The Symbiotic Relationship Between

Information Retrieval and Informetrics

Dietmar Wolfram

School of Information Studies

University of Wisconsin-Milwaukee

P.O. Box 414, Milwaukee, WI U.S.A. 53201

Email: dwolfram@uwm.edu

Phone: 00 1 414 229 6836

Fax: 00 1 414 229 6699

Preprint of forthcoming paper to appear (2015) in *Scientometrics*

DOI: 10.1007/s11192-014-1479-0

ABSTRACT:

Informetrics and information retrieval (IR) represent fundamental areas of study within information science. Historically, researchers have not fully capitalized on the potential research synergies that exist between these two areas. Data sources used in traditional informetrics studies have their analogues in IR, with similar types of empirical regularities found in IR system content and use. Methods for data collection and analysis used in informetrics can help to inform IR system development and evaluation. Areas of application have included automatic indexing, index term weighting and understanding user query and session patterns through the quantitative analysis of user transaction logs. Similarly, developments in database technology have made the study of informetric phenomena less cumbersome, and recent innovations used in IR research, such as language models and ranking algorithms, provide new tools that may be applied to research problems of interest to informetricians. Building on the author's previous work (Wolfram 2003), this paper reviews a sample of relevant literature published primarily since 2000 to highlight how each area of study may help to inform and benefit the other.

Keywords: Information storage and retrieval, Informetrics, Indexing, Research impact, User search patterns

Overview

Information retrieval (IR) and informetrics represent two foundational research areas in information science. IR addresses issues relevant to the collection, representation, storage, indexing, retrieval, and presentation of documentary content, in textual or other media formats. Informetrics and its allied "metric" areas (bibliometrics, scientometrics, cybermetrics web[o]metrics) quantitatively examine the production and use of recorded discourse (Wilson 1999). Both IR and informetrics, which at first glance may appear to be very different in nature, are in reality closely aligned. An understanding of regularities in information production and use, as studied in informetrics, can help to inform information retrieval system design, use, and evaluation. It's also possible to use methods developed for information retrieval to investigate and gain insight into informetric phenomena. As data sets used in metric studies continue to grow, the ability to apply IR techniques used for the representation, comparison and retrieval of data to informetric data becomes increasingly important. One could consider the connection between these two areas as symbiotic—not in a biological sense, but rather as a mutually beneficial relationship. This essay, which builds on an earlier monographic treatment of this topic (Wolfram 2003), focuses on the intersection of informetrics and information retrieval to outline how the former may be used to inform the development and evaluation of information retrieval systems, and how concepts used in information retrieval may help to further the research agenda of informetrics. Key relevant concepts in information retrieval are briefly outlined followed by a review of how research methods in informetrics may be applied to IR and how developments in IR research may be applied to research problems in informetrics.

What is Information Retrieval?

Information retrieval encompasses the processes of how information is represented, stored, accessed, retrieved and presented. IR systems represent implementations of these processes. Traditionally, IR has focused on system-centered aspects of the representation of documents (indexing, document structure), their storage in the system, and mechanisms by which queries made to the system are matched to document sets indexed by the system. In more recent decades, IR has also embraced the user side, examining user thought processes, behaviors, and interactions with IR systems to better inform system design. Early forms of IR centered on corpora of text-based documents or their surrogates. Today's IR systems also provide access to other media formats such as sound, images and video.

The use of an IR system involves an interactive process prompted by a user information need. Various models have been developed to frame the process of user-centered interactive retrieval (Xie 2008). In a nutshell, the user recognizes that the database or information store to which the IR system provides access may contain resources to fill this need. Based on the user's understanding of the subject domain of the information need and search knowledge of the IR system to be used, the user formulates an initial query that is submitted to the IR system. The query encapsulates the topicality of the information need, represented in textual retrieval systems by strings of characters that define one or more terms, operations (e.g., Boolean operators) or limiters (e.g., field searches, document format, language). Interaction with the IR system is facilitated by a user interface. Contents from the database that match the query terms are retrieved for the user to examine. Based on the perusal of candidate items, users may then further refine their queries to identify documents most relevant to their needs. The concept of relevance has been central to the efficacy and evaluation of an information retrieval system (Saracevic 1975), where relevance is an assessment of how well a retrieved document fills the user's information need. This interactive negotiation process with the system is repeated until the information need has been satisfactorily addressed or the user concludes that the IR system cannot fill this need, at which point the search process ends. These ideas are summarized in Figure 1, along with possible data of interest for metrics researchers.

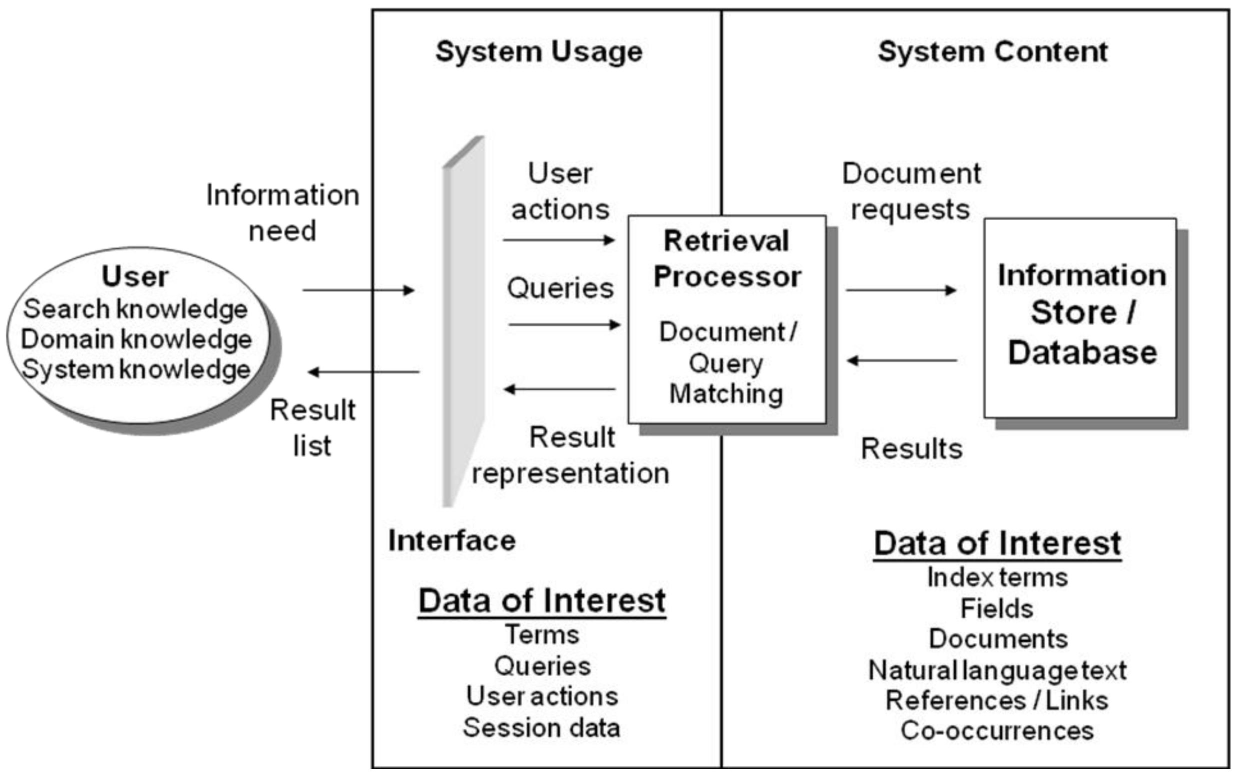

Fig. 1 Information retrieval processes and associated data

The theoretical foundations of IR are based in mathematics, with different types of models having been developed to aid in the matching process between queries and documents. Space limitations prevent a detailed presentation of IR models in this paper, however, three key model categories are briefly acknowledged. Among the earliest, the standard Boolean model makes use of set theory and Boolean operators to match queries to documents. Probabilistic models, which make use of probability for the relevance assessment of documents to queries, have also been long used to facilitate information retrieval (Fuhr 1992). The Vector Space Model (VSM), based on algebraic representations of queries and documents, has been one of the most widely studied approaches (Salton 1989). Essentially, the VSM treats documents and queries as vectors in a multidimensional document space. The relevance of a document to a query is assessed based on the proximity of the document vector to the query vector using measures based on angles or distances. The same measures can be used to assess how closely related documents are to one another. One drawback of this approach is that the retrieval comparison is based on a computationally intensive, high-dimensional space, with the dimensionality is determined by the size of the indexing vocabulary. Another limitation is that traditional IR models have made an assumption that terms co-occurring in queries and documents exist independently of one another. Under this assumption a term like “chocolate” is just as likely to occur with “automobile” as it is with “cake.” More recent models, such as probabilistically-based language models, that address the issue that words in documents do not occur independently of one another, have been applied in IR environments with promising results (Ponte and Croft 1998). As a further extension of language models, IR algorithms that integrate topic models—a type of statistical model initially used in machine learning and natural language processing, where documents are represented as mixtures of latent topics--have been found to be very effective (Blei and Lafferty 2009). The reduction to a comparatively small number of latent topics (perhaps, up to a few hundred), as opposed to a large number of indexing terms (usually, many thousands of terms), also helps to reduce the dimensionality of the document comparison problem.

IR systems have become much more ubiquitous with the wider availability of digital content. Early implementations of IR systems included bibliographic retrieval systems, which index bibliographic citations or the full text of documents, and online public access catalogues (OPACs), which index the contents of physical libraries. The importance of these databases as sources of bibliographic data for informetric studies has long been exploited by researchers. The availability of growing corpora of documents on the Internet prompted the development of search engines, or large scale IR systems that discover, index and provide access to primarily unstructured, full text documents that may be publicly available or privately held.

The processes associated with document indexing and retrieval lend themselves to quantitative study. It should come as no surprise that empirical regularities observed in print documents have also been found in electronic environments, both for system content characteristics and those related to system usage. The application of informetric inquiry to IR environments offers many research avenues for exploration for both system content and usage data.

The Relationship Between Informetrics and Information Retrieval

Quantitative studies of recorded discourse became more closely tied to information retrieval with the development of electronic IR systems. Bibliographic databases containing document surrogates or complete document representations have long provided raw data for informetric studies (e.g., Wormell 1998). Usually the focus of such studies has been on authors, documents, institutions, geographic affiliations, or citations. In particular, citation databases have provided vital datasets for the better understanding of scholarly communication, research impact, and the intellectual structure of disciplines. Thomson Reuters Web of Knowledge (http://wokinfo.com/), SciVerse Scopus (http://www.scopus.com), and Google Scholar (http://scholar.google.com) serve as key sources of data for citation-based studies. For sources not indexed by these providers, other regional or discipline-focused citation databases may be available.

On a more fundamental level, there are theoretical relationships between what IR and informetrics study. Egghe and Rousseau (1997) point to the parallels between the two areas. The concept of duality is central to informetrics, generalized by Egghe (1990) as information production processes, where objects of interest, called *sources*, produce *items*. As examples, authors produce articles, journals produce articles, and articles produce references. The relationship between citing and cited objects is a classic dual relationship. Egghe and Rousseau present parallels in the information retrieval environment, where symmetric or dual relationships exist between documents and queries, indexing and retrieval, and even between relevant and retrieved documents. This concept of duality allows ideas in informetrics to be applied in the IR environment.

Informetric Approaches for Information Retrieval Research

In this section, informetric approaches that can be applied to IR research are discussed. The areas of investigation may be divided into two broad categories: research that addresses aspects of system content, and research that relates to system usage. An earlier treatment of these topics appears in Wolfram (2003), chapters 4 and 5.

IR System Content

IR system content regularities have been investigated for over 50 years. The units of analysis may range from individual words or terms up to the documents themselves or inter-document characteristics such as links. Figure 1 outlines some of the features of IR systems and associated metric data that may be studied related to system content and system usage.

The smallest practical unit of investigation is the occurrence of an index term, which usually consists of a string of one or more alphanumeric characters. As a result of the automatic and possibly manual indexing of documents, one or more terms are identified as access points to a document. In IR environments where the full text of document contents is indexed, term frequency distributions follow a power law. The distribution may be represented in a classic Zipfian form using a rank-frequency display format, or a Lotkaian form using a size-frequency display format (Figure 2). In both forms, the data distribution exhibits an inverse relationship, where few terms occur with high frequency and many terms occur with low frequency.

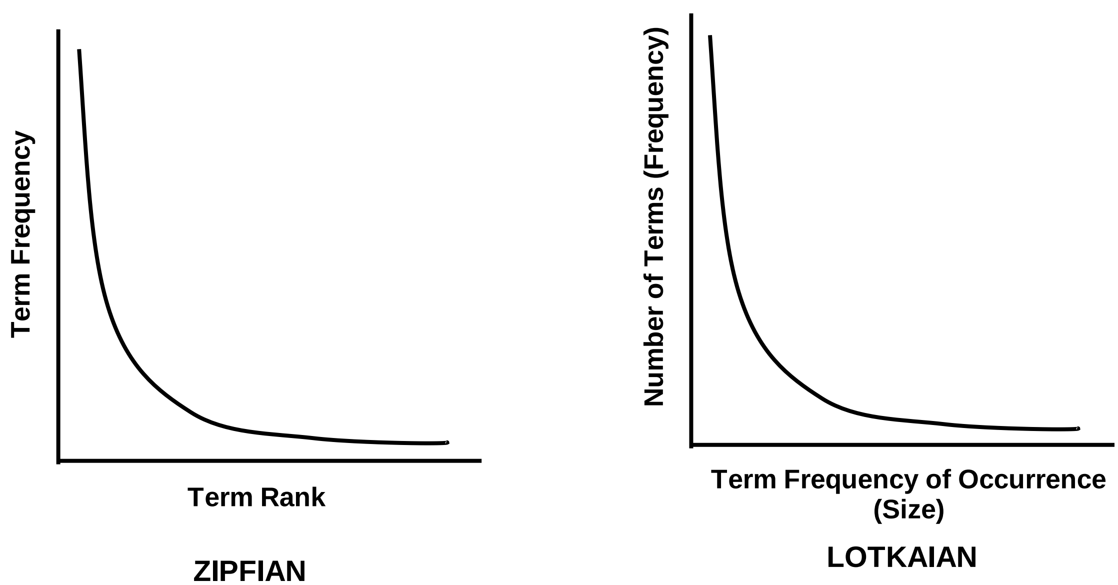

Fig. 2 Zipfian rank-frequency and Lotkaian size-frequency representation

This is also true of subject terms or descriptors that are assigned manually to documents. The frequency of a given term may be measured within a document or across a document set. These frequencies feature prominently in assessing document relevance in traditional IR environments. Term weighting methods are often associated with the statistical characteristics of terms. For example, the commonly applied tf-idf (term frequency - inverse document frequency) measure used to assign a weight to a given term is dependent on a term's frequency within a document and how frequently the same term occurs across the document collection. Those terms that occur at the low end of the size-frequency distribution of index terms represent singly occurring terms that provide the most distinctive access point to a given document, and possess greater specificity. Those terms at the tail end of this distribution, representing the most frequently occurring terms, are likely candidates for inclusion in a stopword list (i.e., terms that will not be indexed because they occur too frequently or provide little substance in describing the content of a document or provide little discriminative power in distinguishing documents from one another).

Associated with indexing specificity is the concept of exhaustivity. Indexing exhaustivity refers to the depth of indexing applied to a document. Higher exhaustivity (i.e., more terms assigned to a document) provides a larger number of access points to a given document. The size-frequency distribution of exhaustivity generally usually follows a unimodal shape, with a small proportion of documents indexed with a small number of terms, a higher proportion of documents with an intermediate number of assigned indexing terms, and a long tail of few documents with high levels of indexing. The resulting shape is similar to a skewed normal distribution (lognormal distribution for continuous data, or Poisson-like for discrete data). Wolfram (2003), chapter 5, provides a more detailed overview of indexing exhaustivity and specificity characteristics.

Knowledge of IR system content regularities can lead to a better understanding of IR processes and potential applications for system design. This can be seen in a number of research studies that have examined regularities in different aspects of IR system content with an eye on system design and evaluation, including term indexing distributions, term co-occurrence, document growth and obsolescence patterns, or Web link studies. For example, Wolfram and Zhang (2008) examined the impact of different levels of indexing exhaustivity and specificity on the characteristics of document sets generated in a vector space environment. The authors demonstrated using computer simulation how different levels of exhaustivity and specificity can influence the distribution of documents within the vector space, called a document space. Characteristics of the document space such as its density, or how closely documents are situated to one another, can in turn impact retrieval by making documents more or less similar to one another. Higher levels of specificity were found to lead to lower document space densities, while the exhaustivity levels examined were found not to play a significant a role in this regard.

The study of how terms co-occur with one another becomes feasible when more than one term is indexed per document. Techniques similar to those used in co-citation analysis may be applied in this case to identify terms that are related in some way based on their frequency of co-occurrence within documents and across document sets. The more frequently two terms co-occur with one another, the stronger the relationship between them is believed to be. Researchers have found that simple co-occurrence has limited application for identifying term or document relationships, but more sophisticated relationships may be established using language models (discussed below). One early application of this relationship was the use of co-occurrence for the automatic development of thesauri, in particular for identifying synonyms. Peat and Willett

(1991) observed, however, that the incidence of frequently co-occurring terms is small, with the frequency distribution of co-occurring term pairs also following a strong inverse relationship. Schneider and Borland (2004, 2005), recognizing the limits of simple term co-occurrence, proposed the use of a combination of document co-citation analysis, citation context analysis, and noun phrase parsing to identify important concepts. They developed a semi-automatic method for identifying primary candidate thesaurus terms within their field of focus (periodontology) with outcomes that were in high agreement with Medical Subject Heading descriptors.

Models for the growth of literature have been developed to describe how disciplines develop. The best fitting models will vary by dataset (Saxena, Gupta, and Jauhari 2007). In simple terms, growth may be: 1) linear, indicating a steady growth; 2) exponential or other non-linear rapid growth; or 3) logistic, with initial rapid growth, followed by declining additions. These models may carry over to IR environments to better estimate database and index growth rates. IR environments in which content is added, such as bibliographic databases, will continue to grow. IR environments in which the content is dynamic, where documents may be added, modified, or removed may actually show a decrease in the number of documents or space usage over time. These environments are much more difficult to model accurately given the complexity of the factors that may contribute the addition or removal of documents over time. The pattern observed for the growth of distinct index terms may follow a logarithmic shape, with rapid initial growth as new documents and terms are added followed by a slow decline in new index terms being added as the full potential vocabulary of the database is reached. The rate of growth in the number of new index terms and maximum size of the index will be influenced by the scope of the indexable vocabulary. Multilingual IR systems and those that also index content such as file names, Uniform Resource Locators (URLs), or misspelled words, will have the potential for essentially unlimited growth.

Webometric studies (Almind and Ingwersen 1997), or informetric studies of the Web, have become increasingly common over the past 20 years with a growing body of literature (Thelwall, Vaughan, and Björneborn 2005; Bar-Ilan 2008). The investigation of hypertext links between Web pages has applied ideas of social network and citation analysis to the Web environment. Methods and tools have since been developed to study Web link phenomena of interest (Thelwall, 2009). The parallels between citation analysis and Web links have been exploited and expanded, although there are some key differences. For example, once a document is published in print form, it is permanently part of the research landscape, with no changes in the list of referenced items. Web documents can be transient due to the dynamic nature of the Web. Linkages within Web documents may change, or the document itself may disappear. References in printed documents also generally point to older documents. On the Web, document age is not an issue, making concurrent linkages possible. It also makes it possible for linkage loops to exist, but this is not possible for citations where older items cannot reference newer items.

The study of persistence of documents on the Web, like growth studies, helps to inform system managers about the dynamics of websites and system developers about Web crawling frequency. Bar-Ilan and Peritz (2009) undertook a long-term study of web documents dealing with informetrics over an eight-year period. The authors observed tremendous growth in the number of documents on the topic over the time frame of the study, but also obsolescence, where documents were no longer accessible or were modified. Interestingly, the study of document persistence also has implications for scholarly communication, where the long term availability of cited documents may be questionable. Wren (2008) noted that the "decay" of URLs continues unabated for sources published in MEDLINE, based on a four-year follow up study.

Techniques used to assess relevance in the IR process have also been applied to Web links. A prime example of this is the PageRank algorithm developed by Page and Brin and employed by the search engine Google (Brin and Page 1998; Page et al. 1999). The developers of Google recognized that in citation analysis, the larger the number of citations received by an author, particularly if the citations are from influential authors, the more influential the cited author must be. The same idea can be applied in a hypertext environment, where the importance of a given document may be assessed by how many other documents that are also deemed important link to the first document. In this case, ideas from citation analysis were applied to the design of the most popular search engine in the Western World.

The integration of informetric methods has been applied to the design of scholarly information systems or digital libraries in the form of bibliometrics-aided retrieval. Mutschke Mayr, Schaer, and Sure (2011) as well as Mayr and Mutschke (2013) outlined how science models used in metrics research may help to inform IR system design and, in turn, how IR can provide a better understanding of scholarly structures. They developed a prototype system that incorporates co-word, bibliometric and network models of science in addition to standard text retrieval approaches. Search terms initially may be selected based on a cloud of visualized term associations. Next, retrieved items may be re-ranked so that journals considered core to a given area are ranked highest. Finally, ranking may also be determined through network-based assessment of author centrality using co-authorship relationships. The effectiveness of re-ranking Bradfordized search results (i.e., results grouped in Bradford zones (Bradford 1934)) is further tested in Mayr (2013) with positive results for search performance.

IR System Usage

IR system usage characteristics can be evaluated from different levels of granularity. At the most fundamental level, one can examine the terms used by searchers within or across queries, where terms are considered to be searchable units on the systems--usually groups of characters such as words. Search term frequency distributions have been studied for a number of retrieval environments. Examinations at the query level look at patterns in how searchers use terms for each submission to the retrieval system. For example, Spink, Jansen, Wolfram and Saracevic (2002) examined the frequency of terms used in queries for three time periods taken at two-year time intervals for the Excite search engine. The distribution of term frequencies follows a shape that can be described as Zipfian or Lotkaian. A term by itself may not provide a context for the query, so conclusions regarding the topicality of searches based on singular terms should not be drawn. An examination of the many singly occurring terms may be revealing of errors in user spelling, or if analyzed over time, of changes in vocabulary usage based on differences in the number of singular terms found.

Query-level analysis provides a richer set of data that permits a focus on the topicality of searches and the characteristics of the query terms taken together. One could examine the frequency of occurrence of queries as a whole within a transaction log to reveal which queries are more popularly entered. Wolfram (2000) examined the frequency of occurrence of queries from the 1997 Excite search service data set, finding that many of the most frequently occurring queries were also found in an analysis of the Alta Vista search engine. Also of interest are the characteristics related to the number of terms used per query. Unlike the power law inverse pattern, this distribution usually follows a unimodal shape, with a mode of two or three terms. Studies have found that for public search systems, the number of terms used per query is usually between two and three terms. Wolfram (2008), in a comparison of four different Web-based search environments, found a mean of 1.78 to 3.66 terms used per query across the systems studied. Also associated with some public search services is the number of pages of results viewed by searchers for a given query. A page may contain anywhere from 10 to 50 results. The number of pages viewed per query follows a power law-like pattern, where most searchers do not browse beyond the first page of results and few searchers will browse many pages (Spink et al. 2002; Wolfram 2008).

Session characteristics provide the richest set of data because they encompass the complete search requests for a user's interaction with an IR system. In systems where there is a distinct start and end based on login and logout information for a given account, these data are indicative of the interaction. With many public search systems, it is not possible to precisely identify session characteristics because identifiers such as system cookies or Internet Protocol (IP) addresses are used, making it necessary to identify methods by which session boundaries may be detected. Cookies may be tied to a given machine, but not a given searcher. Similarly, IP addresses may be shared by multiple machines resulting in the potential for queries by different searchers to be intermingled. The issue of how session boundaries are detected is beyond the scope of this work. Quantitative characteristics of sessions that have been studied include the frequency distribution of session length as measured by the number of queries per session and modifications or changes in queries over the course of a session. This behavior again follows a power law-like pattern with sessions consisting of one query being the most common duration (Spink et al. 2002). Wolfram (2008) found that the majority of sessions for three of the four Web-based search environments he investigated consisted of single queries.

Early studies of query transaction logs were largely descriptive and simply summarized the observed distributions. More recent studies have attempted to move beyond description to better understand groups of users or to predict usage behavior and outcomes. For instance, with sufficient data, investigators may identify patterns in usage behavior by relying on exploratory statistical techniques such as cluster analysis or network analysis. Using forty-seven variables collected from usage transaction logs for the MELVYL OPAC, Chen and Cooper (2001) were able to identify categories of users with a high degree of success based on usage patterns alone, and with no demographic data. By applying a similar approach, Wolfram, Wang and Zhang (2009) were able to use four to six search and usage criteria for three Web search environments to identify three distinct types of sessions common to each environment: 1) "hit and run" sessions on focused topics constituting the majority of sessions; 2) relatively brief sessions on popular topics, and; 3) sustained sessions on focused topics with greater query modification. The last type of session was possibly indicative of struggling searchers. The ability to identify struggling searchers while they are searching could make it possible for the IR system to intervene with assistance. More recently, Han, Joo and Wolfram (2014) have applied network analysis techniques to search session actions collected from an image-based digital library to better understand search and browsing behaviors. The strong bi-directional relationship found I the network between two identified actions pointed to the need for improved browsability of the system interface.

Usage characteristics are not just limited to term, query, and session data. For document access, the number of individual document downloads in a given data set or the number of page visitations in a Web-based environment have also attracted research interest. Patterns in usage can be viewed from two perspectives: 1) the resource provider perspective, where the frequency of use by given users is of interest, and; 2) the requestor perspective, where a user may be accessing different resource providers. Ajiferuke, Wolfram and Xie (2004) examined both perspectives for usage data from a bibliographic database provider. They found a strong power law-like relationship for both the distribution of site visitation frequency by identifiable IP addresses and for the distribution of resource requests, which were modeled using several mathematical functions, including power law models.

Information Retrieval for Informetrics Research

The flow of contributions between IR and informetrics is not unidirectional. Increasingly, information retrieval developments have benefitted metrics research. Informetrics research has become computationally more sophisticated with large datasets available for study. Bibliographic records, which have served as a primary source for informetrics study, lend themselves well to database and IR applications. The process of science has entered an era of "big data" where sophisticated systems are needed for data storage and curation to support research. The same is increasingly true of informetrics research. The techniques used that permit large scale IR or visualization of results also have applications in metric studies. Visualization techniques used to initially identify documents and their inter-relationships in IR environments can also be applied to authors, research groups, fields of study, or geographic areas to better understand the complex dynamics inherent in knowledge production and use. Sophisticated tools that permit the visualization of large amounts of data are making it possible to project high level relationships at the disciplinary level (Börner 2010). As an example, transaction log data from bibliographic and citation databases--in addition to providing a better understanding of user search behavior--can reveal insights into the structure of scientific activity. Bollen et al. (2009)

demonstrated how transaction log clickstream data (i.e., data of all user actions during their search sessions) from several database sources can be used to construct high-resolution journal-based network maps of science using searcher navigation data from one journal to another during search sessions.

The mutual contribution between informetrics and IR is exemplified by the use of PageRank in metrics research. PageRank, as mentioned earlier, was influenced by ideas from citation analysis and was developed and popularized by Google founders Page and Brin to inform the ranking of retrieved Web-based documents. It has since been applied in evaluative metrics research. For example, Liu et al. (2005) relied on a weighted version of PageRank, which they call AuthorRank, to study the co-authorship network of the Digital Libraries community. Similarly, Ding et al. (2009) and Ding (2011) applied weighted PageRank to study author co-citation and citation networks. Furthermore, Bollen, Rodriguez and Van de Sompel (2006), used a weighted PageRank algorithm to assess journal prestige, as opposed to the more traditional journal impact factor, which they argue represented a measure of journal popularity and not prestige. These studies demonstrate how the PageRank relevance ranking algorithm used in IR can be applied to assess authors and journals. These applications highlight an example of the circular beneficial relationship between informetrics and IR, where an idea from citation analysis was adapted to the IR environment. This adaptation from the IR environment was then repurposed for metrics research. An overview of PageRank-related methods for analyzing citation networks may be found in Waltman and Yan (2014).

An example of more classic IR concepts applied to informetric study can be found in White (2007a 2007b), who highlighted the close relationship among metrics, IR and relevance theory. White notes that the term frequency and inverse document frequency of terms, which are readily modeled quantitatively, shape users' perceptions of relevance and its determination in relation to an initial term for retrieval. As terms become less specific (i.e., occur more frequently), their relevance to the seed term also decreases. These classic IR measures were used by White to develop pennant diagrams (or scatterplots) with plot contents divided into sectors. The axes of the plots represent aspects of cognitive effects (tf) and ease of processing (idf) for the data units studied (e.g., themes, journals, authors).

As noted earlier, language models, and in particular topic models (Blei and Lafferty 2009) developed for machine learning and natural language processing, have introduced new ways to support information retrieval. A topic model develops a limited number of explicit or hidden topics based on language usage within a document corpus. Traditionally, documents in an IR system are represented by a subset of a large number of index terms, which are compared to assess document similarity. Using topic modeling, documents are represented as a mixture of topics, which are in turn represented by a mixture of index terms. The more limited number of topics thus reduces the complexity of the similarity determination. This benefit of dimensionality reduction for data comparison has parallels in informetrics research, where today's larger datasets produce higher dimensional spaces that are computationally burdensome to process.

Applications of topic models for informetric and scholarly communication research have begun to be explored in recent years. Among the earliest, Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) introduced an author-topic model that could be used to assess the similarity of authors and the extent to which authors address single or multiple topics in their published research. Mann, Mimno and McCallum (2006) proposed the use of topic analysis for the development of bibliometric impact measures that assess, for example, topical diffusion, diversity, longevity and transfer. Tang, Jin, and Zhang (2008) used topic modeling to model heterogeneous academic networks of papers, authors and publication venues. Lu and Wolfram (2012) have also applied topic model analysis to better understand relationships between authors. Just as IR systems store documents whose relationships are determined by similarities in the document content or citation patterns, the relationships among the authors of these documents themselves may be studied using the same methods. Lu and Wolfram found that a combination of vector space modeling and topic-based approaches made it possible to graphically represent relationships and similarities among authors in a new way that had previously only been used for IR purposes.

Additional techniques may be used in conjunction with topic modeling to reveal disciplinary features or to study research impact. For example, Yan et al. (2012) explored community structures in IR research by combining topic modeling and community detection in IR literature to reveal the changing landscape of IR research. Recently, Yan (2014) has also proposed the used of topic-based PageRank for scientific impact evaluation with superior results to the use of PageRank or a form of topic modeling alone.

The use of topic modeling in metrics research is still relatively new and provides a complementary way to study relationships between objects of interest such as authors or journals that does not require traditional citation data, while employing a more sophisticated comparison method than simple term co-occurrence. Some applications of topic modeling and how it may be used for assessing scholarly impact can be found in Song and Ding (2014).

Conclusion

Information retrieval and informetrics share a common lineage with common interests. The lines between the two have become blurred in research, where the focus is just as much about IR as it is about understanding scholarship and scholarly communication. This is highlighted in Bassecoulard, Lelu and Zitt (2007) and Zitt and Bassecoulard (2006) who applied IR techniques to examine field delineation in scholarship and recognized "… [t]he question of delineation is a particular case of information retrieval application." (Zitt and Bassecoulard 2006, p. 1515).

Regardless of whether the focus of research is on how information is produced, selected, represented, stored, retrieved, or used from an informetrics or information retrieval perspective, the common theme that exists between these two areas is human behavior and the information processes in which humans engage. This review has provided just a brief examination of how metric studies may help to inform information retrieval, and increasingly how IR developments can provide tools and new perspectives for informetric studies. An understanding of the empirical regularities that exist in how information is produced and used can benefit IR system designers to develop more efficient and more effective systems that serve the scholarly community and larger public. Similarly, the vast public and proprietary datasets being generated annually, which are of great interest to metrics researchers, will require new methods that overcome the difficulties associated with large-scale data. Developments in IR research can assist in this regard. The challenge for information science researchers in the 21st century can be met by capitalizing on our knowledge of these two important areas to better serve information users and to benefit society.

References:

Ajiferuke, I., Wolfram, D., & Xie, H. (2004). Modelling website visitation and resource usage characteristics by IP address data. In H. Julien & S. Thompson (Eds.), *CAIS/ACSI 2004 - Access to Information: Technologies, Skills, and Socio-Political Context.* http://www.cais-acsi.ca/proceedings/2004/ajiferuke_2004.pdf. Accessed 25 January 2014.

Almind, P., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to "Webometrics". *Journal of Documentation*, 53, 404-426.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century—A review. *Journal of Informetrics*, 2(1), 1-52.

Bar-Ilan, J., & Peritz, B. (2009). The lifespan of "informetrics" on the Web: An eight year study (1998–2006). *Scientometrics*, 79(1), 7-25.

Bassecoulard, E., Lelu, A., & Zitt, M. (2007). A modular sequence of retrieval procedures to delineate a scientific field: from vocabulary to citations and back. In E. Torres-Salinas & H. F. Moed (Eds.), *11th International Conference on Scientometrics and Informetrics (ISSI 2007)*, Madrid, Spain, 25-27 June 2007, 74–84.

Blei, D.M, & Lafferty, J.D. (2009). Topic Models. In A.N. Srivastava & M. Sahami (Eds.), *Classification, Clustering, and Applications*. (pp. 71- 94). Boca Raton, FL: Chapman & Hall/CRC.

Bollen, J., Rodriquez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69(3), 669-687.

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PLoS One*, 4(3), e4803.

Börner, K. (2010). *Atlas of science: Visualizing What We Know*. Boston: MIT Press.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, 137, 8-96.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30, 107-117.

Chen, H.-M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11), 888-904.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2), 236-245.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11), 2229-2243.

Egghe, L. (1990). The duality of informetrics systems with applications to the empirical laws. *Journal of Information Science*, 16, 17-27.

Egghe, L., & Rousseau, R. (1997). Duality in information retrieval and the hypergeometric distribution. *Journal of Documentation*, 53(5), 488–496.

Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35(3), 243-255.

Han, H.J., Joo, S., & Wolfram, D. (2014). Using transaction logs to better understand user search session patterns in an image-based digital library. *Journal of the Korean Biblia Society for Library and Information Science*, 25(1), 19-37.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6), 1462-1480.

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based and author co-citation approaches. *Journal of the American Society for Information Science and Technology*, 63(10), 1973-1986.

Mann, G.S., Mimno, D., & McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. The *ACM Joint Conference on Digital Libraries*, June 11-15, 2006, Chapel Hill, North Carolina, USA.

Mayr, P. (2013). Relevance distributions across Bradford Zones: Can Bradfordizing improve search? In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (Eds.), *14th International Society of Scientometrics and Informetrics Conference* (pp. 1493–1505). Vienna, Austria.

Mayr, P., & Mutschke, P. (2013). Bibliometric-enhanced retrieval models for big scholarly information systems. *In Big Data, 2013 IEEE International Conference on* (pp. 5-8). IEEE.

Mutschke, P., Mayr, P., Schaer, P., & Sure, Y. (2011). Science models as value-added services for scholarly information systems. *Scientometrics*, 89(1), 349-364.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf. Accessed 10 June 2014.

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5), 378-383.

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. In

W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275-281). New York: ACM Press.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In C. Meek & J. Halpern (Eds.), *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487-494). AUAI Press.

Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer*. Reading, MA: Addison-Wesley Publishing Company.

Saracevic, T. (1975). RELEVANCE: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6), 321-343.

Saxena, A., Gupta, B. M., & Jauhari, M. (2007). Exploring models for the growth of literature data. *DESIDOC Bulletin of Information Technology*, 27(3), 3-12.

Schneider, J. W., & Borlund, P. (2004). Introduction to bibliometrics for construction and maintenance of thesauri: methodical considerations. *Journal of Documentation*, 60(5), 524-549.

Schneider, J. W., & Borlund, P. (2005). A bibliometric-based semi-automatic approach to identification of candidate thesaurus terms: parsing and filtering of noun phrases from citation contexts. In F. Crestani & I. Ruthven (Eds.), *Information Context: Nature, Impact, and Role: 5th*

International Conference on Conceptions of Library and Information Sciences, CoLIS 2005 (pp. 226-237). Springer: Berlin.

Song, M., & Ding, Y. (2014). Topic modeling: Measuring scholarly impact using a topical lens. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 235-257). New York: Springer.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer Magazine*, 35(3), 107-109.

Tang, J., Jin, R., & Zhang, J. (2008, December). A topic modeling approach and its integration into the random walk framework for academic search. In F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, & X. Wu (Eds.), *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 1055-1060). IEEE.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative Web research for the social sciences*. Morgan & Claypool (electronic book).

Thelwall, M., Vaughan, L, & Björneborn, L. (2005). Webometrics. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 39, pp. 81-135). Medford, NJ: Information Today.

Waltman, L., & Yan, E. (2014). PageRank-related methods for analyzing citation networks. In

Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 83-100). New York: Springer.

White, H. D. (2007a). Combining bibliometrics, information retrieval, and relevance theory, part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology*, 55(4), 536-559.

White, H. D. (2007b). Combining bibliometrics, information retrieval, and relevance theory, part 2: Some implications for information science. *Journal of the American Society for Information Science and Technology*, 55(4), 583-605.

Wilson, C. S. (1999). Informetrics. In M. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 34, pp. 107-247). Medford, NJ: Information Today.

Wolfram, D. (2000). A query-level examination of end user searching behaviour on the Excite search engine. In H. Olson (Ed.), *Proceedings of the 28th Annual Conference of the Canadian Association for Information Science*. http://www.cais-acsi.ca/proceedings/2000/wolfram_2000.pdf. Accessed 10 June 2014.

Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, CT: Libraries Unlimited.

Wolfram, D. (2008). Search characteristics in different types of Web-based IR environments:

Are they the same? *Information Processing & Management*, 44, 1279-1292.

Wolfram, D., & Zhang, J. (2008). The influence of indexing practices and term weighting algorithms on document spaces. *Journal of the American Society for Information Science and Technology*, 59(1), 3-11.

Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology*, 60(5), 896-910.

Wormell, I. (1998). Informetrics: Exploring Databases as analytical tools. *Database*, 21(5), 25-30.

Wren, J. D. (2008). URL decay in MEDLINE—a 4-year follow-up study. *Bioinformatics*, 24(11), 1381-1385.

Xie, I. (2008). *Interactive information retrieval in digital environments*. Hershey, PA: IGI Publishing.

Yan, E. (2014). Topic-based Pagerank: Toward a topic-level scientific evaluation. *Scientometrics*, 100(2), 407-437.

Yan, E., Ding, Y., Milojević, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics*, 6(1), 140-153.

Zitt, M., & Bassecoulard, E. (2006), Delineating complex scientific fields by hybrid lexical-citation method: An application to nanoscience, *Information Processing & Management*, 42(6), 1513–1531.