taxonomy initiatives, and artifacts within the organisation and across the industry.
- Users refers to the target audience for the taxonomy, user profiles, and user characteristics in terms of information usage patterns.
- Content is the type of information that will be covered by the taxonomy or that the taxonomy will be built upon.

3.7 Indexing and Searching

- Indexing is the process of creating a representation of a document, primarily of its topic or content, although formal representation, of elements such as authorship, title, bibliographic context etc., is sometimes included in the term.
- The index language which comes out of the conflation algorithm and this index language is used to describe documents and requests. The elements of the index language are index terms, which may be derived from the text of the document to be described.
- For example, preprocessing of document i.e. raw documents must be converted into bag of terms representation. These expressions are sometimes called **document representatives**.
- To make these representatives for each document we first collect the words and create the file containing words except the stop words, then we stem the words with in a file thus we have all the terms important to our search in their root forms.
- After this we count the frequency of each word and the word having frequency above a threshold (based on a formula consisting file size) is selected as an index term. Collection of all such terms creates our index table (document representative) for that document.

- Same process will be repeated for phrase based search but the difference is that all work will be done on phrases identified by a phrase identification program which identify phrases by counting their frequencies; phrase having a suitable count will be selected for processing. Thus an index table will be generated containing phrases as its indices.
- Automatic indexing is the process of representing the topical content of digitally stored documents or document surrogates, performed without, or with very modest, human intervention.

➥ 3.7.1 Inverted Indexes

- Each document is assigned a list of keywords or attributes. Each keyword (attribute) is associated with operational relevance weights.
- An inverted file is the sorted list of keywords (attributes), with each keyword having links to the documents containing that keyword.
- **Penalty :** The size of inverted files ranges from 10 % to 100 % of more of the size of the text itself. It need to update the index as the data set changes.
- Inverted file is composed of two elements :

  a. Vocabulary b. Occurrences
- **Vocabulary** is the set of all different words in the text. For each such word a list of all the text positions where the word appears is stored. The set of all those lists is called the occurrences .
- A controlled vocabulary which is the collection of keywords that will be indexed. Words in the text that are not in the vocabulary will not be indexed. A list of stop-words that for reasons of volume will not be included in the index.
- A set of rules that decide the beginning of a word or a piece of text that is indexable. A list of character sequences to be indexed (or not indexed).
- Fig. 3.7.1 shows a sample text with inverted index file. Each entry in the vocabulary has **the word**, a **pointer into the postings structure** and **word metadata**.

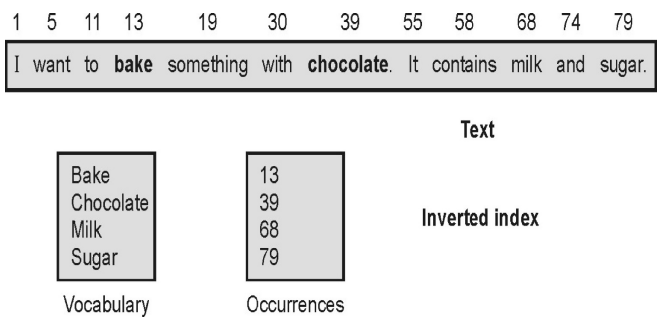


Fig. 3.7.1 : Sample text with inverted index file

- Every query goes here first
  a. Keep in memory or a separate file
  b. Search should be fast
  c. Support prefix or pattern matching
  d. Support updating
- The positions can refer to words or characters. The word positions simplify phrase and proximity queries, while character positions facilities direct access to the matching text positions. Time needed to access posting lists is a function of their length and their allocation.
- The space required for the **vocabulary** is rather small. According to Heaps law the vocabulary grows as O(nB), where B is a constant between 0 and 1 dependent on the text. The **occurrences** required much more space. Each word appearing in the text is referenced once in that structure, the extra space is O(n).
- Block addressing is used to reduce the space requirement. The text is divided in blocks and the occurrences point to the blocks where the word appears. The classical indices which point to the exact occurrences are called **full inverted indices**.
- The definition of inverted files does not require that the addresses in the directory are in any order. However, to facilitate operations such as conjunction ('and') and disjunction ('or') on any two inverted lists, the addresses are normally kept in record number order. This means that 'and' and 'or' operations can be performed with one pass through both lists. The penalty we pay is of course that the inverted file becomes slower to update.
- Fig. 3.7.2 shows the sample text split into blocks and an inverted index using block addressing built on it.

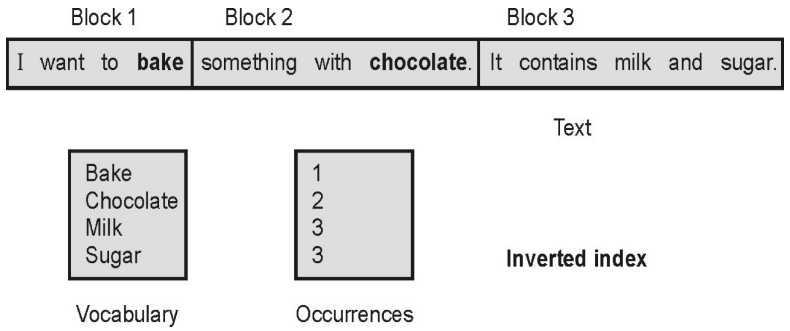


Fig. 3.7.2 : Sample text split into blocks and an inverted index using block addressing

- The blocks can be of fixed size or they can be defined using the natural division of the text collection into files, documents, web pages etc. Concept of block improves the retrieval efficiency.

3.7.2 Searching

- The search algorithm on an inverted index follows three steps :
  1. **Vocabulary search :** The words and pattern presents in the query are isolated and searched in the vocabulary.
  2. **Retrieval of occurrences :** The list of the occurrences of all the words found is retrieved.
  3. **Manipulation of occurrences :** The occurrences are processed to solve phrases, proximity or Boolean operations. If block addressing is used it may be necessary to directly search the text to find the information missing from the occurrences.
- Structures used in inverted files are sorted arrays, hashing structures, Tries (digital search trees) and combinations of these structures. Single word queries can be searched using any suitable data structure to speed up the search.
- If the index stores character positions the phrase query cannot allow the separators to be disregarded and the proximity has to be defined in terms of character distance.

3.7.3 Construction

- Building and maintaining an inverted index is a relatively low cost task. An inverted index on a text of "n" characters can be built in O(n) time. All the vocabulary is kept into the trie data structure, storing for each word a list of its occurrences.
- Each word of the text is read and searched in the trie. If it is found, it is added to the trie with an empty list of occurrences. Once it is in the trie, the new position is added to the end of its list of occurrences. Fig. 3.7.3 shows building an inverted index for the sample text.

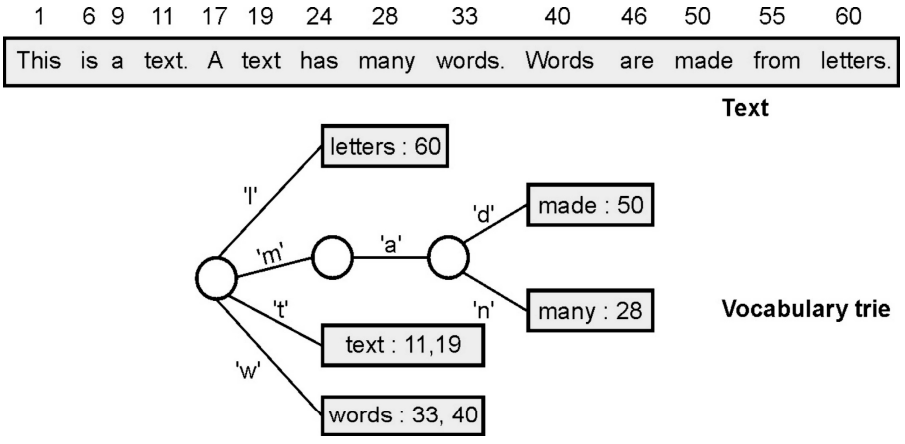


Fig. 3.7.3 : Building an inverted index for the sample text

- For simplicity, split the index into two files.
  1. First file contains list of occurrences that are stored contiguously. This file is typically called a **posting file**.
  2. **Second file :** The vocabulary is stored in lexicographical order and for each word; a pointer to its list in the first file is also included. This allows the vocabulary to be kept in memory at search time.

3.8 Part A : Short Answered Questions [2 Marks Each]

Q.1 Mention types of classifier techniques.

Ans. : Types of classifier techniques are back-propagation, support vector machines, and k-nearest-neighbor classifiers.

Q.2 What is called Bayesian classification ?

Ans. : Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.

Q.3 Define decision tree.

Ans. : A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome(categorical or continues value). A decision tree or a classification tree is a tree in which each internal node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature.

Q.4 Define information gain. **AU : March-17**

Ans. : • Entropy measures the impurity of a collection. Information Gain is defined in terms of Entropy.

- Information gain tells us how important a given attribute of the feature vectors is.
- Information gain of attribute A is the reduction in entropy caused by partitioning the set of examples S.

$$\text{Gain } (S, A) = \text{Entropy } (S) - \sum_{v \in \text{ Values}(A)} \frac{|S_v|}{S} \text{Entropy } (S_v)$$

where Values (A) is the set of all possible values for attribute A and S_v is the subset of S for which attribute A has value v.

Q.5 Define pre pruning and post pruning. **AU : Dec.-16**

▣ **Ans. :** • In prepruning, a tree is "pruned" by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples.

- In the postpruning, it removes subtrees from a "fully grown" tree. A subtree at a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced.

Q.6 Why tree pruning useful in decision tree induction ?

▣ **Ans. :** When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of *overfitting* the data. Such methods typically use statistical measures to remove the least reliable branches.

Q.7 What is tree pruning ?

▣ **Ans. :** Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

Q.8 What are Bayesian classifiers ?

▣ **Ans. :** Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.

Q.9 What is meant by naive Bayes classifier ?

▣ **Ans. :** A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points.

Q.10 What are the characteristics of k-nearest neighbors algorithm ?

▣ **Ans. : Characteristics :**

- The unknown tuple is assigned the most common class among its k nearest neighbours.
- Nearest-neighbor classifiers use distance-based comparisons that intrinsically assign equal weight to each attribute.
- Nearest-neighbor classifiers can be extremely slow when classifying test tuples.
- Distance metric is calculated by using Euclidean distance and Manhattan distance.
- It does not use model building.
- It relies on local information.

Q.11 What is dimensionality reduction ?

▣ **Ans. :** In dimensionality reduction, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data. If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless.

Q.12 Define similarity.

▣ **Ans. :** The similarity between two objects is a numerical measure of the degree to which the two objects are alike. Similarities are usually non-negative and are often between 0 and 1. A small distance indicating a high degree of similarity and a large distance indicating a low degree of similarity.

Q.13 What is recall ?

▣ **Ans. :** Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection.

Q.14 What is precision ?

▣ **Ans. :** Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved.

Q.15 Define an inverted index. **AU : May-17**

▣ **Ans. :** An inverted index is an index into a set of documents of the words in the documents. The index is accessed by some search method. Each index entry gives the word and a list of documents, possibly with locations within the documents, where the word occurs. The inverted index data structure is a central component of a typical search engine indexing algorithm.

Q.16 What is zone index ? **AU : Dec.-17**

▣ **Ans. :** A zone is a region of the document that can contain an arbitrary amount of text, e.g., Title, Abstract and References. Build inverted indexes on zones as well to permit querying. Zones are similar to fields, except the contents of a zone can be arbitrary freetext.

3.9 Multiple Choice Questions

Q.1 Cluster is a group of _____ that belong to the same class.

[a] data [b] objects [c] information [d] data and information

Q.2 _____ clustering method works by grouping data objects into a tree of clusters.

[a] PAM [b] Density-based method

[c] Hierarchical [d] Grid-based method

Q.3 _____ can be cast as a problem of searching through a large predefined space of potential hypotheses.

[a] Machine learning [b] Concept learning

[c] Training set [d] Find-S

Q.4 A _____ is a flowchart-like tree structure, where each internal node denotes a test on an attribute.

[a] decision tree [b] binary tree

[c] cluster [d] none of these

Q.5 CART stands for _____.

[a] Concept and Regression Trees [b] Classification and Regression Trees

[c] Cluster and Regression Trees [d] None

Q.6 In theory, Bayesian classifiers have the _____ error rate in comparison to all other classifiers.

[a] equal [b] maximum [c] minimum [d] zero

Q.7 Error rate = 1 – (_____ rate)

[a] recall [b] precision [c] specificity [d] accuracy

Q.8 _____ classifiers are statistical classifiers.

[a] Bayesian [b] CART [c] Information gain [d] All of these

Q.9 _____ is a technique used to predict group membership for data instances.

[a] Clustering [b] Steaming

[c] Classification [d] None

Q.10 _____ matrix is also called a contingency table.

[a] Plain [b] Diffusion [c] Performance [d] Confusion

Q.11 A\An _____ is an index into a set of documents of the words in the documents.

[a] inverted index [b] decision tree

[c] confusion matrix [d] all of these

➤ Answer Keys for Multiple Choice Questions

Q.1	b	**Q.2**	c	**Q.3**	b	**Q.4**	a
Q.5	b	**Q.6**	c	**Q.7**	d	**Q.8**	a
Q.9	c	**Q.10**	d	**Q.11**	a		

❑❑❑

Notes

UNIT - IV

4 Web Retrieval and Web Crawling

Syllabus

The Web - Search Engine Architectures - Cluster based Architecture - Distributed Architectures - Search Engine Ranking - Link based Ranking - Simple Ranking Functions - Learning to Rank - Evaluations - Search Engine Ranking - Search Engine User Interaction - Browsing - Applications of a Web Crawler - Taxonomy - Architecture and Implementation - Scheduling Algorithms - Evaluation.

Contents

4.1 The Web

- Web is collection of millions of files stored on thousands of servers all over the world. These files represent documents, pictures, videos, sounds, programs, interactive environments.
- Following are hardware, software, protocols that make up the web.
  1. A web server is a computer connected to the Internet that runs a program that takes responsibility for storing, retrieving, distributing some of the web files. A web client (web browser) is a computer that requests files from the web.
  2. Well-defined set of languages and protocols independent of the hardware or operating system are required to run on the computers.
  3. The Hyper Text Markup Language (HTML) is the universal language of the Web.
  4. Java is a language for sending small applications over the Web. Java script is a language for extending HTML to embed little programs called scripts in web pages. The primary purpose of Java and scripts is to speed up the interactivity of web pages.
  5. Pictures, drawings, charts, diagrams are displayed on Web using image formats such as JPEG and GIF formats. The Virtual Reality Modeling Language is the web's way of describing three-dimensional objects.
- A web page is an HTML document that is stored on a web server. A Web site is a collection of web pages belonging to a particular organization. URL of these pages share a common prefix, which is the address of the size's home page.
- Search engines are a bottom-up approach for finding your way around the Web. Some search engines search only the titles of web pages. While another search every word. Keywords can be combined with Boolean operations, such as AND, OR , NOT, to produce rather complicated queries.

➤ Web Browsers

- A web browser is a program. Web browser is used to communicate with web servers on the Internet, which enables it to download and display the Web pages. Netscape Navigator and Microsoft Internet Explorer are the most popular browser software's available in the market.
- Browser interacts with Web as well as computer operating system and with other programs. Most browser windows have the same basic layout. Some of the essential elements are menu bar, toolbar, address or location window, viewing window status bar.
- The purpose of the web browser is to display web pages, which may either arrive over the Internet. Web browser can be used to view files of any common web format that are stored on the user system.

- WWW uses client-server interaction. The browser program acts as a client that uses the Internet to contact a remote server for a copy of the requested page. The server on the remote system returns a copy of the page along with the additional information.

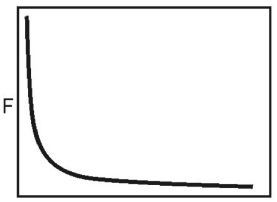
➥ 4.1.1 Characteristics

- In characterizing the structure and content of the Web, it is necessary to establish precise semantics for Web concepts.
- Two characteristics make retrieval of relevant information from the Web a really hard task :
  a. The large and distributed volume of data available
  b. The fast pace of change
- Internet and particular Web is dynamic in nature, so it is a difficult task to measure.
- In the mid-1980s, you might spend all afternoon visiting your friends beforedropping by the bank and grocery, and then go out to dinner and a show after.
- Today, your banking, shopping and chat with your friends are all readily handledfrom your tablet or phone, and if you're not in the mood for a fancy outing, Netflix and a quick Internet-ordered pizza take care of the evening's entertainment needs and all of it can be accomplished in less time than it takes to say "The Internet made me a hermit."
- It was something of a privilege to be online, but the net gained traction really quickly. While just 16 million people were using the web in 1995, this figure had leapt to 50 million three years later, one billion by 2009 and more than twice that last year.
- In 1993 there was an estimated 130 websites online, which jumped to ten thousand by 1996. Fast-forward to 2012, and there are 634 million websites competing for your attention.

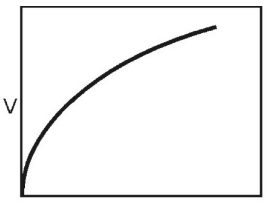
➥ 4.1.2 Modeling the Web

- Can we model the document characteristics of the whole web ? **Answer is "Yes"**.
- The Heap's and Zipf's laws are also valid in the web. Normally the vocabulary grows faster and the word distribution should be more biased. But there are no such experiments on large Web collections to measure these parameters.
- **Heaps' law :** An empirical rule which describes the vocabulary growth as a function of the text size.
- It establishes that a text of n words has a vocabulary of size $O(n^{\beta})$ for $0 < \beta < 1$.
- **Zipf's law :** An empirical rule that describes the frequency of the text words. It states that the i-th most frequent word appears as many times as the most frequent one divided by $i^{-\theta}$, for some $\theta > 1$.

- One more model is related to the distribution of document size. According to this model, the document sizes are self similar, they have a large variance. The probability of finding a document of size x bytes is given by :



(a) Distribution of sorted word frequencies



(b) Size of the vocabulary

Fig. 4.1.1

$$p(x) = \frac{1}{x_{\sigma}\sqrt{2\pi}} \exp -(ln\, x - \mu)^2 2\sigma^2$$

where the average (μ) and standard deviation (σ) are 9.357 and 1.318 respectively.

- The majority of the documents are small, but there is a non trivial number of large documents. This is intuitive for image or video files, but it is also true for HTML pages. A good fit is obtained with the Pareto distribution

$$p(x) = \frac{\alpha k^{\alpha}}{x^{1+\alpha}}$$

where x is measured in bytes and k and α are parameters of the distribution.

- So what languages dominate the Web ? It should come as no surprise that English still dominates the Web, with more than two-thirds of the Web's pages being in English. According to a study by, a Web site in the language Catalan, Japanese is the second most popular language of Web sites.

Web pages by language		
Language	**Web pages**	**Percent of total**
English	214,250,996	68.39
Japanese	18,335,739	5.85
German	18,069,744	5.77
Chinese	12,113,803	3.87
French	9,262,663	2.96
Spanish	7,573,064	2.42
Russian	5,900,956	1.88
Italian	4,883,497	1.56
Portuguese	4,291,237	1.37
Korean	4,046,530	1.29
Dutch	3,161,844	1.01
Sweden	2,929,241	0.93

- As you can see, the growth is impressive and unimpeded. Also :
  a. The total number of web sites seems to follow Moore's Law and double every 18-24 months.
  b. At the current rate, we will hit 1 billion sites in 2013 and 2 billion sites in 2015.
  c. Over the years, the number of web sites seems to be roughly equal to the number of people on the internet.
  d. If WordPress continues on its current trajectory, there will be 300-500 million WordPress sites by 2015.

4.1.3 Link Analysis

- The goal of information retrieval is to find all documents relevant for a user query in a collection of documents. With the advent of the web new sources of information became available, one of them being the hyperlinks between documents and records of user behaviour.
- Collections of documents connected by hyperlinks. Hyperlinks provide a valuable source of information for web information retrieval. This area of information retrieval is commonly called link analysis.
- A hyperlink is a reference of a web page B that is contained in web page A. When the hyperlink is clicked on in a web browser, the browser displays page B. This functionality alone is not helpful for web information retrieval.
- Link analysis has been used successfully for deciding which web pages to add to the collection of documents and how to order the documents matching a user query.
- Link analysis for web search has intellectual antecedents in the field of citation analysis, aspects of which overlap with an area known as bibliometrics.
- Link analysis are of three types : Microscopic level, mesoscopic level and macroscopic level.
  1. Microscopic level : It is related to the statistical properties of links and individual nodes.
  2. Mesoscopic level : It is related to the properties of areas or regions of the web.
  3. Macroscopic level : It is related to the structure of the web at large.

4.2 Search Engine Architectures

- Search engines are becoming the primary entry point for discovering web pages. Ranking of web pages influences which pages users will view. Exclusion of a site from search engines will cut off the site from its intended audience.
- A search engine is a program designed to help find information stored on a computer system such as the World Wide Web or a personal computer. The search engine allows one to ask for content meeting specific criteria and retrieves a list of references that match those criteria.

4.2.1 Cluster based Architecture

- Centralized crawler indexer architecture is used by most of the search engine. Using crawlers, information is gathered into a single site, where it is indexed; the site then processes all user queries.
- This system has its own data collecting mechanism, and all the data are stored and indexed in a conventional database system. Although many web search engines download web pages and provide service by thousands of servers, they all belong to this kind according to their basic architecture.
- Centralized architecture consists of following components :

  1. Crawlers 2. Index 3. Query engine 4. User interface.

1. Crawlers

- Crawlers are programs i.e. software agents that traverse the web sending new or updated pages to a main server where they are indexed. It sends requests to Web sites and downloads Web pages. The local system sends requests for Web pages to remote Web servers.
- In the downloaded documents it discovers new pages and new servers. The effect of such repeated requests is of a "robot" that moves from site to site, gathering information at every site it visits. In reality, this "crawler" never leaves the local system.
- Crawlers are also called robots, spiders, wanderers and walkers.
- Fig. 4.2.1 shows the software architecture of a search engine based on Altavista architecture.

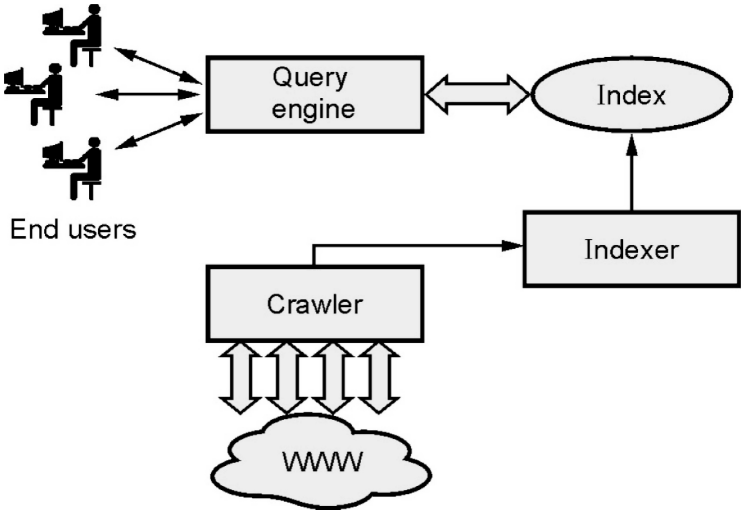


Fig. 4.2.1 : Typical crawler indexer architecture

- Figure consists of two parts : One deal with users, consisting of the user interface and the query engine and another that consists of the crawler and indexer modules.

➤ 2. Indexer

- The index is used in a centralized fashion to answer queries submitted from different places in the web. Index the downloaded pages. Each downloaded page is processed locally.
- The indexing information is saved and the page is discarded.
- A module that takes a collection of documents or data and builds a searchable index from them. Common practices are inverted files, vector spaces, suffix structures and hybrids of these.
- *Exception :* Some search engines keep a local cache copy of many popular pages indexed in their database, to allow for faster access and in case the destination server is temporarily in-accessible.

➤ 3. User interface :

Solicit queries and deliver answers. All requests are submitted to a single site.

➤ 4. Query engine :

- It process queries against the index. All processing is done locally. It requires a massive array of computers and storage.
- Following table gives idea about search engine with URL and web page indexed upto May 1998.

Search engine	URL	Web page indexed
AltaVista	www.altavista.com	140
AOL Netfind	www.aol.com/netfind/	-
Excite	www.excite.com	55
Google	google.stanford.edu	25
GoTo	goto.com	-
HotBot	www.hotbot.com	110
Infoseek	www.infoseek.com	30
Lycos	www.lycos.com	30
Magellan	www.mckinley.com	55
Microsoft	search.msn.com	-
northernLight	www.nlsearch.com	67
WebCrawler	www.webcrawler.com	2

➤ **Problems using this architecture :**

1. Difficult to gathering the data because of dynamic nature of the Web.
2. The high load on Web servers.
3. Large volume of data : Could be difficult for crawler to cope with Web growth
4. Communication link problem.

- Most of the search engines are based in the United States and focuses on document in English. Some search engines specialized in different countries and different languages. Some of the search engines retrieve only specific web pages such as personal or institutional home pages or specific objects such as electronic mail address, images etc.

➥ 4.2.2 Distributed Architecture

- When the data source is large enough that even the metadata can't be efficiently managed in a database system, we can choose distributed system. Distributed information retrieval system does not have its own actual record database. It just indexes the interface of sub database system.
- When receiving a query from a user, main system will instantly obtain the records from sub databases through their search interfaces. The limitation of this system is that the number of sub databases can't be many, otherwise the search speed can't be ensured. A famous system is InfoBus system in Stanford digital library project.
- Harvest is an example of distributed architecture.

➤ **Harvest**

- Harvest is a distributed crawler-indexer architecture which addresses the main problems in crawling and indexing the Web :
  1. Web servers get requests from different crawlers of search engines which increase the server's load;
  2. Most of the entire objects retrieved by the crawlers are useless and discarded;
  3. No coordination exists among the crawlers.
- Harvest is designed to be a distributed search system where machines work together to handle the load which a single machine could not handle.
- Harvest also can be used to save bandwidth by deploying gatherers near the data source and exchanging the summarized data which usually is much smaller than the original data.
- But it seems most of further Harvest applications are in the field of caching Web objects instead of providing advanced internet search services. State of the art indexing techniques can reduce the size of an inverted file to about 30 % of the size of the text.

- Fig. 4.2.2 shows Harvest architecture.
- Harvest does not suffer from some of common problems of the crawler-indexer architectures, such as

1. Increased servers load caused by reception of simultaneous requests from different crawlers

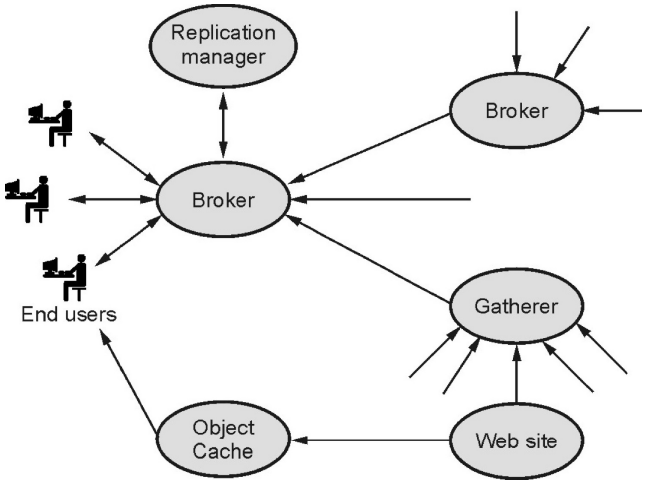


Fig. 4.2.2 : Harvest architecture

2. Increased web traffic, due to crawlers retrieving entire objects, while most content is not retained eventually
3. Lack of coordination between engines, as information is gathered independently by each crawler

➤ Goals of harvest :

- One of the goals of the Harvest is to build topic specific brokers, focusing the index contents and avoiding many of the vocabulary and scaling problems of generic indices. It includes a distinguished broker that allows other brokers to register information about gathers and brokers.
- This is useful to search for an appropriate broker or gatherer when building a new system.

➤ Components :

1. Gatherers : It extracts information from the documents stored on one or more Web servers. It can handle documents in many formats like HTML, PDF, Postscript, etc.
2. Broker : Provides the indexing mechanism and query interface.
3. Replicator : To replicate servers. It enhances user base scalability.
4. Object cache : Reduces network and server load. It also reduces response latency when accessing Web pages.

➤ **Features of harvest :**

- Harvest is a modular, distributed search system framework with working set components to make it a complete search system. The default setup is to be a web search engine, but it is also much more and provides following features :
  1. Harvest is designed to work as distributed system. It can distribute the load among different machines. It is possible to use a number of machines to gather data. The full-text indexer doesn't have to run on the same machine as broker or web server.
  2. Harvest is designed to be modular. Every single step during collecting data and answering search requests are implemented as single programs. This makes it easy to modify or replace parts of Harvest to customize its behaviour.
  3. Harvest allows complete control over the content of data in the search database. It is possible to customize the summarizer to create desired summaries which will be used for searching. The filtering mechanism of Harvest allows making modifications to the summary created by summarizers. Manually created summaries can be inserted to the search database.
  4. The Search interface is written in Perl to make customization easy, if desired.
- For 100 million pages, this implies about 150 GB of disk space. Assuming that 500 bytes are required to store the URL and the description of each Web page, we need 50 GB to store the description for 100 million pages.
- The use of Meta search engines is justified by coverage studies that show that a small percentage of Web pages are in all search engines. Moreover, less than 1 % of the Web pages indexed by AltaVista, HotBot, Excite and Inforseek are in all of those search engines.

➤ **Advantages of distributed architecture :**

1. *Server load reduced :* A gatherer running on a server reduces the external traffic (i.e., crawler requests) on that server.
2. *Network traffic reduced :* Crawlers retrieve entire documents, whereas Harvest moves only summaries.
3. *Redundant work avoided :* A gatherer sending information to multiple brokers reduces work repetition.

4.3 Search Engine Ranking

Ranking refers to the process search engines use to determine where a particular piece of content should appear on a SERP. Search visibility refers to how prominently a piece of content is displayed in search engine results.

➥ 4.3.1 Link based Ranking

- The primary challenge of a search engine is to return results that match a user's needs. A word will potentially map to millions of documents. How to order them ?
- Most of the search engines use variations of the Boolean or vector model to do ranking. Every search engine strives to return relevant web pages that will satisfy your requests. Each search engine uses a proprietary 'ranking algorithm' that attempts to instantly build a list of highly appropriate responses to your query.
- Since each search engine applies its own formula to a unique database of information, results and relevancy rankings will always vary from search engine to search engine.
- Yuwono and Lee propose three ranking algorithms in addition to the classical scheme. They are called

  1. Boolean spread 2. Vector spread 3. Most cited
- Boolean spread and vector spread are the normal ranking algorithms of the Boolean and vector model extended to include pages pointed to by a page in the answer. The last most cited is based only on the terms included in pages having a link to the pages in the answer.
- Hyperlink information is also used by some of the new ranking algorithms. The number of hyperlinks that point to a page provides measures of its popularity and quality. Also many links in common between pages or pages referenced by the same page often indicates a relationship between those pages. Here we will discuss three examples of ranking techniques.

➤ a. WebQuery :

- It allows visual browsing of Web pages. WebQuery takes a set of web pages and ranks them based on how connected each web page is.

➤ b. HITS algorithm :

- Algorithm developed by Kleinberg in 1998.
- Second method is used in Hypertext Induced Topic Search (HITS). This ranking scheme depends on the query and considers the set of pages S that point to or are pointed by pages in the answer.
- Pages that have many links pointing to them in S are called authorities. Pages that have many outgoing links are called hubs. Gives each page a hub score and an authority score. A good authority is pointed to by many good hubs. A good hub points to many good authorities. Users want good authorities.

- Computes hubs and authorities for a particular topic specified by a normal query. First determines a set of relevant pages for the query called the base set S. Analyze the link structure of the web sub-graph defined by S to find authority and hub pages in this set.
- Hubs : It contains many outward links and lists of resources.
- Authorities : It contains many inward links and provides resources, content.

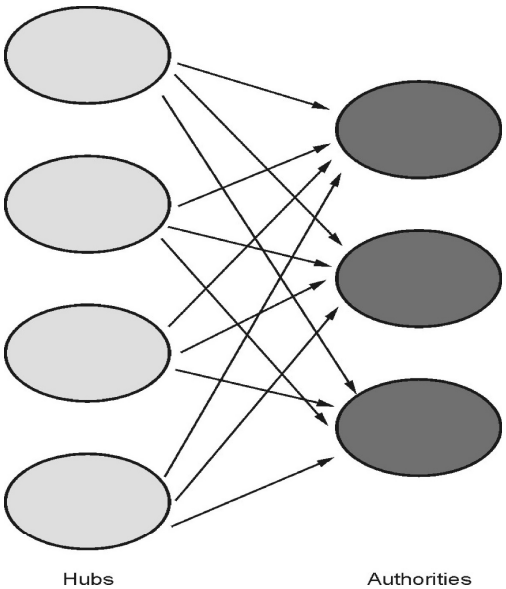


Fig. 4.3.1

- Let H(p) and A(p) be the hub and authority value of pages p. These values are defined such that the following equations are satisfied for all pages p. Authorities are pointed to by lots of good hubs :

$$a_p = \sum_{q : q \to p} h_q$$

- Hubs point to lots of good authorities :

$$h_p = \sum_{q : p \to q} a_q$$

➤ C. PageRank

- Numeric value to measure how important a page is. PageRank (PR) is the actual ranking of a page, as determined by Google.
- A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50 % chance" of something happening. Hence, a PageRank of 0.5 means there is a 50 % chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank.

- Purpose : To increase the quality of search results and has left all of its competitors for dead.
- The Google Page Rank is based on how many links you have to and from your pages, and their respective page rank.
- We update our index every four weeks. Each time we update our database of web pages, our index invariably shifts : We find new sites, we lose some sites, and sites ranking may change. Your rank naturally will be affected by changes in the ranking of other sites.
- The Google PageRank (PR) is calculated for every webpage that exists in Google's database. It's real value varies from 0,15 to infinite, but for representation purposes it is converted to a value between 0 and 10 (from low to high). The calculation of the PR for a page is based on the quantity and quality of web pages that contain links to that page.
- Let C(a) be the number of outgoing links of page "a" and suppose that page "a" is pointed to by pages p1 to pn. Then, the PageRank PR(a) of "a" is defined as

$$PR(a) = 1 + (1 - q) \sum_{i=1}^{n} PR(p_i) / C(p_i)$$

  Where q must be set by the system.
- It is obvious that the PageRank algorithm does not rank the whole website, but it's determined for each page individually. Furthermore, the PageRank of page A is recursively defined by the PageRank of those pages which link to page A.
- PageRank can be computed using an iterative algorithm. This means that each page is assigned an initial starting value and the PageRanks of all pages are then calculated in several computation circles based on the equations determined by the PageRank algorithm. The iterative calculation shall again be illustrated by the three-page example, whereby each page is assigned a starting PageRank value of 1.

➥ 4.3.2 Simple Ranking Functions

- Simplest ranking scheme use a global ranking function such as PageRank. In this case, quality of a Web page in the result set is independent of the query. The query only selects pages to be ranked.
- To elaborate ranking scheme, use a linear combination of different ranking signals.
- To illustrate, consider the pages p that satisfy query Q Rank score R (p, Q) of page p with regard to query Q can be computed as :

$$R(p, Q) = \alpha\, BM25(p, Q) + (1 - \alpha)\, PR(p)$$

Where $\alpha = 1$: text-based ranking, early search engines

$\alpha = 0$: link-based ranking, independent of the query

- Current engines combine a text-based ranking with a link-based ranking.

4.3.3 Learning to Rank

- Various methods are used for computing a Web ranking. Some of them are given below :
  1. Apply machine learning techniques to learn the ranking of the results
  2. Use a learning algorithm fed with training data that contains ranking information
  3. Loss function to minimize: number of mistakes done by learned algorithm
- Given query Q, three types of training data can be used :
  1. Pointwise : a set of relevant pages for Q
  2. Pairwise : a set of pairs of relevant pages indicating the ranking relation between the two pages
  3. Listwise : a set of ordered relevant pages: $p_1 > p_2 \cdots > p_m$

4.3.4 Evaluations

- To evaluate quality, Web search engines typically use
  a. human judgements of which results are relevant for a given query
  b. some approximation of a ground truth inferred from user's clicks
  c. combination of both
- To evaluate search results, use precision-recall metrics. Precision of Web results should be measured only at the top positions in the ranking, say P@5, P@10, and P@20. Each query-result pair should be subjected to 3-5 independent relevant assessments.
- An advantage of using click-through data to evaluate the quality of answers derives from its scalability. Note that users' clicks are not used as a binary signal but in significantly more complex ways such as :
  a. Considering whether the user remained a long time on the page it clicked (a good signal),
  b. Jumped from one result to the other (a signal that nothing satisfying was found), or
  c. The user clicked and came back right away (possibly implies Web spam)
- These measures and their usage are complex and kept secret by leading search engines. An important problem when using clicks is to take into account that the click rate is biased by the ranking of the answer and the user interface.

4.4 Search Engine User Interaction

Most search engine users have very little technical background. So design of the interface has been heavily influenced by extreme simplicity rules.

➤ Search Rectangle Paradigm

- Users are now accustomed with specifying their queries in a search rectangle commonly referred to as the search box. Fig. 4.4.1 shows various search engines.

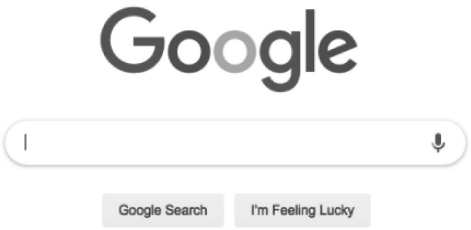


Fig. 4.4.1 : Google home page looked in Feb 2019

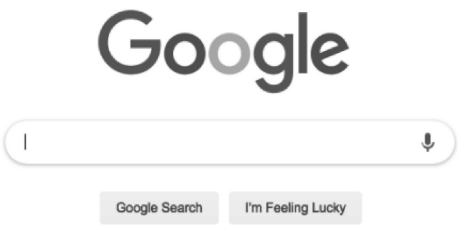


Fig. 4.4.2 : Google home page looked in March 2019

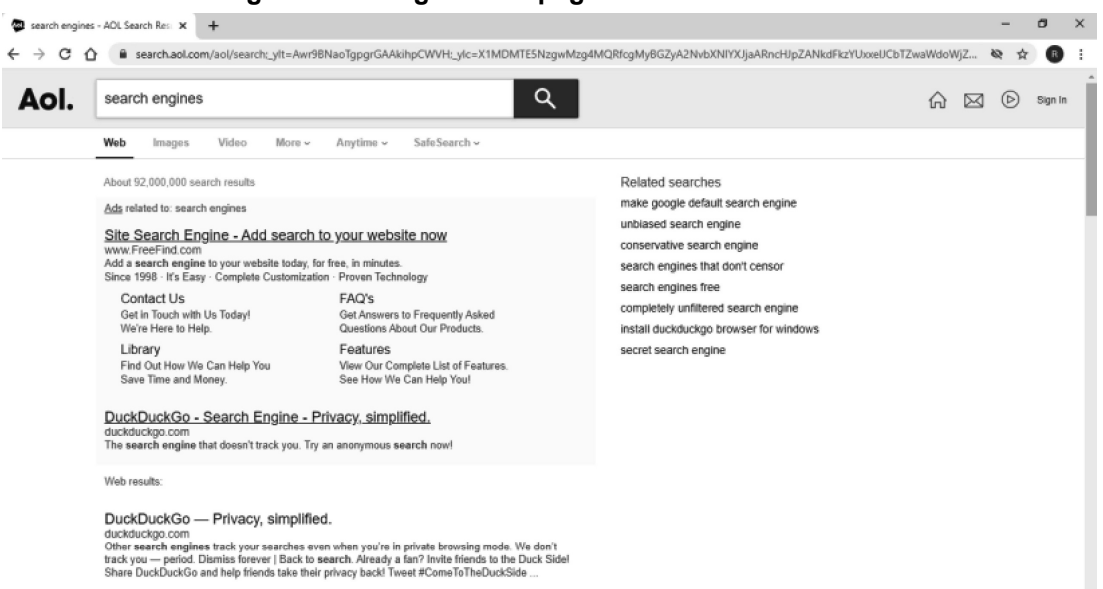


Fig. 4.4.3 : Aol search engine homepage

- While search rectangle is the favored layout style, there are alternatives :
  a. Many sites include an Advanced Search page (rarely used)
  b. Search toolbars provided by most search engines as a browser plug-in can be seen as a version of the search rectangle.
  c. Ultimate rectangle, introduced by Google's Chrome omnibox, merges the functionality of the address bar with that of the search box.

➤ Query Language

- Queries in relational or object-oriented database systems are based on an exact match mechanism, by which the system is able to return exactly those tuples satisfying some well specified criteria given in the query expression.
- When the query is submitted, the features of the query object are matched with respect to the features of the objects stored in the database and only the objects that are more similar to the query one is returned to the user.
- In designing a multimedia query language, following points are considered :
  1. How the user enters his/her request to the system, i.e. which interfaces are provided to the user for query information ?
  2. Which conditions on multimedia objects can be specified in the user request.
  3. How uncertainty, proximity and weights impact the design of the query language.

➤ Search Query Operators

- Search operator is classified as
  1. Punctuation based search operator.
  2. Boolean search operator.
  3. Advanced search operators.
- Boolean search is a search that uses the logical operators (AND, OR, NOT, -) in addition to the keywords.
  a. AND : The AND operator tells the search engine to return only documents with all the keywords you entered.
  b. OR : The OR operator tells the search engine to return documents if they contain one or more keywords.
  c. NOT : The NOT operator tells the search engine to exclude documents from a search if they contain the keywords.
  d. - Operator : The "-" operator is the same as the NOT operator and tells the search engine to exclude documents from a search if they contain the keywords.

4.5 Browsing

- Browsing and searching are the two main paradigms for finding information on line. The search paradigm has a long history; search facilities of different kinds are available in all computing environments.
- The browsing paradigm is newer and less ubiquitous, but it is gaining enormous popularity through the World-Wide Web.
- Both paradigms have their limitations. Search is sometimes hard for users who do not know how to form a search query so that it is limited to relevant information.
- Browsing can make the content come alive, and it is therefore more satisfying to users who get positive reinforcement as they proceed. However, browsing is time-consuming and users tend to get disoriented and lose their train of thoughts and their original goals.
- **Flat browsing :** User explores a document space that flows a flat organization. Documents might be represented as dots in a two-dimensional plane or as elements in a single dimension list, which might be ranked by alphabetical or by any other order.

4.5.1 Web Directories

- Web directory is a classification of Web pages by subject. One of best and oldest Web directory is Yahoo, which is most used searching tool. eBLAST, LookSmart Magellan and NewHoo are the example of Web directories. Some of them are hybrid in nature.
- Web Directories also called catalogs, yellow pages or subject directories.

➤ Principles :

- Classification is by a hierarchical taxonomy.
- Directory may be specific to a subject, a region, a language.
- Pages are submitted and reviewed before they are included.
- Automatic classification is not successful enough.
- Some subcategories are also available in the main page of Web directories.
- **Advantage :** If found, the answer will be useful in most cases;
- **Disadvantage :** Classification is not specialized enough; Not all Web pages are classified;

4.6 Applications of a Web Crawler

- A web crawler is a program which browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

- Web crawlers is also called ant, bot, worm or Web spider. The process of scanning the WWW is called Web crawling or spidering.
- Web crawling is the process by which we gather pages from the Web, in order to index them and support a search engine. The objective of crawling is to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them.
- It starts with a list of URLs to visit. As it visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, recursively browsing the Web according to a set of policies.
- Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine, which will index the downloaded pages to provide fast searches.
- Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses.
- When a search engine's web crawler visits a web page, it "reads" the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich Meta tags.
- Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information. The website is then included in the search engine's database and its page ranking process.

➤ **Crawling process:**

- The crawler begins with a seed set of URLs to fetch. The crawler fetches and parses the corresponding Web-pages and extracts both text and links. The text is fed to a text indexer; the links (URL) are added to a URL frontier.
- Initially, the URL frontier contains the seed set; as pages are fetched, the corresponding URLs are deleted from the URL frontier. The entire process may be viewed as traversing the web graph. In continuous crawling, the URL of a fetched page is added back to the frontier for fetching again in the future.

➥ 4.6.1 Web Crawler Architecture

- Fig. 4.6.1 shows web crawler architecture.
- The **URL frontier**, containing URLs yet to be fetched in the current crawl. In the case of continuous crawling, a URL may have been fetched previously but is back in the frontier for re-fetching.

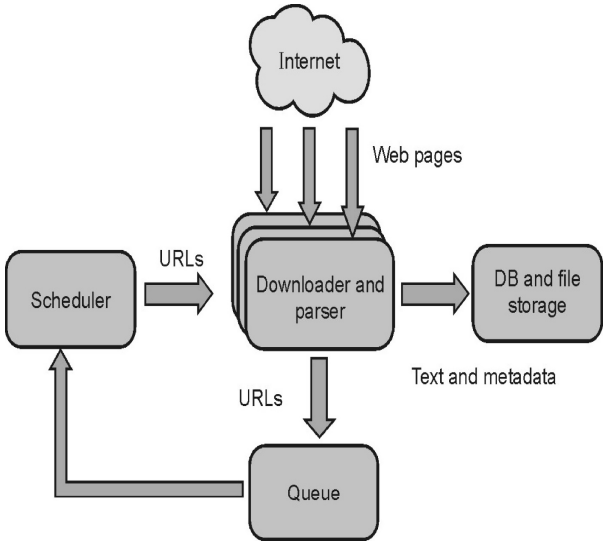


Fig. 4.6.1 : Web crawler architecture

- A **DNS resolution** module that determines the web server from which to fetch the page specified by a URL.
- A fetch module that uses the http protocol to retrieve the web page at a URL.
- A parsing module that extracts the text and set of links from a fetched web page.
- A duplicate elimination module that determines whether an extracted link is already in the URL frontier or has recently been fetched.
- **Robots.txt file :** A web server administrator can use a file / robots.txt to notify a crawler what is allowed and what is not.

➤ **Politeness policies :**

- Within website www.cs.sfu.ca, many pages contain many links to other pages in the same site. Downloading many pages simultaneously from one website may overwhelm the server.
- A reasonable web crawler should use only a tiny portion of the bandwidth of a website server, not fetching more than one page at a time.
- **Implementation :** Logically split the request queue into a single queue per web server; a server queue is open only if it has not been accessed within the specified politeness window.
- Suppose a web crawler can fetch 100 pages per second and the politeness policy dictates that it cannot fetch more than 1 page every 30 seconds from a server then we need URLs from at least 3,000 different servers to make the crawler reach its peak throughput.

- **Detecting updates :** If a web page is updated, the page should be crawled again. In HTTP, request HEAD returns only header information about the page but not the page itself. A crawler can compare the date it received from the last GET request with the Last-Modified value from a HEAD request.

4.6.2 Taxonomy of Crawler

➤ Features of a crawler :

1. Robustness : Ability to handle spider-traps (cycles, dynamic web pages). Robustness should not fall into spider traps which generate an infinite number of pages.
2. Politeness : Policies about the frequency of robot visits
3. Distribution : Crawling should be distributed within several machines
4. Scalability : Crawling should be extensible by adding machines, extending bandwidth, etc.
5. Efficiency : Clever use of the processor, memory, bandwidth.
6. Quality : Should detect the most useful pages, to be indexed first
7. Freshness : Should continuously crawl the web
8. Extensibility : Should support new data formats (e.g. XML-based formats), new protocols (e.g. ftp), etc.

➤ Web page Types :

- **Static web pages** don't change content or layout with every request to the web server. They change only when a web author manually updates them with a text editor or web editing tool like Adobe Dreamweaver.
- Dynamic web pages can adapt their content or appearance depending on the user's interactions, changes in data supplied by an application, or as an evolution over time, as on a news web site.

4.7 Scheduling Algorithms

- Scheduling algorithm uses three policies with different goals :
  1. Selection policy
  2. Revisit policy
  3. Politeness policy
  - **Selection policy :** As only a percent of the Web can be downloaded, a web crawler must use a selection policy to determine which resources are relevant to download. This is more useful than downloading a random portion of the Web.

- An example of a selection policy is the PageRank policy where the importance of a page is determined by the links to and from that page.
- Common selection policies are restricting followed links, path-ascending crawling, focused crawling and crawling the deep web.
- **Revisit policy :** Web crawlers use revisiting policies to determine the cost associated with an outdated resource. The goal is to minimize this cost. This is important because resources in the Web are continually created, updated or deleted; all within the time it takes a web crawler to finish its crawl through the Web.
- Two re-visit policies are uniform policy and proportional policy.
- The **politeness policy** is used so that the performance of a site is not heavily affected whist the web crawler downloads a portion of the site. The server may be overloaded as it has to handle the requests of the viewers of the site as well as the web crawler.

4.8 Part A : Short Answered Questions [2 Marks Each]

Q.1 Define web server.

▣ **Ans. :** A web server is a computer connected to the Internet that runs a program that takes responsibility for storing, retrieving and distributing some of the web files

Q.2 What is web browsers?

▣ **Ans. :** A web browser is a program. Web browser is used to communicate with web servers on the Internet, which enables it to download and display the web pages. Netscape Navigator and Microsoft Internet Explorer are the most popular browser software's available in market.

Q.3 Explain paid submission of search services.

▣ **Ans. :** In paid submission, user submit website for review by a search service for a preset fee with the expectation that the site will be accepted and included in that company's search engine, provided it meets the stated guidelines for submission. Yahoo! is the major search engine that accepts this type of submission. While paid submissions guarantee a timely review of the submitted site and notice of acceptance or rejection, you're not guaranteed inclusion or a particular placement order in the listings.

Q.4 Explain paid inclusion programs of search services.

▣ **Ans. :** Paid inclusion programs allow you to submit your website for guaranteed inclusion in a search engine's database of listings for a set period of time. While paid inclusion guarantees indexing of submitted pages or sites in a search database, you're not guaranteed that the pages will rank well for particular queries.

Q.5 Define search engine optimization.

▣ **Ans. :** Search Engine Optimization (SEO) is the act of modifying a website to increase its ranking in organic (vs paid), crawler-based listings of search engines. There are several ways to increase the visibility of your website through the major search engines on the Internet today. The two most common forms of Internet marketing Paid(Sponsored) Placement and Natural Placement

Q.6 What is the purpose of web crawler ? **AU : Dec.-16**

▣ **Ans. :** A web crawler is a program which browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

Q.7 Define focused crawler.

▣ **Ans. :** A focused crawler or topical crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics.

Q.8 What are the near-duplicate detection?

▣ **Ans. :** Near-duplicate detection is the task of identifying documents with almost identical content. Near-duplicate web documents are abundant. Two such documents differ from each other in a very small portion that displays advertisements, for example. Such differences are irrelevant ant for web search

Q.9 Define web crawling. **AU : May-17**

▣ **Ans. :** A web crawler is a program which browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

Q.10 What are politeness policies used in web crawling ? **AU : Dec.-17**

▣ **Ans. :** Politeness policies are as follows :

a. To adjust the crawl frequency

b. To revisit pages with the same frequency, ignoring the change rate of individual pages.

Q.11 What is snippet generation ? **AU : Dec.-17**

▣ **Ans. :** Snippet generation is a special type of extractive document summarization, in which sentences are selected for inclusion in the summary on the basis of the degree to which they match the search query. This process was given the name of query based summarization.

Q.12 What is PageRank ?

▣ **Ans. :** A method for rating the importance of web pages objectively and mechanically using the link structure of the web

Q.13 Define dangling link.

▣ **Ans. :** This occurs when a page contains a link such that the hypertext points to a page with no outgoing links. Such a link is known as dangling link.

Q.14 Define snippets.

▣ **Ans. :** Snippets are short fragments of text extracted from the document content or its metadata. They may be static or query based. In static snippet, it always shows the first50 words of the document, or the content of its description metadata, or a description taken from a directory site such as dmoz.org. A query-biased snippet is one selectively extracted on the basis of its relation to the searcher's query.

Q.15 Define hubs.

▣ **Ans. :** Hubs are index pages that provide lots of useful links to relevant content pages(topic authorities). Hub pages for IR are included in the home page.

Q.16 Define authorities. **AU : Dec.-16**

▣ **Ans. :** Authorities are pages that are recognized as providing significant, trustworthy, and useful information on a topic. In-degree (number of pointers to a page) is one simple measure of authority. However in-degree treats all links as equal.

4.9 Multiple Choice Questions

Q.1 _____ is a program which browses the World Wide Web in a methodical, automated manner.

[a] Web crawler [b] HTML [c] URL [d] Web

Q.2 _____ is the act of modifying a website to increase its ranking in organic, crawler-based listings of search engines.

[a] Google search engine [b] Search engine optimization

[c] Web crawler optimization [d] All of these

Q.3 _____ occurs when a page contains a link such that the hypertext points to a page with no outgoing links.

[a] Deep link [b] Dark link

[c] Dangling link [d] All of these

Q.4 _____ are index pages that provide lots of useful links to relevant content pages.

[a] Authorities [b] Hub [c] Web page [d] None

Q.5 SEO copywriting aims to produce _____ on a website.

[a] page [b] rank [c] value [d] content

Q.6 Which of the following is not the factors that affect the search engine ranking?

[a] Visible on-page factors [b] Invisible on-page factors

[c] Time-based factors [d] Internal factors

Q.7 _____ web pages do not change content or layout with every request to the web server.

[a] Dynamic [b] static [c] dynamic and static [d] None

Q.8 Boolean search is a search that uses the _____ operators.

[a] binary [b] relational [c] logical [d] bitwise

Q.9 _____ Link analysis is related to the statistical properties of links and individual nodes.

[a] Mesoscopic level [b] Macroscopic level

[c] Microscopic level [d] All of these

Q.10 The graph model in _____ link analysis is induced from two kinds of relationships, that is, block-to-page and page-to-block.

[a] page level [b] text level [c] block level [d] document level

Q.11 _____ ranking methods use the query to rank all documents in the order of relevance.

[a] Text [b] Word [c] Information [d] Document

➤ **Answer Keys for Multiple Choice Questions**

Q.1	a	**Q.2**	b	**Q.3**	c	**Q.4**	b
Q.5	d	**Q.6**	d	**Q.7**	b	**Q.8**	c
Q.9	c	**Q.10**	c	**Q.11**	d		

❑❑❑

UNIT - V

5 Recommender System

Syllabus

Recommender Systems Functions - Data and Knowledge Sources - Recommendation Techniques - Basics of Content-based Recommender Systems - High Level Architecture - Advantages and Drawbacks of Content-based Filtering - Collaborative Filtering - Matrix factorization models - Neighborhood models.

Contents

5.1 Recommender Systems Functions

- Recommender systems are a way of suggesting like or similar items and ideas to a users specific way of thinking. Recommender systems are widely used on the Web for recommending products and services to users
- Recommender systems try to automate aspects of a completely different information discovery model where people try to find other people with similar tastes and then ask them to suggest new things.
- The goal of a recommender system is to generate meaningful recommendations to a collection of users for items or products that might interest them. Suggestions for books on Amazon, or movies on Netflix, are real-world examples of the operation of industry strength recommender systems.
- Fig. 5.1.1 shows recommendation systems concept.

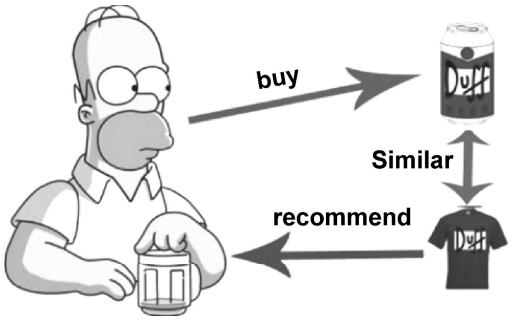


Fig. 5.1.1 : Recommendation systems

- Recommendation systems are a key part of almost every modern consumer website. The systems help drive customer interaction and sales by helping customers discover products and services they might not ever find themselves.
- Recommender systems predict the preference of the user for these items, which could be in form of a rating or response. When more data becomes available for a customer profile, the recommendations become more accurate.
- There are a variety of applications for recommendations including movies (e.g. Netflix), consumer products (e.g., Amazon or similar on-line retailers), music (e.g. Spotify), or news, social media, online dating and advertising.
- Fig. 5.1.2 shows how Amazon uses recommendation concept.

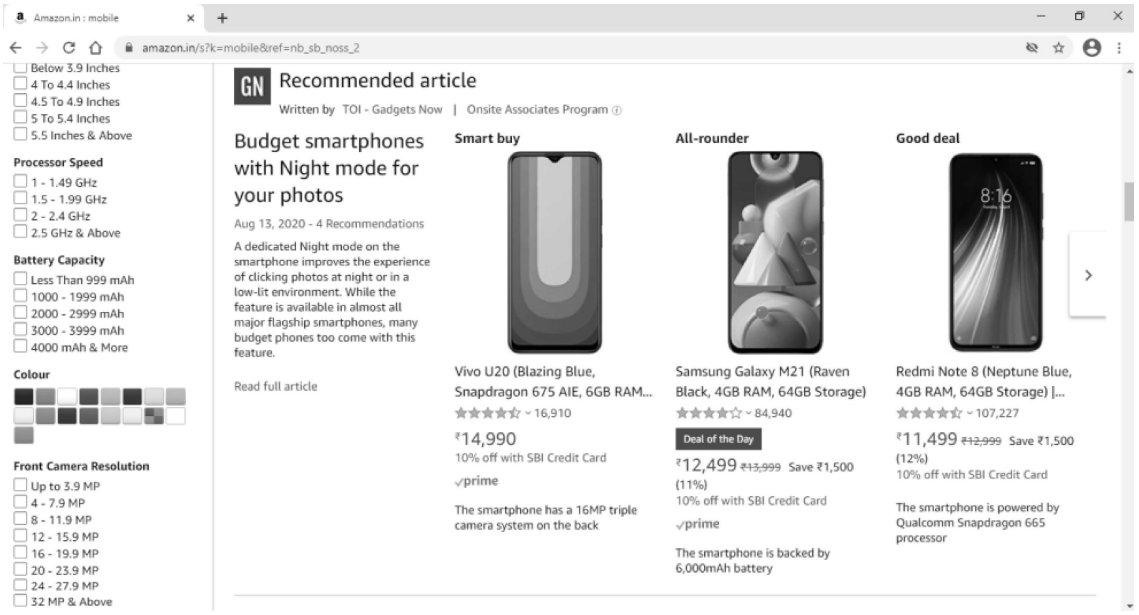


Fig. 5.1.2 How Amazon uses recommendation concept

- When you searching mobile on Amazon, it display various mobile and it also display recommended article for mobiles.
- These systems serve two important functions.
  1. They help users deal with the information overload by giving them recommendations of products, etc.
  2. They help businesses make more profits, i.e., selling more products
- Various reasons why service providers increase the use of recommendation systems :
  1. It increases the number of product/items sold.
  2. Sell more diverse products/items
  3. User satisfaction is increases
  4. Increase user fidelity
  5. Better understand what the user wants
- The most common scenario is the following :
  a) A set of users has initially rated some subset of movies (e.g., on the scale of 1 to 5) that they have already seen.
  b) These ratings serve as the input. The recommendation system uses these known ratings to predict the ratings that each user would give to those not rated movies by him/her.
  c) Recommendations of movies are then made to each user based on the predicted ratings.

➤ Recommendation process :

- Every recommendation system follows a specific process in order to produce product recommendations.

- The recommendation approaches can be classified based on the information sources they use. Three possible sources of information can be identified as input for there commendation process. The available sources are the user data (demographics), the item data(keywords, genres) and the user-item ratings.

5.1.1 Challenges

- Following are the challenges for building recommender systems :
  1. Huge amounts of data, tens of millions of customers and millions of distinct catalog items.
  2. Results are required to be returned in real time.
  3. New customers have limited information.
  4. Old customers can have a glut of information.
  5. Customer data is volatile.

5.2 Data and Knowledge Sources

- Recommender systems are information processing systems which actively collect/gather various types of data for designing recommendations system. Data is primarily about the items to suggest and the users who will receive these recommendations.
- Items are the objects that are recommended. Items may be characterized by their complexity and their value or utility.
- Transactions are log-like data that store important information generated during the human-computer interaction and which are useful for the recommendation generation algorithm that the system is using.

5.3 Recommendation Techniques

In general, there are three types of recommender system :

1. Collaborative recommender system is a system that produces its result based on past ratings of users with similar preferences
2. Content based recommender system is a system that produces its result based on the similarity of the content of the documents or items.
3. Knowledge based recommender system is a system that produces its result based on additional and means end knowledge.
4. Demographic based recommender system: This type of recommendation system categorizes users based on a set of demographic classes. This algorithm requires market research data to fully implement. The main benefit is that it doesn't need history of user ratings
5. Hybrid recommender systems combine various inputs and different recommendation strategies to take advantage of the synergy among them.

5.4 Basics of Content-based Recommender Systems

- Content-based recommenders refer to such approaches, that provide recommendations by comparing representations of content describing an item to representations of content that interests the user. These approaches are sometimes also referred to as content-based filtering.
- Content-based recommendation systems try to recommend items similar to those a given user has liked in the past.
- In a movie recommendation application, a movie may be represented by such features as specific actors, director, genre, subject matter, etc.
- The user's interest or preference is also represented by the same set of features, called the user profile.
- Recommendations are made by comparing the user profile with candidate items expressed in the same set of features. The top-k best matched or most similar items are recommended to the user.
- The simplest approach to content-based recommendation is to compute the similarity of the user profile with each item.

5.4.1 High Level Architecture Content-based Recommender Systems

- Fig. 5.4.1 shows High Level Architecture Content-based Recommender Systems.

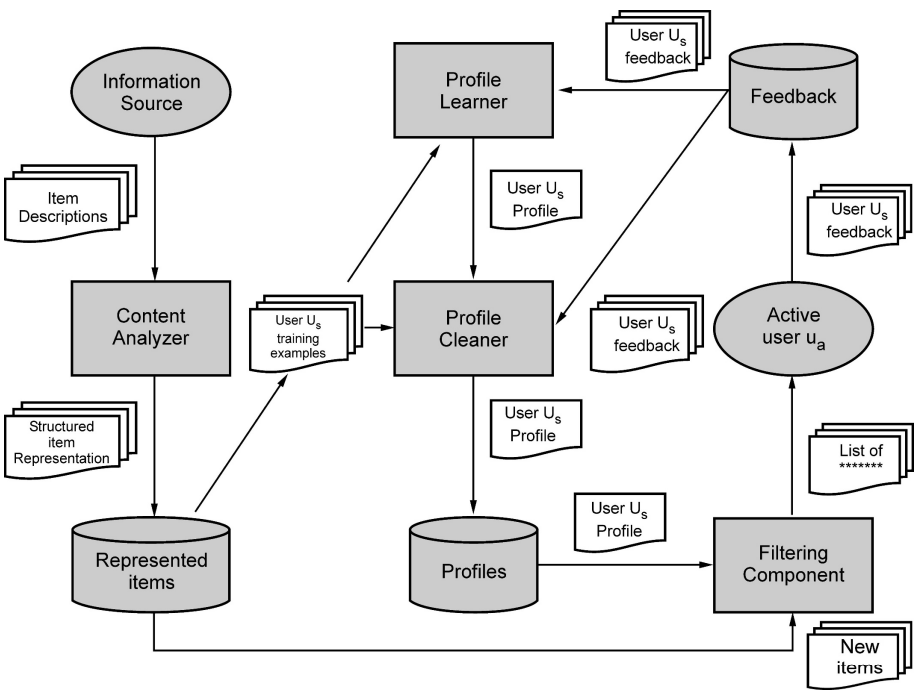


Fig. 5.4.1 : High Level Architecture Content-based Recommender Systems

➤ 1. Content Analyzer

- Extracts the features (keywords, n-grams) from the source.
- Conversion from unstructured to structured item.
- Data stored in the repository Represented Items

➤ 2. Profile Learner

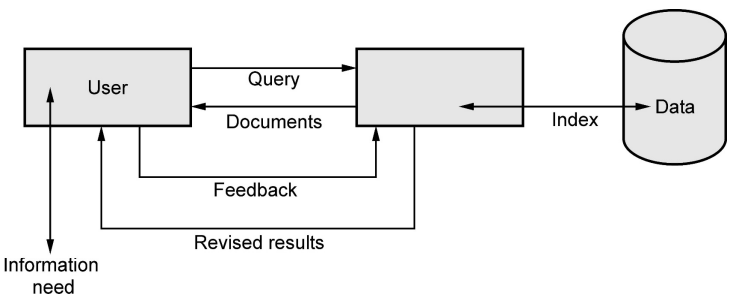
- To build user profile
- Updates the profile using the data in Feedback repository

➤ 3. Filtering Component

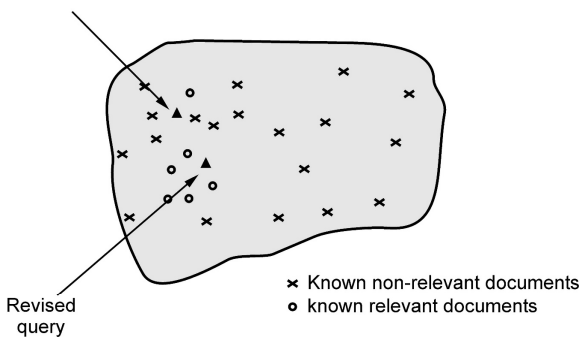
- Matching the user profile with the actual item to be recommended
- Uses different strategies
- Users have no detailed knowledge of collection makeup and the retrieval environment. Most users often need to reformulate their queries to obtain the results of their interest.

➥ 5.4.2 Relevance Feedback

- Thus, the first query formulation should be treated as an initial attempt to retrieve relevant information. Documents initially retrieved could be analyzed for relevance and used to improve the initial query.
- Fig. 5.4.2 shows relevance feedback on initial query.



(a) Relevance feedback



(b) Relevance feedback on initial query

Fig. 5.4.2

- The process of query modification is commonly referred as Relevance feedback, when the user provides information on relevant documents to a query.
- The process of query modification is commonly referred as Query expansion, when information related to the query is used to expand it.
- The user issues a short and simple query. The search engine returns a set of documents. User marks some docs as relevant, some as non-relevant.
- **Characteristics of relevance feedback :**
  1. It shields the user from the details of the query reformulation process.
  2. It breaks down the whole searching task into a sequence of small steps which are easier to grasp.
  3. Provide a controlled process designed to emphasize some terms (relevant ones) and de-emphasize others (non-relevant ones).
- **Issues with relevance feedback :**
  1. The user must have sufficient knowledge to form the initial query.
  2. This does not work too well in cases like : Misspellings, CLIR and mismatch in user's and document's vocabulary.
  3. Relevant documents have to be similar to each other while similarity between relevant and non-relevant document should be small.
  4. Long queries generated may cause long response time.
  5. Users are often reluctant to participate in explicit feedback.
- **Advantages of relevance feedback**
  1. Relevance feedback usually improves average precision by increasing the number of good terms in the query.
  2. It breaks down the search operation into a sequence of small search steps, designed to approach the wanted subject area gradually.
  3. It provides a controlled query alteration process designed to emphasize some terms and to deemphasize others, as required in particular search environments.
- **Disadvantages of relevance feedback**
  1. More computational work
  2. Easy to decrease precision
- **Two basic approaches of feedback methods :**
  1. **Explicit feedback :** The information for query reformulation is provided directly by the users. However, collecting feedback information is expensive and time consuming.

     The accuracy of recommendation depends on the quantity of ratings provided by the user.

2. **Implicit feedback :** The information for query reformulation is implicitly derived by the system. Implicit feedback reduces the burden on users by inferring their user's preferences from their behavior with the system.

5.4.3 Advantages and Drawbacks of Content-based Filtering

Advantages :

1. User Independence : Recommends only the items that interest the user
2. Transparency : Recommendation is based on the item features, explicitly list the contents features
3. New Item : Helps in recommending new items that are not yet rated by other users

Drawbacks :

1. The user will never be recommended for different items.
2. Business cannot be expanded as the user does not try a different type of product.
3. Overspecialization: Recommends those items that score high with the user profile
4. Cold Start Problem: For a new user, systems don't have historical information to recommend items

5.5 Collaborative Filtering

AU : May-17, Dec.-17

- Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a single user by collecting preferences or taste information from many users (collaborating).
- Collaborative filtering (CF) uses given rating data by many users for many items as the basis for predicting missing ratings and/or for creating a top-N recommendation list for a given user, called the active user. Formally, we have a set of users $U = \{u_1, u_2, \ldots, u_m\}$ and a set of items $I = \{ i_1, i_2, \ldots, i_n\}$. Ratings are stored in a $m \times n$ user-item rating matrix.
- The problem of collaborative filtering is to predict how well a user will like an item that he has not rated given a set of historical preference judgments for a community of users.

5.5.1 Type of CF

- There are two types of collaborative filtering algorithms : user based and item based.

1. User based

- User-based collaborative filtering algorithms work off the premise that if a user (A) has a similar profile to another user (B), then A is more likely to prefer things that B prefers when compared with a user chosen at random.

- The assumption is that users with similar preferences will rate items similarly. Thus missing ratings for a user can be predicted by first finding a neighborhood of similar users and then aggregate the ratings of these users to form a prediction.
- The neighborhood is defined in terms of similarity between users, either by taking a given number of most similar users (k nearest neighbors) or all users within a given similarity threshold. Popular similarity measures for CF are the Pearson correlation coefficient and the Cosine similarity.
- For example, a collaborative filtering recommendation system for television tastes could make predictions about which television show a user should like given a partial list of that user's tastes (likes or dislikes).
- Note that these predictions are specific to the user, but use information gleaned from many users. This differs from the simpler approach of giving an average score for each item of interest, for example based on its number of votes.
- User-based CF is a memory-based algorithm which tries to mimics word-of-mouth by analyzing rating data from many individuals.
- The two main problems of user-based CF are that the whole user database has to be kept in memory and that expensive similarity computation between the active user and all other users in the database has to be performed.

➤ 2. Item-based collaborative filtering

- Item-based CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix. The assumption behind this approach is that users will prefer items that are similar to other items they like.
- The model-building step consists of calculating a similarity matrix containing all item-to-item similarities using a given similarity measure. Popular are again Pearson correlation and Cosine similarity. All pair-wise similarities are stored in n × n similarity matrix S.
- Item-based collaborative filtering has become popularized due to its use by YouTube and Amazon to provide recommendations to users. This algorithm works by building an item-to-item matrix which defines the relationship between pairs of items.
- When a user indicates a preference for a certain type of item, the matrix is used to identify other items with similar characteristics that can also be recommended.

- Item-based CF is more efficient than user-based CF since the model is relatively small ($N \times k$) and can be fully pre-computed. Item-based CF is known to only produce slightly inferior results compared to user-based CF and higher order models which take the joint distribution of sets of items into account are possible. Furthermore, item-based CF is successfully applied in large scale recommender systems (e.g., by Amazon.com).

➥ 5.5.2 Collaborative Filtering Algorithms

➤ 1. Memory-based algorithms :

- Operate over the entire user-item database to make predictions.
- Statistical techniques are employed to find the neighbors of the active user and then combine their preferences to produce a prediction.
- Memory-based algorithms utilize the entire user-item database to generate a prediction. These systems employ statistical techniques to find a set of users, known as neighbors that have a history of agreeing with the target user.
- Once a neighborhood of users is formed, these systems use different algorithms to combine the preferences of neighbors to produce a prediction or top-N recommendation for the active user. The techniques, also known as nearest-neighbor or user-based collaborative filtering are more popular and widely used in practice.
- Dynamic structure. More popular and widely used in practice.

Advantages

1. The quality of predictions is rather good.
2. This is a relatively simple algorithm to implement for any situation.
3. It is very easy to update the database, since it uses the entire database every time it makes a prediction.

Disadvantages

1. It uses the entire database every time it makes a prediction, so it needs to be in memory it is very, very slow.
2. Even when in memory, it uses the entire database every time it makes a prediction, so it is very slow.
3. It can sometimes not make a prediction for certain active users/items. This can occur if the active user has no items in common with all people who have rated the target item.
4. Overfits the data. It takes all random variability in people's ratings as causation, which can be a real problem. In other words, memory-based algorithms do not generalize the data at all.

➤ **2. Model-based algorithms :**

- Input the user database to estimate or learn a model of user ratings, then run new data through the model to get a predicted output.
- A prediction is computed through the expected value of a user rating, given his/her ratings on other items.
- Static structure. In dynamic domains the model could soon become inaccurate.
- Model-based collaborative filtering algorithms provide item recommendation by first developing a model of user ratings. Algorithms in this category take a probabilistic approach and envision the collaborative filtering process as computing the expected value of a user prediction, given his/her ratings on other items.
- The model building process is performed by different machine learning algorithms such as Bayesian network, clustering and rule-based approaches. The Bayesian network model formulates a probabilistic model for collaborative filtering problem.
- The clustering model treats collaborative filtering as a classification problem and works by clustering similar users in same class and estimating the probability that a particular user is in a particular class C and from there computes the conditional probability of ratings.
- The rule-based approach applies association rule discovery algorithms to find association between co-purchased items and then generates item recommendation based on the strength of the association between items

Advantages

1. Scalability : Most models resulting from model-based algorithms are much smaller than the actual dataset, so that even for very large datasets, the model ends up being small enough to be used efficiently. This imparts scalability to the overall system.
2. Prediction speed : Model-based systems are also likely to be faster, at least in comparison to memory-based systems because, the time required to query the model is usually much smaller than that required to query the whole dataset.
3. Avoidance of over fitting : If the dataset over which we build our model is representative enough of real-world data, it is easier to try to avoid over-fitting with model-based systems.

Disadvantages

1. Inflexibility : Because building a model is often a time- and resource-consuming process, it is usually more difficult to add data to model-based systems, making them inflexible.

2. Quality of predictions : The fact that we are not using all the information (the whole dataset) available to us, it is possible that with model-based systems, we don't get predictions as accurate as with model-based systems. It should be noted, however, that the quality of predictions depends on the way the model is built. In fact, as can be seen from the results page, a model-based system performed the best among all the algorithms we tried.

➥ 5.5.3 Advantages and Disadvantages

➤ Advantages

1. Collaborative filtering application is to recommend interesting or popular information as judged by the community.
2. Collaborative filtering system can make more personalized recommendation by analyzing information from your past activity or the history of other users of similar taste.

➤ Disadvantages

1. Many commercial recommender systems are based on large datasets. As a result, the user-item matrix used for collaborative filtering could be extremely large and sparse, which brings about the challenges in the performances of the recommendation.
2. As the numbers of users and items grow, traditional CF algorithms will suffer serious scalability problems.
3. Gray sheep refers to the users whose opinions do not consistently agree or disagree with any group of people and thus do not benefit from collaborative filtering.
4. A collaborative filtering system doesn't automatically match content to one's preferences

➥ 5.5.4 Difference between Collaborative Filtering and Content based Filtering

Collaborative Filtering	Content Filtering
Collaborative-Filtering systems focus on the relationship between users and items.	Content-Based systems focus on properties of items.
Example : Netflix movie recommendations	Example : Pandora.com music recommendations
Pro : Does not assume access to side information about items	Con : Assumes access to side information about items
Cannot recommend new items	It can recommend new items

Collaborative Filtering	Content Filtering
Item features are inferred from ratings.	Match the item features with user preferences.
Con: Does notwork on new items that have no ratings	Pro: Got a new item to add? No problem, just be sure to include the side information.

University Questions

1. Explain collaborative filtering and content based recommendation system with an example. **AU : May-17, Marks 16**
2. Explain in detail, the collaborative filtering using clustering technique. **AU : Dec.-17, Marks 10**
3. Explain in detail, the collaborative filtering using clustering technique. **AU : Dec-17, Marks 10**
4. Explain collaborative filtering recommendation system with an example. **AU : May-17, Marks 8**

5.6 Matrix Factorization Models

- Matrix factorization (MF) models are based on the latent factor model. MF approach is most accurate approach to reduce the problem from high levels of sparsity in RS database.
- Matrix factorization is a simple embedding model. Given the feedback matrix $A \in R^{m \times n}$, where m is the number of users (or queries) and n is the number of items, the model learns :
  1. A user embedding matrix $U \in R^{m \times d}$, where row i is the embedding for user i.
  2. An item embedding matrix $V \in R^{n \times d}$, where row j is the embedding for item j.

5.6.1 Singular Value Decomposition (SVD)

- SVD is a matrix factorization technique that is usually used to reduce the number of features of a data set by reducing space dimensions from N to K where K < N.
- The matrix factorization is done on the user-item ratings matrix.

$$\underset{m \times n}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}} = \underset{m \times r}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}} \underset{r \times r}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}} \underset{r \times n}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}$$

- The matrix S is a diagonal matrix containing the singular values of the matrix X. There are exactly r singular values, where r is the rank of X.
- The rank of a matrix is the number of linearly independent rows or columns in the matrix. Recall that two vectors are linearly independent if they can not be written as the sum or scalar multiple of any other vectors in the space.

➤ Incremental SVD Algorithm (SVD++)

- The idea is borrowed from the Latent Semantic Indexing (LSI) world to handle dynamic databases.
- LSI is a conceptual indexing technique which uses the SVD to estimate the underlying latent semantic structure of the word to document association.
- Projection of additional users provides good approximation to the complete model
- SVD based recommender systems has following limitations
  a. It can not be applied on sparse data
  b. Does not have regularization

5.7 Neighbourhood Models

The most common approach to CF is the neighborhood-based approach. Its original form, which was shared by virtually all earlier CF systems, is the user-oriented approach. Such user-oriented methods estimate unknown ratings based on recorded ratings of like minded users.

5.7.1 Similarity Measures

- In order to cluster the items in a data set, some means of quantifying the degree of association between them is required. This may be a distance measure, or a measure of similarity or dissimilarity.
- The relationship between the document is described by
  a. Similarity : These values indicate how much two documents or objects are near to each other.
  b. Association : It is same as similarity but difference is objects which are considered for comparison are object characterized by discrete state attributes.
  c. Dissimilarity : Dissimilarity value shows that how much far the objects are.
- The measure of similarity is designed to quantify the likeness between objects so that if one assumes it is possible to group objects in such a way that an object in a group is more like the other members of the group than it is like any object outside the group, then a cluster method enables such a group structure to be discovered.

- A measure of association increases as the number or proportion of shared attribute states increases. In information retrieval system, two documents will be similar to each other if they have more number of common index terms.
- If two documents are having less number of common index terms then obviously they will be semantically far from each other. So we will not include such documents in the same group.
- There are five commonly used measures of association in information retrieval. Since in information retrieval documents and requests are most commonly represented by term or keyword lists. A query is also represented by a list of keywords. Each list of keyword is considered as one set.

5.8 Part A : Short Answered Questions [2 Marks Each]

Q.1 What is collaborative filtering ?

▣ **Ans. :** Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a single user by collecting preferences or taste information from many users (collaborating). It uses given rating data by many users for many items as the basis for predicting missing ratings and/or for creating a top-N recommendation list for a given user, called the active user.

Q.2 What do you mean by item-based collaborative filtering ? **AU : May-17**

▣ **Ans. :** Item-based CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix. The assumption behind this approach is that users will prefer items that are similar to other items they like.

Q.3 What are problem of user based CF ?

▣ **Ans. :** The two main problems of user-based CF are that the whole user database has to be kept in memory and that expensive similarity computation between the active user and all other users in the database has to be performed.

Q.4 Define user based collaborative filtering. **AU : Dec.-16**

▣ **Ans. :** User-based collaborative filtering algorithms work off the premise that if a user (A) has a similar profile to another user (B), then A is more likely to prefer things that B prefers when compared with a user chosen at random.

Q.5 What is cosine similarity ?

▣ **Ans. :** This metric is frequently used when trying to determine similarity between two documents. Since there are more words that are in common between two documents, it is useless to use the other methods of calculating similarities

Q.6 What are the characteristics of relevance feedback ?

▣ **Ans. : Characteristics of relevance feedback :**

1. It shields the user from the details of the query reformulation process.
2. It breaks down the whole searching task into a sequence of small steps which are easier to grasp.
3. Provide a controlled process designed to emphasize some terms (relevant ones) and de-emphasize others (non-relevant ones)

Q.7 Write goal of recommender system.

▣ **Ans. :** The goal of a recommender system is to generate meaningful recommendations to a collection of users for items or products that might interest them.

Q.8 Define recommender systems.

▣ **Ans. :** Recommender Systems are software tools and techniques providing suggestions for items to be of use to a user. The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read.

Q.9 What is demographic based recommender system ?

▣ **Ans. :** This type of recommendation system categorizes users based on a set of demographic classes. This algorithm requires market research data to fully implement. The main benefit is that it doesn't need history of user ratings.

Q.10 What is Singular Value Decomposition (SVD) ?

▣ **Ans. :** SVD is a matrix factorization technique that is usually used to reduce the number of features of a data set by reducing space dimensions from N to K where K < N.

Q.11 What is Content-based recommender ?

▣ **Ans.:**Content-based recommenders refer to such approaches, that provide recommendations by comparing representations of content describing an item to representations of content that interests the user. These approaches are sometimes also referred to as content-based filtering.

Q.12 What is matrix factorization model ?

▣ **Ans. :** Matrix factorization is a class of collaborative filtering algorithms used in recommender systems. Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices

5.9 Multiple Choice Questions

Q.1 _____ factorization is a class of collaborative filtering algorithms used in recommender systems.

[a] Matrix [b] Row [c] Columns [d] All of these

Q.2 SVD is a matrix factorization technique that is usually used to reduce the number of features of a data set by reducing space dimensions from N to K where _____.

[a] K > N [b] K = N [c] K < N [d] K >= N

Q.3 _____ CF is a model-based approach which produces recommendations based on the relationship between items inferred from the rating matrix.

[a] User based [b] Item based

[c] Both (a) and (b) [d] None

Q.4 _____ filtering is a method of making automatic predictions about the interests of a single user by collecting preferences or taste information from many users.

[a] Information [b] Data [c] Collaborative [d] None

Q.5 _____ filtering can recommend new items.

[a] Content [b] Collaborative [c] Information [c] Data

Q.6 Which of the following is the components of High Level Architecture of Content-based Recommender Systems ?

[a] Content Analyzer [b] Profile Learner

[c] Filtering Component [d] All of these

➤ Answer Keys for Multiple Choice Questions

Q.1	a	**Q.2**	c	**Q.3**	b	**Q.4**	c
Q.5	a	**Q.6**	d				

❑❑❑

Notes

SOLVED MODEL QUESTION PAPER

(As Per New Syllabus)

Information Retrieval Techniques

Semester - VIII (CSE/IT) Professional Elective - V

Time : Three Hours] **[Maximum Marks : 100**

Answer All Questions

PART - A **(10 × 2 = 20 Marks)**

Q.1 *Explain the type of natural language technology used in information retrieval.*
(Refer Two Marks Q.5 of Chapter - 1)

Q.2 *What is conflation ?* **(Refer Two Marks Q.7 of Chapter - 1)**

Q.3 *What are the assumption of vector space model ?*
(Refer Two Marks Q.6 of Chapter - 2)

Q.4 *Explain Luhn's ideas.* **(Refer Two Marks Q.9 of Chapter - 2)**

Q.5 *Define pre pruning and post pruning.* **(Refer Two Marks Q.5 of Chapter - 3)**

Q.6 *Define information gain.* **(Refer Two Marks Q.4 of Chapter - 3)**

Q.7 *Define search engine optimization.* **(Refer Two Marks Q.5 of Chapter - 4)**

Q.8 *What are politeness policies used in web crawling ?*
(Refer Two Marks Q.10 of Chapter - 4)

Q.9 *What is collaborative filtering ?* **(Refer Two Marks Q.1 of Chapter - 5)**

Q.10 *What are the characteristics of relevance feedback ?*
(Refer Two Marks Q.6 of Chapter - 5)

PART - B **(5 × 13 = 65 Marks)**

Q.11 a) *i) Explain in detail about the components of IR.* **(Refer section 1.4)** **[6]**

ii) Explain various method used for visualization of search engine .

(Refer section 1.8) **[7]**

OR

b) *i) What is difference between data retrieval and information retrieval ?*

(Refer section 1.3.1) **[5]**

ii) Appraise the history of information retrieval. **(Refer section 1.1)** **[8]**

Q.12 a) *Explain following IR models.*

a. Boolean model *b. Vector model* **(Refer sections 2.1.2 and 2.1.3)** **[13]**

OR

b) *i) What is relevance feedback ? Explain with an example an algorithm for relevance feedback.* **(Refer section 2.6)** **[7]**

ii) What is TREC collection ? List benefits. **(Refer section 2.7.1)** **[6]**

Q.13 a) *i) What is machine learning ? Explain supervised, unsupervised learning.* **(Refer section 3.1)** **[6]**

ii) Explain K-mean clustering. **(Refer section 3.2.2)** **[7]**

OR

b) *i) What is dimensionality reduction ? Explain its advantages and disadvantages.* **(Refer section 3.4)** **[8]**

ii) Write a short note on inverted indexes. **(Refer section 3.7.1)** **[5]**

Q.14 a) *i) What is web and web browser ? Explain characteristics of web.* **(Refer section 4.1)** **[6]**

ii) Draw and explain Harvest architecture. Also explain features. **(Refer section 4.2.2)** **[7]**

OR

b) *i) Describe search engine user interaction.* **(Refer section 4.4)** **[6]**

ii) What is web crawler ? Explain web crawler architecture. **(Refer section 4.6.1)** **[7]**

Q.15 a) *i) What is recommender systems ? What are the challenges of recommender systems ?* **(Refer section 5.1)** **[6]**

ii) Explain various collaborative filtering algorithms. **(Refer section 5.5.2)** **[7]**

OR

b) *What is content-based recommender systems ? Draw and explain high level architecture content-based recommender systems.* **(Refer section 5.4)** **[13]**

PART - C **(1 × 15 = 15 Marks)**

Q.16 a) *Consider the following six training examples, where each example has three attributes : color, shape and size. Color has three possible values : red, green and blue. Shape has two possible values : square and round. Size has two possible values : big and small.*

Example	Color	Shape	Size	Class
1	red	square	big	+
2	blue	square	big	+
3	red	round	small	–
4	green	square	small	–
5	red	round	big	+
6	green	square	big	–

Which is best attribute for the root node of decision tree ? **(Refer example 3.3.1)** **[15]**

OR

b) *i) Assume the following : A database contains 80 records on a particular topic. A search was conducted on that topic and 60 records were retrieved of the 60 records retrieved, 45 were relevant. Calculate the precision and recall scores for the search.*

(Refer example 3.5.1) **[7]**

ii) Discuss about search interfaces today. How query specification process is used ?
(Refer section 1.7) **[8]**

❑❑❑

Notes

Notes

Notes