Bibliometrics, Citation Analysis…

INSC 702: Advanced Topics in Information Science

Shimelis Assefa

Outline

- Bibliometrics, Scientometrics, Informetrics, Webometrics, Altmetric
- Citation Analysis
- Bradford’s law of bibliographic Scattering
- Lotka’s Law of Scientific Productivity
- Zipf’s Law of Word Frequency
- Webometrics

Bibliometrics

- Biblio= books/bibliographies; metrics = measurement.
- Qualitative and quantitative methods to examine & analyze artifacts of communication.
- A method based on Information-as-thing.
- The theoretical objects appropriate to the method include –author, titles, institutions, & citations.

Bibliometrics…

- The quantitative application may identify patterns of:
  - Word use, vocabulary.
  - Locating high or low frequency occurrences of specific phrases, words, or structures.
  - Recurring patterns in data, subject, queries, citations, authorship, publication data, themes, characters.
  - Suggest researchable issues or relationships.

Bibliometrics…

- Bibliometrics may also be used to identify – title clusters, journal clusters, discipline, & diffusion networks.
- Wide range of citation studies –mapping scientific domains, including growth, diffusion, specialization, collaboration, impact, & obsolescence of literature and concepts.

Bibliometrics…

- Practical uses of bibliometrics include contributing information for decision-making
  - Collection development
  - Weeding
  - Cataloguing & classification
  - Circulation patterns
  - Document retrieval systems
  - User preferences

Bibliometrics…

- Measurement provides
  - A method for describing an entity.
  - Descriptive information, such as how frequently a word is used or how frequently an author publishes.
  - Comparisons based on quantity, frequency, length, & characteristics of quality
- Bibliometrics studies deal with 3 components:
  - The physical object,
  - Its creation & subject content, and
  - Its use.

Bibliometrics…

- Physical object – e-communication & media, such as radio, TV, film, e-mail, e-journals,

  e-publications
  - Permit retrospective examination & evaluation
- Creation & content evaluation – examines the productivity of authors, currency of content, content coverage, & spread of content.
- Studying uses & users –most complex of bibliometrics studies.
  - There are compound factors involved that seem to defy quantification, such as human info seeking behavior.

Bibliometrics…

- Bibliometric studies potentially reveal:
  - Maps of literature or communication.
  - Who cites whom, & overall paths of information. exchange or communication networks, & patterns of scientific scholarly communication.
- Measurement: objects & representatives of information
  - One method evaluates the history of the objects.
  - How many of X objects in Y context are produced now versus previously?
  - Predicting future shelf space & storage problems.

Bibliometrics…

- Citation half-life and use half-life
  - Method to evaluate the usefulness of specific journals.
  - Attempts to evaluate the length of time in which a journal is used as a resource as indicated by citation, or physical use as indicated by circulation.
  - To determine when half of all the active literature of a field has been published, or the time period that includes half or more of the references made.
- Half-life studies may suggest
  - Which journals are most used.
  - Which journals are least used, for discontinuation purpose.

Bibliometrics…

- Measurement: creation & content
  - Counting the no. of times a specific article or author is cited.
  - This may indicate the author has some authority on a subject.
  - Author productivity.
  - Simple count of publications, however, doesn’t indicate quality.
- Central interest for Bibliometrics is the study of disciplinary development as signified by measures of production & use of scholarly literature.

Informetrics, Scientometrics, Altmetrics

- Journal of Informetrics
- Journal of Scientometric Research or
- Scientometrics
- Webometrics – e.g., ranking of universities
- Altmetric - https://www.altmetric.com/

Citation Analysis

- The study of a particular kind of sign within texts that indicates information has been used.
- Citation establishes relationship between the citing and cited documents.
- This relationship may also clarify aspects of aboutness of each document.
- CA is a powerful means of investigating the intellectual & paradigmatic structure of a scholarly discipline.
- It can reveal conflicts & alliances, agreed-upon & contested ideas, influences, schools of thought, & institutional arrangements.
- A text’s citation (both from it & to it) locate that text in an intellectual space.

Citation…

- Four fundamental methods pertinent to CA:
  - Simple counting of the no. of citations to a given item.
  - Counting the average no. of citations to a given item in some specified length of time (citation impact).
  - Bibliographic coupling
  - Co-citation
- Every one of the 4 phenomena can be measured & ranked quantitatively.
- Each of the phenomena is indicator of the relevance of the documents to one another & potentially of documents to authors & readers.
- Each can be used to provide insight into the social & intellectual structure of subject disciplines.

Citation…

- The nature and meaning of the relation between citing & cited documents is subject to questions.
- Not all citations are of equal value.
- There are many reasons why an author may cite another text.
- The citation may be positive, negative, necessary, or pointless.
- It may be done to authenticate a claim, give credit where credit is due, levy a criticism, or serve less than noble reasons.

Citation…

- Six assumptions underlie the performance of CA
  - How cited documents are used
  - A cited text be related to the content of the citing text
  - A document that is frequently cited is assumed to possess some quality that makes it seminal for the social & intellectual practice of a given discipline.
  - Conversely, an infrequently cited text is without merit
  - Citations should always be made to the most important & most useful texts available.
  - Some citations are ‘more equal’ than others.

Citation…

- In general, CA is used to identify:
  - The most discussed issues in a field & topic area.
  - It may also be a way to establish what the focus of interests, or new topics, in a discipline were at a given time or the influences affecting the field.
  - Evaluating word & phrase frequencies may provide some insight into the ‘aboutness’ or the subject matter.
  - Is also used to evaluate documents for subject content
  - Determining the subject content of an item theoretically may be accomplished by examining & distilling the content of a document into representative characteristics.

eigenFACTOR.org™

RANKING AND MAPPING SCIENTIFIC KNOWLEDGE

Bradford Law

- Samuel Bradford, 1934, Bradford’s law of scatter.
- Identified a pattern about author concentration.
- A small no. of journals in a specific field will contain the highest concentration of articles, while a larger no. of journals will have a lower concentration.
- This ‘core & scatter’ translates into a pattern wherein a small core of journals contain the majority of articles on a given topic.
- Bradford theorized that a core of journals & the zones of scatter for a subject could be determined.
- This permits one to select the fewest journals that would provide the best coverage of a topic.

Bradford…

- Bradford’s law of scatter is based on the observation that collection use is rarely evenly distributed.
- Generally, a few documents will be used quite often, some documents will be used sometimes, & a great number of documents will hardly be used at all.
- Thus, for any given subject & collection of relevant journals, Bradford’s law states that it is possible to divide these journals into **zones** that, while containing the same number of articles on the subject, the number of journals needed to yield that number of articles increases substantially from one zone to the next.

- **Mathematically, it is expressed as:**

  If R(*n*)= *total # of journal articles*

  N= total # of journals, and

  s= a constant specific to a given subject discipline, then

  $R(n) = N \log {}^{n} / {}_{s} \; (1 \leq n \leq N)$

Bradford…

- Examples of Bradford distribution

Zone	#of Journals	#of Articles
1	10	400
2	60	400
3	250	400

- Bradford’s law is the basis of the well known 80/20 rule.
- Most of the information relevant to any given topic will be found in a few sources.

Lotka’s Law

- Lotka’s law of author productivity, Alfred Lotka.
- For any body of scientific literature, represented by articles in scientific journals,
  - a substantial number of authors will contribute
  - a single publication,
  - a smaller number will contribute some articles, and
  - a few will contribute a great number of articles.
- Mathematically, Lotka’s law is expressed as:

  Number of authors writing **n articles = 1/n^2 (number of authors writing one article).**
- **Inverse square relation.**

Lotka’s…

- Lotka’s law implies that any scientific discipline tends to be dominated by the work of few people.
- Knowing these people will be very important for collection development & IR purposes.
- Lotka’s law has its own limitations.

Zipf’s Law

- Zipf’s law of the distribution of words in texts, George Zipf.
- Based on the intuitive notion that people tend to use only a small part of the vocabulary available to them for communication.
- For any text, it is possible to count the appearances of any word and to rank words by their frequency of appearance.
- Zipf’s law says that for any given text, on any subject, if a rank order of the word’s frequency is multiplied by the number of times it appears in the text, then the result is a constant.

Zipf's…

- Mathematically, it is expressed as:

If f = frequency of appearance of a word, and R = rank of the frequency of appearance of a word, then

(f) (r) = c

Where c is a constant for a given text.

Zipf’s…

- Zipf’s law has implications for indexing and IR.
- Underlying Zipf’s law is economic of effort.
- Language is a phenomena that seeks to convey the most information with the fewest number of words.
- Lotka’s law of scientific productivity
- Bradford’s law of bibliographic scattering
- Zipf’s law of word frequencies in texts.

H-index

- A method to characterize or evaluate the scientific output of a researcher.
- A scientist has index h if h of his or her N_p papers have at least h citations each and the other ($N_p - h$) papers have ≤h citations each.
- The research reported here concentrated on physicists.

H-index…

- Measures productivity.
- Does not measure importance or impact of papers.
- The relation between $N_{c,\ tot}$ and ***h*** will depend on the detailed form of the particular distribution.
- And it is useful to define the proportionality constant **a**
  - $N_{c,\ tot} = ah^2$
  - A ranges between 3 and 5.

Webometrics

- The quantitative study of Web-related phenomena.
- The method has its roots in Bibliometric.
- The object of study is – the Web.
- The challenge is to ascertain the extent to which useful information about science communication can be extracted from hyperlinks.
- Hyperlinks as a potential source for new information.
- Quantitative studies of hyperlinks have broader potential applications compared to bibliometric studies of citations.
- Improving Web search engine performance.

Webometrics…

- Current Webometric research focus on
  - Web page content analysis
  - Web link structure analysis
  - Web usage analysis (exploiting log files of users' searching & browsing behavior)
  - Web technology analysis (including search engine performance)
- Data is gathered from search engines.
- Webometrics, Cybermetrics, Scientometrics, Informetrics

- According to Björneborn and Ingwersen (2004) webometrics is “The study of quantitative aspects of construction and use of information resources, structures and technologies on the Web drawing on bibliometric and infometric approaches.” Webometrics has four areas of research: webpage content analysis, web link structure analysis, web usage analysis and web technology analysis. These four areas of research are described by Jala, Biswas and Mukhopadhyay (2009) as the following:

  “webpage analysis content includes automatic categorization of web pages and texts using different search engines and tools for web analysis;

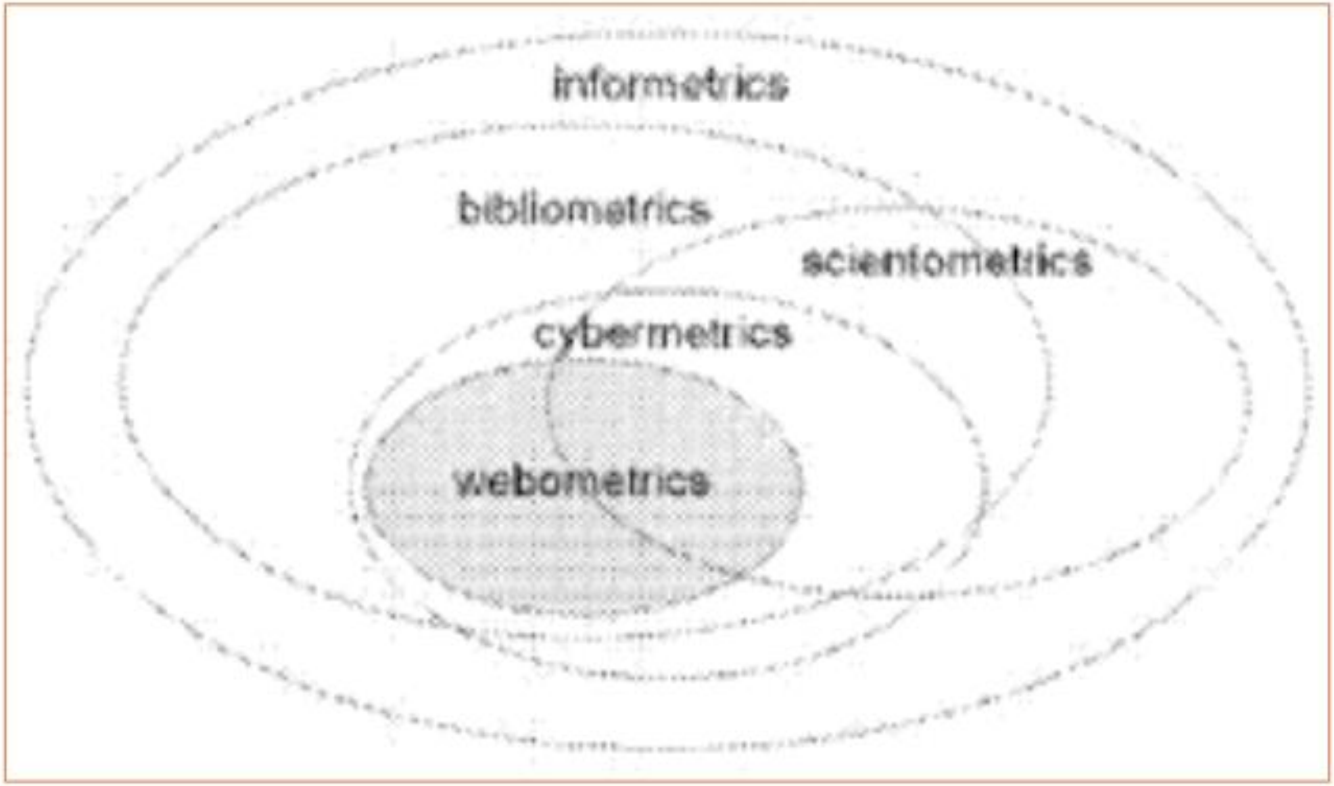
  web link structure analysis includes the categorization of hyperlinks and inlinks, self-links and external links to a particular website, patterns of linking, etc;

  web usage analysis which includes the exploitation of log files for users’ searching and browsing behavior; and

  web technology analysis includes the performance of search engines with respect to information retrieval and supporting webometric analysis.”

Webometrics…

- Relationships between the various metrics in IS.



Webometrics…

CYBER *metrics*

International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

- **Cybermetrics** focus is on methodologies and results of webometric, scientometric, bibliometric or informetric research with emphasis placed on aspects related to Internet:
- Application of **scientometric methodology** to the analysis of scientific communication in the Web and the Usenet, including the development of new Internet R&D indicators.

Webometrics…

- **Analyses** of hypertext linking phenomena, the informetric laws and distributions, and any mathematical model derived.
- **Impact of the Internet** on scientific co-operation and other aspects related to science organization, information flows and interdisciplinary connections.
- Evaluation of **electronic scientific journals** and peer-review processes in the World Wide Web.