

Identifying Research Trends in IS

Emergent Research Forum Papers

Abstract

The Information Systems (IS) discipline is a relatively young and rapidly evolving field of study, and its roots can be found within diverse disciplines. While these roots have been studied thoroughly and discussed at length, little is known about the emergent areas of research within the discipline. In this study, we map the IS discipline based on all publications in the top IS journals from 2005 through 2014. Our results provide a holistic view of the field and identify active and emergent areas of research.

Introduction

Numerous studies have investigated IS in the context of its foundational and reference disciplines. Other studies have examined what the core paradigms, theories or methodologies of the field are or what they should be (Agarwal and Lucas, 2005; Benbasat and Zmud, 2003; Hirschheim and Klein, 2003; Taylor et al, 2010). More recently, several studies have been conducted to provide a historic view of the field and its evolution during the last 40 to 50 years (Banker et al, 2004; Hirschheim and Klein, 2010; Hirschheim and Klein, 2012). These studies provide an invaluable resource in understanding the core of IS and its identity. This study builds upon these contributions. Rather than examining the established cores and foundations of the discipline, or focusing on the historic evolution of the field, we seek to identify the active and emergent areas of research based on the recent publications in IS.

As the underlying technology continues to evolve, our research to inform the community on how to take advantage of this evolution will invariably need to change too. This implies that there is a constant need to revisit past classifications, and it is a challenge that is inherent to the IS discipline. Our research attempts to address this issue. We devise a method that visualizes the recent evolution of the field and the associations between different research areas.

The paper is organized as follows: In the next section, we explain the limitations of the current document retrieval methods and tools. Section 3 summarizes the current scholarly document retrieval and classification methods that guided our research - co-citation analysis, latent semantic analysis, and bibliographic coupling. In section 4, we create an aggregate measure of similarity between a pair of documents. We use our measure to determine the association between all publications in the "Basket of 8" journals over the past ten years, and then represent the data graphically in a network format.

Limitations of Current Document Retrieval Systems

Current document retrieval systems (DRS), while extremely useful, are beset with several shortcomings, especially when used for retrieving scholarly documents. Document retrieval is the problem of finding a relevant document from a collection of documents based on the user's information need (Blair 1985; Croft 1987). Scholars in any field of study rely on DRS heavily to find scientific documents especially when conducting literature reviews. The common way of searching for a scientific document on current platforms is by searching for a specific term in the documents, while limiting the search domain or consequently refining the search results. This retrieval method is appropriate for initiating a search process or when the exact term included in the document is known. However, term mismatch caused by synonymy and polysemy represents a fundamental challenge to this model. Therefore, researchers are not able to find all relevant articles using search terms, unless they are familiar with all the synonym terms. On the other hand, a single term may refer to two completely different concepts.

In addition to being inefficient in finding relevant documents, current document retrieval and processing tools are unable to provide a holistic view of the documents' domain, or the field or subfield of study they are representing, since they are not designed for such a task. The above mentioned document retrieval methods are designed to report a list of documents based on the user's query which represents the user's information needs. Providing a holistic view of a field of study has traditionally been the domain of senior scholars who, thanks to their vast experience, have manually or empirically classified the literature.

Representing a holistic view of (hereafter called mapping) a scientific discipline manually is only possible when the field is in its initial stages of development or based on a relatively small subsample of its publications. For example, Ives (1980) manually examined 331 Management Information Systems (MIS) doctoral dissertation written between 1973 and 1979 in order to categorize research areas in this what was then new field of study. Similarly, Swanson and Ramiller (1993) manually classified 397 submissions to the journal of Information Systems Research based on their research questions to reveal the themes in the field of IS.

In this study we propose a hybrid document retrieval and classification method which can not only identify documents that are relevant to a given document efficiently, but also present a holistic map of a scientific field or subfield of study. We define an aggregated measure of documents similarity which incorporates the outputs from latent semantic analysis (LSA) as well as citation analysis. Applying this measure to all pairs of documents within a field or subfield of study allows us to develop a new document retrieval system. In addition, presenting and analyzing the results in a network format enables us to visually map the field of study.

Scholarly Document Retrieval and Classification Methods

We first review the extant scholarly DRS platforms. We group them based on their scopes and purposes, and then discuss their use, advantages, and shortcomings. We then review the existing document retrieval methods in the context of document clustering. We explore the underlying idea behind each method and show their application in mapping scientific fields or subfields of study by presenting some examples.

Scholarly Document Retrieval Systems

Currently, available document retrieval tools used by scholars are mainly in one of the two following forms: 1) search functions provided by academic publications or societies such as the ones provided by MISQ, INFORMS, ACM, or Wiley, and 2) scholarly citation indexing and search engine such as Google Scholar, ISI Web of Science, ScienceDirect, Scopus and PubMed.

The scope of the search functions provided by academic publications varies significantly and is often specific to the publication or its affiliations. For example, the search feature within the MISQ website is a simple DRS covering only MISQ publications, while the ones within ACM Digital Library and the Wiley Online Library cover tens of thousands of documents within their associated journals or conferences. However, the functions are mostly similar and are based on searching for a term (or terms) while limiting the scope of search using filter options. The search is typically performed based on title, authors, abstract, or the body (text) of the document, and the filters are often based on the name of the journal, the year of publication, the journal volume, or the journal issue. The results are presented as a list typically ranked by the "relevance" to the search term. The relevance is a measure defined and used by the individual retrieval system and is based on where in each document the search term appears or how frequently it appears in each document. Although helpful, these systems are very limited in their use. For example, term mismatch caused by synonymy (multiple words having the same meaning) and polysemy (the same word having several meanings) pose a fundamental challenge to these systems.

While DRS provided by academic publications are limited in their scope, scholarly citation indexing and search engines are often broad and span over multiple disciplines. For example, it is estimated that Google Scholar and Web of Science cover 160 and 57 million documents respectively. These systems are primarily focused on citation indexing and analysis. The results of queries are often sorted by the number of citations for a document, and a list of citing documents are also provided for each returned document. This functionality has enabled scholars to retrieve more recent documents similar to a given document,

based on direct citation. However, the effectiveness of this method is low since not all related articles to a given article will necessarily cite that article nor are all citing articles necessarily related to the article.

In the following section, we describe major document retrieval methods used for the purpose of document clustering and mapping scientific disciplines.

Constructing an Aggregated Measure of Similarity

Scholarly document clustering methods appear mostly in one of three forms: 1) latent semantic analysis (LSA), 2) bibliographic analysis, and 3) co-citation analysis (CoCit). The underlying idea behind these methods are explained in the appendix.

LSA identifies similar documents which contain words close in meaning with a high level of precision. However, two types of error may occur as with any automated classification method. First, LSA may misclassify two documents to be dissimilar when they are actually quite similar. Second, it may identify two documents as being similar when they are not. The first type of error is reduced by aggregating the similarity measures derived from LSA and CoCit. For the set of IS publications described in the data source section, we use a max-aggregator where the CoCit is higher than 0.6. This value balances the hit rate and precision of the measure. The second type of error is reduced by filtering out uncoupled documents from the set of identified similar documents. The step-by-step procedure to derive the final measure of similarity is presented in the following algorithm.

Algorithm

- Steps 1 and 2: Perform LSA to find each document representation in the term-document space and determine the similarity for each pair of documents based on the cosine of the angles between their vectors in the space.
- Steps 3 and 4: Represent each document by the set of its citations (outgoing) and use co-citation analysis to determine the similarity between each pair of documents.
- Steps 5 and 6: Represent each document by the set of its references (incoming) and identify bibliographically “uncoupled” pairs.
- Step 7: Create set S containing all pairs of similar documents by aggregating the two measures of similarities found in steps 2 and 4.
- Step 8: Remove all pairs from set S which are bibliographically “uncoupled” as determined by step 6.

These steps are summarized in Figure 1.

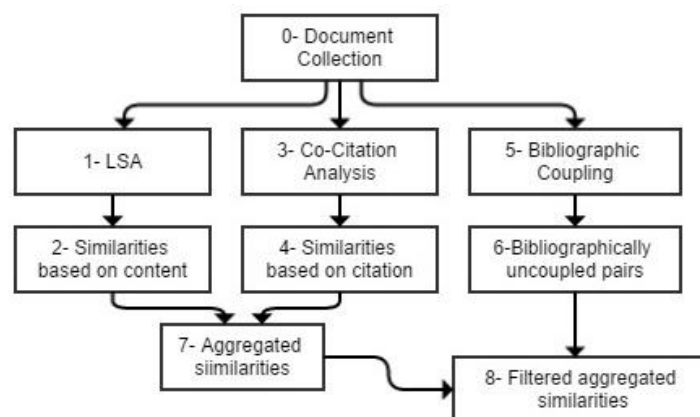


Figure 1. Algorithm for constructing an aggregated measure of similarity

5. Network Representation and Community Detection

Date Source

We analyzed the contents and references of all articles published in the top eight IS journals included in the “Senior Scholars’ Basket of Journals” from 2005 through the present (2014). Excluding editorial notes and introduction to special issues, our collection contains 2,423 articles. In addition, we collected and analyzed information on all papers citing (31,530 papers) and cited by (165,282 papers) these 2,423 articles.

Having one aggregated measure of similarity between each pair of documents allow us to represent the data in a network format. We then employ a robust algorithm for community detection in large networks, based on Blondel et al. (2008). Incorporating our proposed aggregated similarity measure in these algorithms identified the communities for 97.3% of the documents in our data set of IS publications. For the remaining 2.7% of the documents, we assigned them to the community with the highest number of “similar” documents.

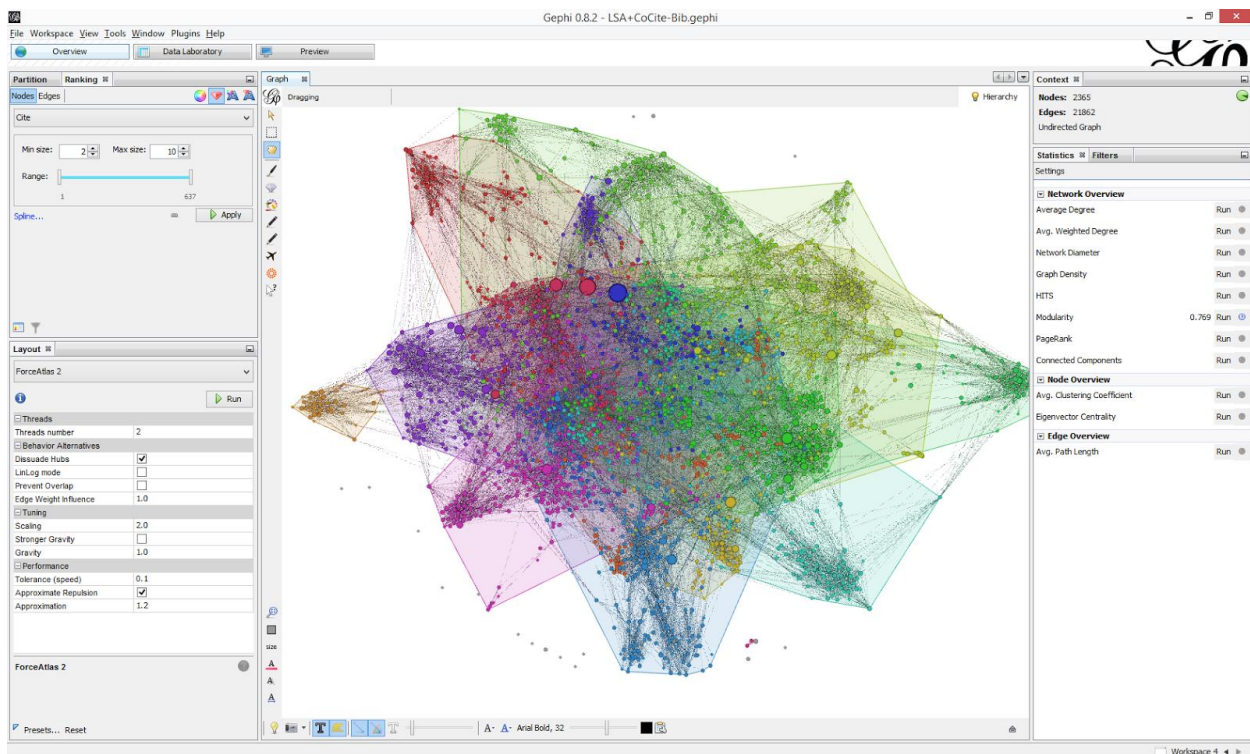


Figure 2. Community detection based on the proposed aggregated similarity measure

Active Areas of Research in IS

As illustrated in Figure 2, the clustering results contain several clearly separated clusters as well as some interrelated ones. In either case, we were able to identify and label each cluster relatively easily (based on the title of the similar articles). Identified communities are classified into five categories as shown in Figure 3.

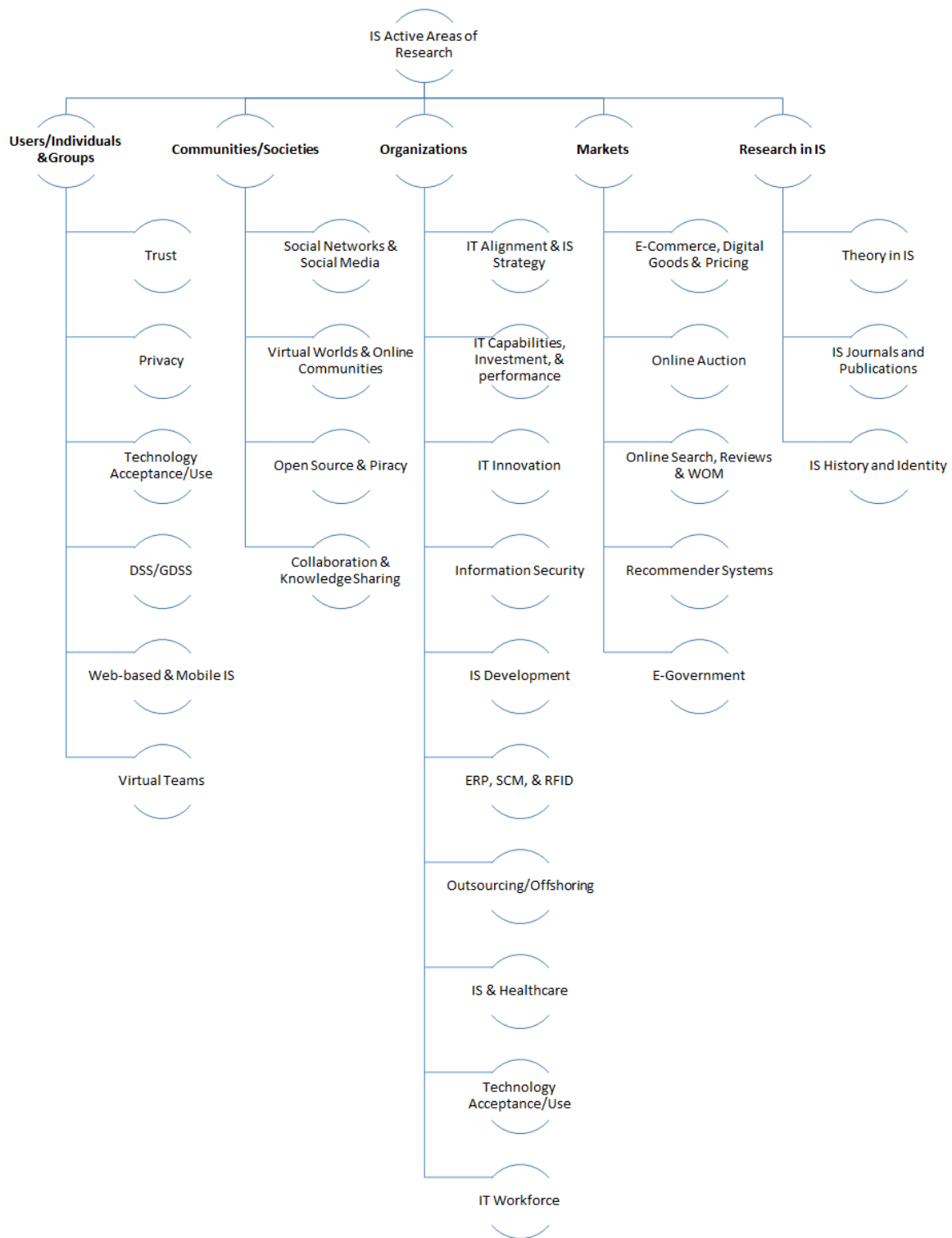


Figure 3. IS active areas of research

Classification Verification

We have compared our classification results with the pioneering work of Sidorova et al. (2008), Sidorova et al. (2013) and a set of recent review articles in IS literature that may indicate the key topics in IS (Hirschheim and Heinz, 2010; Hirschheim and Heinz, 2012; Serenko and Jiao, 2012; Banker and Kauffman 2004)

From the 28 categories identified in this study (shown in Figure 7), 26 are explicitly included in Sidorova et.al. (2013). The two categories which are treated differently in our classification are titled “Social Networks & Social Media” and “Open Source.” Social Networks and Social Media in our classification does not directly correspond to any of the categories in Sidorova (2013). The category containing articles on social networks in Sidorova’s classification is labeled as Networks and contains topics such as Artificial Neural Networks and telecommunication and electronic networks. This can be a result of LSA’s tendency to group articles with the keyword of “network” together, while they actually refer to different concepts. In comparison, our classification suggests a strong association of articles on social networks and social media based on text as well as citation patterns.

Another subtle but important difference between our proposed classification scheme and the previous ones in on the topic of Open Source. While Open Source is classified under Software Development and Use in previous classifications, we have classified it under the Communities/Societies class. This is because of the strong association between many articles on Open Source and other categories under the Communities/Societies class, such as Online Communities and Collaboration. We subsequently observed that social and collaborative aspects of open source movement has been a focus of study recently, in the IS literature, which indicated that the proposed classification scheme was “alert” enough to catch up on this development.

Discussion and Future Work

Our contribution to the literature is two-fold. First, we contribute to the knowledge base in information science and bibliometrics by proposing a scholarly document retrieval and classification system. In this document retrieval and classification system, we suggest a novel approach to aggregate measures of document similarities that is rooted within a Bayesian probabilistic framework; and we design a network representation of a collection of documents within a discipline.

Third, we contribute to the research area of IS History. The current study complements Sidorova et al. (2008) and Sidorova et al. (2013). A collection of 2,423 publications in top IS journals is analyzed using the developed system to map recent IS literature; the properties of the map is presented and important articles are identified; and a classification of the recent IS literature is presented and compared to some of the previous classifications.

We are currently working to first statistically evaluate our classification performance. Then, we seek to identify the most important recent articles in IS, using the developed network of IS articles.

Appendix

Scholarly Document Clustering and Classification

Scholarly document clustering methods appear mostly in one of three forms: 1) semantic analysis, 2) bibliographic analysis, and 3) co-citation analysis. We describe the underlying ideas behind these three methods and their advantages and disadvantages. We then explain the concept of hybrid modeling which is the foundation of our aggregation method.

a. Latent Semantic Analysis

Latent semantic analysis (LSA) is “a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text” (Landauer et al. 1998). The underlying idea is that in a collection of documents, similar documents contain words that are close in meaning. In addition, similarity of meaning of words and documents to each other can be determined using simultaneous equations (Evangelopoulo et al. 2012). The analysis is based on factorization of word-counts-per-document matrix into a lower space, using singular value decomposition (SVD). Association between two documents or two terms may then be calculated by taking the cosine of the angle between their representation vectors in the lower space. It has been shown that LSA reduces the adverse effects of synonymy and polysemy in automated document clustering (Evangelopoulo et al. 2012).

b. Co-Citation

Document co-citation is another measure of semantic association between documents. The definition of co-citation between two documents A and B is based on the number of documents citing both A and B. Therefore, it is a measure of association between two documents established by the citing authors (Small, 1973). The co-citation approach is founded on the premise that peer-reviewed publications provide the most valid representation of a school of research (Gmur 2003).

Several variations of co-citation have been developed since its introduction. The CoCit score is a measure of co-citations which has demonstrated a considerably higher degree of robustness than other measures of co-citation when used for classification (Gmur 2003). The CoCit score between document A and document B is defined as the co-citation count divided by the product of minimum counts and mean counts of the two individual documents' citations.

$$\text{CoCit}_{AB} = \frac{(\text{co-citation}_{AB})^2}{\text{minimum}(\text{citation}_A; \text{citation}_B) \times \text{mean}(\text{citation}_A; \text{citation}_B)}$$

Although co-citation analysis is the dominant method for empirical study of scientific disciplines, its effectiveness for topic discovery and clustering has been questioned due to its low recall rate. The problem is more severe when using co-citation for analysis including recent publications that contain a significant number of documents with low or no citations.

c. Bibliographic Coupling

Documents are bibliographically coupled if they cite at least one other document in common. The bibliographic coupling approach for document classification is founded on the premise that “a number of scientific papers bear a meaningful relation to each other when they have references in common” (Weinberg 1974). Bibliographic coupling and co-citation analysis investigate two completely different sources of information to find evidence of two documents' association. As visually displayed in Figure 4, bibliographic coupling is a backward-looking approach based on cited documents while co-citation analysis is forward-looking and based on citing documents.

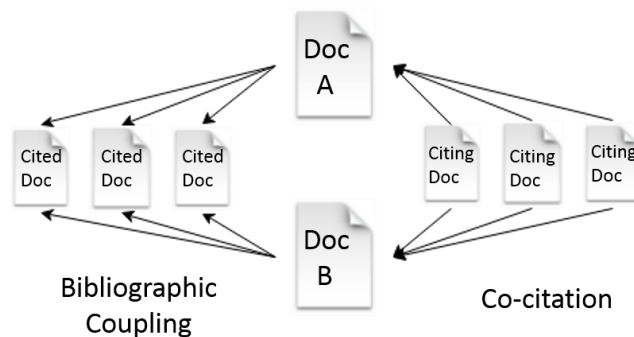


Figure 4. Co-citation and Bibliographic Coupling between two documents

Bibliographic coupling has been used for clustering of scientific documents for decades. However, the method's precision (the fraction of similar documents that are correctly identified as similar), when used by itself, is of concern since two papers having a reference in common may refer to different pieces of information in the cited document (Weinberg 1974). Bibliographic coupling has a low precision but a high recall (the fraction of dissimilar documents that are correctly identified as dissimilar). In Section 4, we explain how we take advantage of this high recall rate by using it as a filter in our proposed system.

References

- Agarwal, Ritu, and Henry C. Lucas Jr. "The Information Systems Identity Crisis: Focusing on High-Visibility and High-Impact Research." *MIS Quarterly*, 2005, 381–98.
- Banker, Rajiv D., and Robert J. Kauffman. "50th Anniversary Article: The Evolution of Research on Information Systems: A Fiftieth-Year Survey of the Literature in Management Science." *Management Science* 50, no. 3 (2004): 281–98.
- Banville, Claude, and Maurice Landry. "Can the Field of MIS Be Disciplined?" *Communications of the ACM* 32, no. 1 (1989): 48–60.
- Baskerville, Richard L., and Michael D. Myers. "Information Systems as a Reference Discipline." *MIS Quarterly*, 2002, 1–14.
- Benbasat, Izak, and Ron Weber. "Research Commentary: Rethinking 'diversity' in Information Systems Research." *Information Systems Research* 7, no. 4 (1996): 389–99.
- Benbasat, Izak, and Robert W. Zmud. "The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties." *MIS Quarterly*, 2003, 183–94.
- Blair, David C., and Melvin E. Maron. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System." *Communications of the ACM* 28, no. 3 (1985): 289–99.
- Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008, no. 10 (2008): P10008.
- Boyack, Kevin W., and Richard Klavans. "Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?" *Journal of the American Society for Information Science and Technology* 61, no. 12 (2010): 2389–2404.
- Boyack, Kevin W., Henry Small, and Richard Klavans. "Improving the Accuracy of Co-Citation Clustering Using Full Text." *Journal of the American Society for Information Science and Technology* 64, no. 9 (2013): 1759–67.
- Burton Swanson, E., and Neil C. Ramiller. "Information Systems Research Thematics: Submissions to a New Journal, 1987–1992." *Information Systems Research* 4, no. 4 (1993): 299–330.
- Chen, WenShin, and Rudy Hirschheim. "A Paradigmatic and Methodological Examination of Information Systems Research from 1991 to 2001." *Information Systems Journal* 14, no. 3 (2004): 197–235.

- Cheon, Myun J., C. Lee, Varun Grover, and others. "Research in MIS—points of Work and Reference: A Replication and Extension of the Culnan and Swanson Study." *ACM SIGMIS Database* 23, no. 2 (1992): 21–29.
- Croft, W. Bruce. "Combining Approaches to Information Retrieval." In *Advances in Information Retrieval*, 1–36. Springer, 2000.
- Croft, W. Bruce, and Roger H. Thompson. "I3R: A New Approach to the Design of Document Retrieval Systems." *Journal of the American Society for Information Science* 38, no. 6 (1987): 389–404.
- Culnan, Mary J. "Mapping the Intellectual Structure of MIS, 1980-1985: A Co-Citation Analysis." *MIS Quarterly*, 1987, 341–53.
- . "The Intellectual Development of Management Information Systems, 1972–1982: A Co-Citation Analysis." *Management Science* 32, no. 2 (1986): 156–72.
- Culnan, Mary J., and E. Burton Swanson. "Research in Management Information Systems, 1980-1984: Points of Work and Reference." *MIS Quarterly*, 1986, 289–302.
- Evangelopoulos, Nicholas, Xiaoni Zhang, and Victor R. Prybutok. "Latent Semantic Analysis: Five Methodological Recommendations." *European Journal of Information Systems* 21, no. 1 (2012): 70–86.
- Gmür, Markus. "Co-Citation Analysis and the Search for Invisible Colleges: A Methodological Evaluation." *Scientometrics* 57, no. 1 (2003): 27–57.
- Grover, Varun, Rahul Gokhale, Jaejoo Lim, John Coffey, and Ramakrishna Ayyagari. "A Citation Analysis of the Evolution and State of Information Systems within a Constellation of Reference Disciplines." *Journal of the Association for Information Systems* 7, no. 5 (2006): 13.
- Hirschheim, Rudy, and K. Heinz. "A Short and Glorious History of the Information Systems Field." *Journal of the Association of Information Systems*, 2010.
- Hirschheim, Rudy, and Heinz K. Klein. "A Glorious and Not-so-Short History of the Information Systems Field." *Journal of the Association for Information Systems* 13, no. 4 (2012): 188–235.
- Jacomy, Mathieu, Sebastien Heymann, Tommaso Venturini, and Mathieu Bastian. "Forceatlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization." *Medialab Center of Research* 560 (2011).
- Klein, Heinz K. "Crisis in the IS Field? A Critical Reflection on the State of the Discipline." *Journal of the Association for Information Systems* 4, no. 1 (2003): 10.
- Lambiotte, Renaud, J.-C. Delvenne, and Mauricio Barahona. "Laplacian Dynamics and Multiscale Modular Structure in Networks." *arXiv Preprint arXiv: 0812.1770*, 2008.
- Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25, no. 2–3 (1998): 259–84.
- Li, Yibai, and K. D. Joshi. "The State of Social Computing Research: A Literature Review and Synthesis Using the Latent Semantic Analysis Approach," 2012.
- Lyytinen, Kalle, and John Leslie King. "Nothing at the Center?: Academic Legitimacy in the Information Systems Field." *Journal of the Association for Information Systems* 5, no. 6 (2004): 8.
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank Citation Ranking: Bringing Order to the Web," 1999.
- Palvia, Prashant, David Leary, En Mao, Vishal Midha, Parveen Pinjani, and A. F. Salam. "Research Methodologies in MIS: An Update." *Communications of the Association for Information Systems* 14, no. 1 (2004): 24.
- Pritchard, Alan. "Statistical Bibliography or Bibliometrics?" *Journal of Documentation*, no. 25 (1969): 348–49.
- Robey, Daniel. "Research Commentary: Diversity in Information Systems Research: Threat, Promise, and Responsibility." *Information Systems Research* 7, no. 4 (1996): 400–408.
- Serenko, Alexander, and Changquan Jiao. "Investigating Information Systems Research in Canada." *Canadian Journal of Administrative Sciences/Revue Canadienne Des Sciences de l'Administration* 29, no. 1 (2012): 3–24.
- Sidorova, Anna, Nicholas Evangelopoulos, Russell Torres, and Vess Johnson. *A Survey of Core Research in Information Systems*. Springer, 2013.
- Sidorova, Anna, Nicholas Evangelopoulos, Joseph S. Valacich, and Thiagarajan Ramakrishnan. "Uncovering the Intellectual Core of the Information Systems Discipline." *MIS Quarterly*, 2008, 467–82.
- Small, Henry. "Co-Citation in the Scientific Literature: A New Measure of the Relationship between Two Documents." *Journal of the American Society for Information Science* 24, no. 4 (1973): 265–69.

- Straub, Detmar. "The Value of Scientometric Studies: An Introduction to a Debate on IS as a Reference Discipline." *Journal of the Association for Information Systems* 7, no. 5 (2006): 241–46.
- Taylor, Hazel, Stuart Dillon, and Melinda Van Wingen. "Focus and Diversity in Information Systems Research: Meeting the Dual Demands of a Healthy Applied Discipline." *MIS Quarterly* 34, no. 4 (2010): 647–67.
- Vessey, Iris, Venkataraman Ramesh, and Robert L. Glass. "Research in Information Systems: An Empirical Study of Diversity in the Discipline and Its Journals." *Journal of Management Information Systems* 19, no. 2 (2002): 129–74.
- Wei, Chih-Ping, Paul Jen-Hwa Hu, Chia-Hung Tai, Chun-Neng Huang, and Chin-Sheng Yang. "Managing Word Mismatch Problems in Information Retrieval: A Topic-Based Query Expansion Approach." *Journal of Management Information Systems* 24, no. 3 (2007): 269–95.
- Weinberg, Bella Hass. "Bibliographic Coupling: A Review." *Information Storage and Retrieval* 10, no. 5 (1974): 189–96.
- Weiss, Ron, Bienvenido Vélez, and Mark A. Sheldon. "HyPursuit: A Hierarchical Network Search Engine That Exploits Content-Link Hypertext Clustering." In *Proceedings of the Seventh ACM Conference on Hypertext*, 180–93. ACM, 1996.