View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by AIS Electronic Library (AISeL)

ISTopic: Understanding Information Systems Research through Topic Models

Research-in-Progress

Hailiang Chen
Department of Information Systems
College of Business
City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong, Hong Kong
hailchen@cityu.edu.hk

J. Leon Zhao
Department of Information Systems
College of Business
City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong, Hong Kong
hailchen@cityu.edu.hk

Abstract

What are the fundamental research questions in Information Systems? How do various research topics relate with one another to form the IS research landscape and how do they evolve over time? This study is an initial attempt to answer these questions using topic models to investigate the topics examined by the premier IS journals in 1977-2014. We present an IS Topic Graph that contains 33 research areas, 31 of which are closely connected with one another. Further analyses of this graph reveal how different IS research areas are intertwined to the extent that they are almost inseparable. Looking into IS research at a finer level, we identify 300 research topics, and a chronological analysis reveals a trend of topic diversification and externalization. To guide future research, an intelligent literature search tool called ISTopic is built and is available at http://www.istopic.org for public access.

Keywords: ISTopic, topic model, IS discipline, correlated topic model, latent dirichlet allocation

Introduction

Prior studies in information systems have undergone different paths to assess the intellectual structure of the information systems discipline. Besides a few studies that involve manual coding of research articles (e.g., Alavi and Carlson 1992; Vessey et al. 2002), a majority of existing studies adopt the dimensionality reduction approach such as co-citation analysis (e.g., Culnan 1986; Culnan 1987; Taylor et al. 2010) and latent semantic analysis (e.g., Sidorova et al. 2008), so that a large number of research articles in the discipline can be described by using a much lower number of unobserved (latent) variables called factors (i.e., these methods are different forms of factor analysis). Researchers then manually interpret and name these factors by summarizing the research by foundational authors (co-citation analysis) or the terms and documents (latent semantic analysis) loaded on the same factors.

One major limitation with the dimensionality reduction approach as discussed in these prior studies is that not all research areas or themes are uncovered. For example, Taylor et al. (2010) point out that "up to 2005, design science has not emerged as an independent factor, or even at the fine-grained level, as a research theme" (p. 665), as design science researchers are absent in the foundational author list. Similarly, even though all the research abstracts published in the top 3 IS journals from 1985 to 2006 are included in the latent semantic analysis conducted by Sidorova et al. (2008), "for the 100-factor solution, 25.6 percent of the documents failed to load on any of the 100 factors" (p. A2). This suggests that a substantial amount of research work is left out and cannot be characterized by the factors discovered.

A second major limitation with prior studies on mapping out the intellectual structure of the IS discipline is that relationships among different research areas or subfields are not explicitly modeled, so up until this date there is no systematic study of how different research areas interact and glue together to form the information systems research landscape. Thirdly, existing approaches typically assign only one topic to a document, which is quite restrictive, because it is common that multiple topics could be covered together in one research article (e.g., an article about healthcare security covers both the healthcare and security topics). To address these limitations and paint a complete picture of the intellectual structure of the information systems field, this study employs the Correlated Topic Model (CTM) proposed by Blei and Lafferty (2007) to discover the main research areas (topics) examined in premier IS journals and at the same time describe their relationships. CTM is an extension of the generative probabilistic model[1], Latent Dirichlet Allocation (LDA), developed by Blei et al. (2003). The intuition behind LDA is that documents are generated by choosing a distribution over a set of latent topics and that each topic is in turn characterized by a distribution over words. Topic models also assume that each document can contain multiple topics. In addition, once the model is trained, all documents (whether new or in the training dataset) can be described by a distribution over a set of topics, where each topic is represented by a list of most likely words. In this way, all research topics and their associated research papers in the information systems discipline can be identified. However, the LDA model implicitly assumes independence among topics (by the assumption of the Dirichlet distribution on the topic proportions) and fails to model the correlation between topics. CTM addresses this limitation and replaces the Dirichlet with a more flexible logistic normal distribution, which allows the presence of one latent topic to be correlated with the presence of another. The estimation results from the CTM model can be further utilized to draw a topic graph to identify the connections among topics.

In this paper, we report our research progress by employing topic modeling techniques to study an academic discipline and offer deeper insights into the discipline both at a topic relationship level and at the chronology level. Our data include 2,789 abstracts of all research studies published from 1977 to 2014 in the three well regarded top IS journals: *Information Systems Research* (*ISR*), *Journal of Management Information Systems* (*JMIS*), and *MIS Quarterly* (*MISQ*). We find that using a total of 40 topics achieves the best fit for CTM according to the model selection results, although 7 of them are mostly about nontopical content. Among the remaining 33 topics, 31 topics are connected with one another, and only 2 topics are isolated topics. Although LDA's independence assumption about latent topics may be

[1] Generative probabilistic models assume that observed data is generated by some parameterized random process; hidden parameters are estimated with the goal of providing the best fitting of the observed data.

unrealistic in many text corpuses, it produces a collection of more research topics (100 or 300 topics) at a finer level in our context.

Our first main contribution to the IS community is to present the IS Topic Graph that reveals the connections among different research areas in IS. This graph provides useful insights into each research area based on the other areas that are connected to it. In addition, replacing each topic area with its most productive researchers, we characterize how top researchers under each research area connect with one another based on proximity in research focuses.

Our second main contribution is to shed light on the evolution of IS research topics examined in premier journals. We find that IS research is more geared towards internal corporate operations in the earlier years, but gradually IS topics have become more diversified and moved to external factors such as markets, customers and relationships. Our results also indicate that over the years IS research has shifted from data room concerns to boardroom concerns.

Our last main contribution is to build an intelligent literature search tool called ISTopic for the IS community to utilize the results of our research. ISTopic is available online for public access at http://www.istopic.org. Following Baskerville and Myers (2009), the trend of a research topic can be visualized through its associated number of abstracts over the years to show the "fashion waves" in IS research. On an article level, a graphic account of the topic distribution within each abstract can reveal how different research topics are covered by one research work. On an author level, the research interests of an author in our sample can be identified by aggregating the topic distributions over all the abstracts written by that author.

Data

To get a complete picture of all research topics covered by the three top "pure MIS" journals (ISR, JMIS, and MISQ) suggested by Rainer and Miller (2005), we set the beginning of our sample period in 1977, when MIS Quarterly was established. The data was collected through the official website of each journal. Title, author names, keywords, and abstracts of each publication were downloaded. A total of 2,789 research articles (including research notes) in a 38-year period are included in the analyses. Table 1 provides a brief summary of our dataset. As the title and author-supplied keywords of an article are indicative of its topic focus, we put the title and keywords together with the abstract as the text input for topic models. Following standard practice, we filter out stop words (such as "and", "an", and "are") and also drop rare words that appear in less than four abstracts. Stemming in topic models is often unnecessary, as it sometimes combines terms that may have different meanings. On the other hand, different forms or variations of the same word will usually show up in the word list for the same topic. For this reason, we only present the results without the stemming procedure. Results with stemming are available upon request.

Journal	#Years	Starting Year	#Abstracts
Information Systems Research	25	1990	698
Journal of Management Information Systems	31	1984	1,043
MIS Quarterly	38	1977	1,048
Total	38	1977	2,789

Table 1. Summary Statistics

IS Topic Graph

Assuming that two topics could potentially be correlated in one abstract, we first apply the Correlated Topic Model (CTM) to identify latent topics in the information systems discipline. We use the variational expectation-maximization (EM) method developed by Blei and Lafferty (2007) for training and inference. To evaluate the number of topics that have been studied over the 38-year period, we fit the CTM model with different number of topics ranging from 10 to 300 and compute the held out log probability using 10-fold cross validation. At 40 topics, the held out log probability reaches the peak, indicating that using 40

topics achieves the best model fit. Among the 40 topics identified by CTM, 7 of them are however related with nontopical contents. The list of the meaningful 33 topics identified by CTM is available on the ISTopic site.

Following Blei and Lafferty (2007), we construct the IS Topic Graph by adopting the technique introduced in Meinshausen and Bühlmann (2006). This technique is adapted from the Lasso (Tibshirani 1996) and uses the topic weights of the 40 CTM topics in 2,789 articles as the input. The idea is to regress each topic's weight on all the other topics' weights, while imposing a penalty parameter on the regression coefficients to encourage sparsity of the topic graph. The output is a 40 by 40 matrix indicating whether there is an edge between any two topics or not. After some trial and error, we set the penalty parameter at 0.04 so that the graph is still sparse but the number of "isolated" nodes is minimized. Figure 1 presents this topic graph. Each topic is represented by a circle (node) in the graph. Topic ID and the top 5 most likely words are listed within the circle. The number of abstracts associated with each topic is shown in parenthesis. When two topics are correlated, there is a solid line (edge) between them. Among the 33 topics about meaningful contents, we find that 31 topics are connected with one another and the remaining 2 topics (Topic 07 and Topic 38 at the bottom of the graph) are "isolated" from the others, indicating that the presence of these two topics is independent from the presence of other topics. The IS Topic Graph not only covers the main research topics examined by the IS field but also describes the relationships among them. To the best of our knowledge, no prior study has ever explicitly analyzed the connections among all IS research topics; our study is the first one that reveals the network structure of research topics in the information systems discipline. As an ongoing research effort, we intend to refine the IS Topic Graph with intuitive topic representations by building on the MIS research framework proposed by Berthon et al. (2002) that consists of Problem, Theory, Method, and Context.

One immediate benefit of the IS Topic Graph is that we can learn important aspects of a topic based on the other topics that are connected to it in the graph. Due to space limit, we will only describe two examples here. First, Topic 27 is about the coordination and collaboration issues for virtual teams. This topic is connected with Topic 17 (group communication and group work) and Topic 35 (software/system development project) in the graph, which means that these two topics are closer to Topic 27 than all the remaining topics. This makes sense because virtual teams are a special type of groups and open source software development projects are often carried out by virtual teams. Second, Topic 05 is about IT outsourcing, governance, and contracts. It is connected with Topic 13 (feedback and reputation mechanisms in the vendor/client relationship), Topic 14 (information risk and competitive advantage), and Topic 28 (firm performance and software industry) in the graph. These connections provide us with various insights into the relevant but different research topics in IS research surrounding the outsourcing topic.

To infer a big picture about IS research from the IS Topic Graph in Figure 1, we can see that topics can be clustered into strategic issues, behavioral issues, economic issues, and design issues. However, it is difficult to cleanly separate a particular type of issues from other issues. Although some IS activities turn to cluster researchers into camps based on research paradigms, the IS Topic Graph shows that IS research topics are very intertwined, indicating that researchers from so called different camps can work together to resolve research problems by collaborating with people in their neighborhoods of research.

As an initial attempt to understand the connections between topics in the IS Topic Graph, in Figure 2 we replace each topic node with the last names of its five most productive researchers and illustrate how top researchers under each research area connect with one another based on proximity in research focuses. For each of the 33 topics, we identify the top 15 researchers who have published most in the corresponding topic. 364 distinct researchers are found, and 75 of them show up in more than one topic's list of top 15 researchers. Interestingly, we find that many connected two topics in the IS Topic Graph do not share any top researcher in common. This at least implies that productive researchers publishing on two different topics is not the main reason why two topics are connected in the IS Topic Graph.

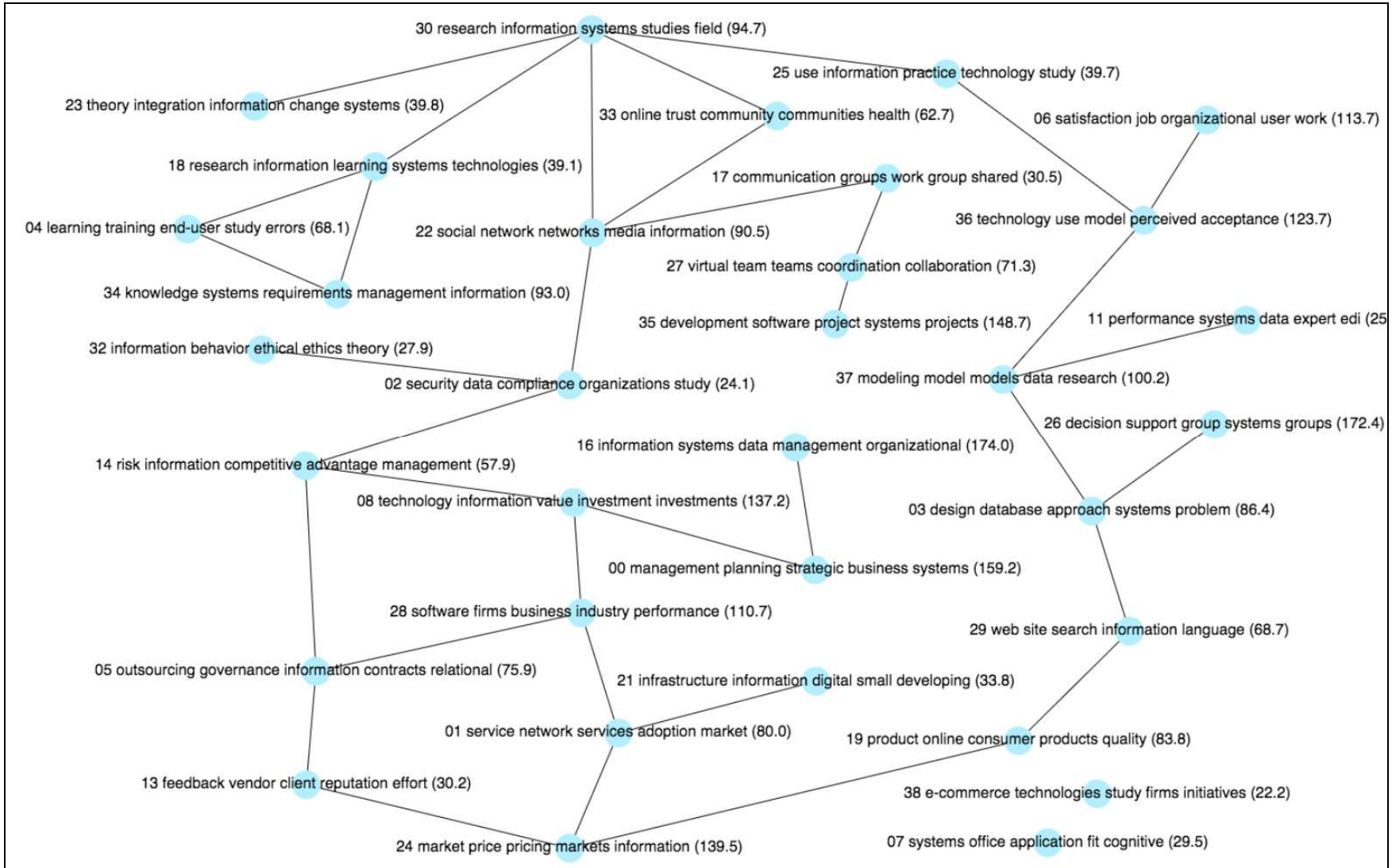

Figure 1. The IS Topic Graph

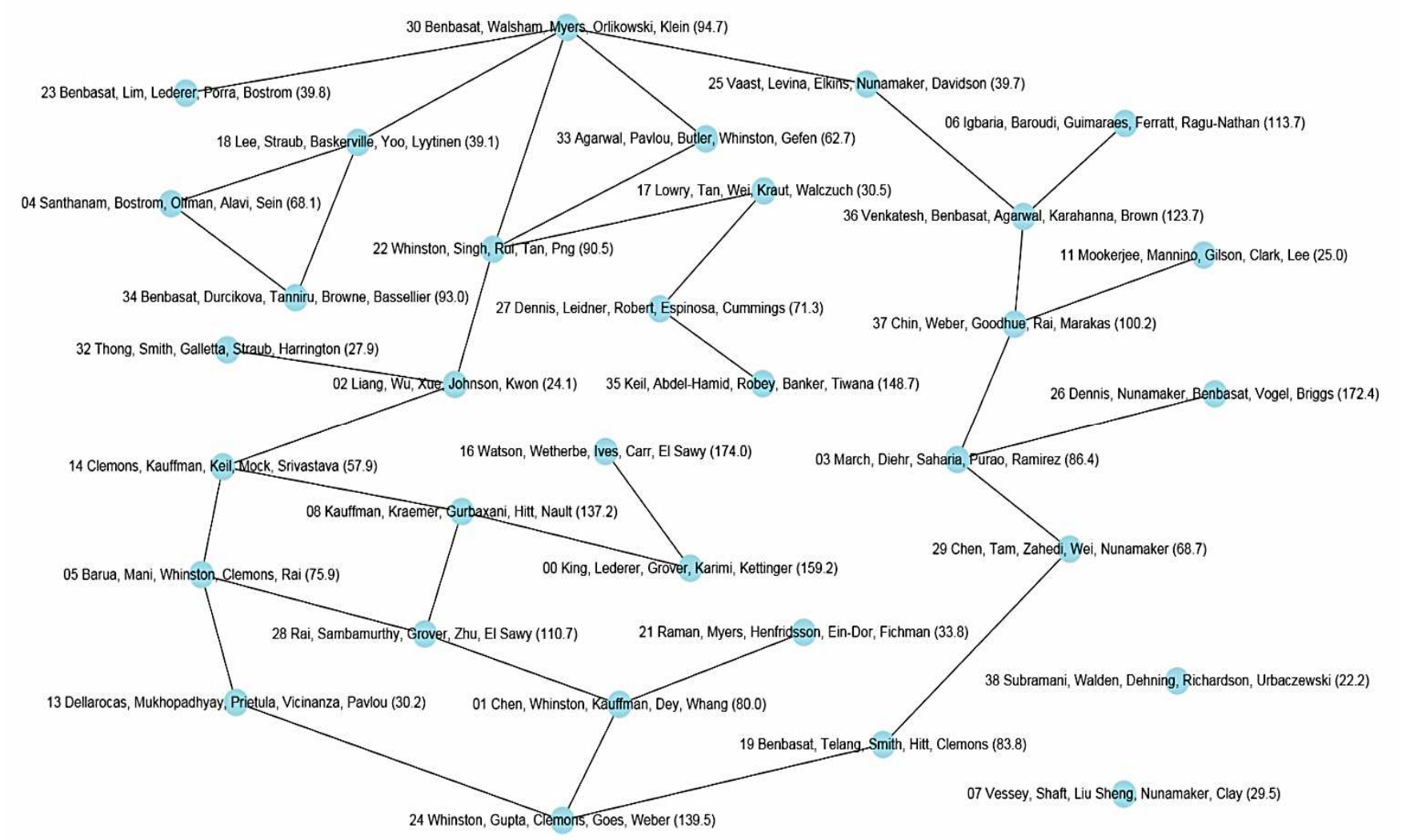

Figure 2. Top Researchers on the IS Topic Graph

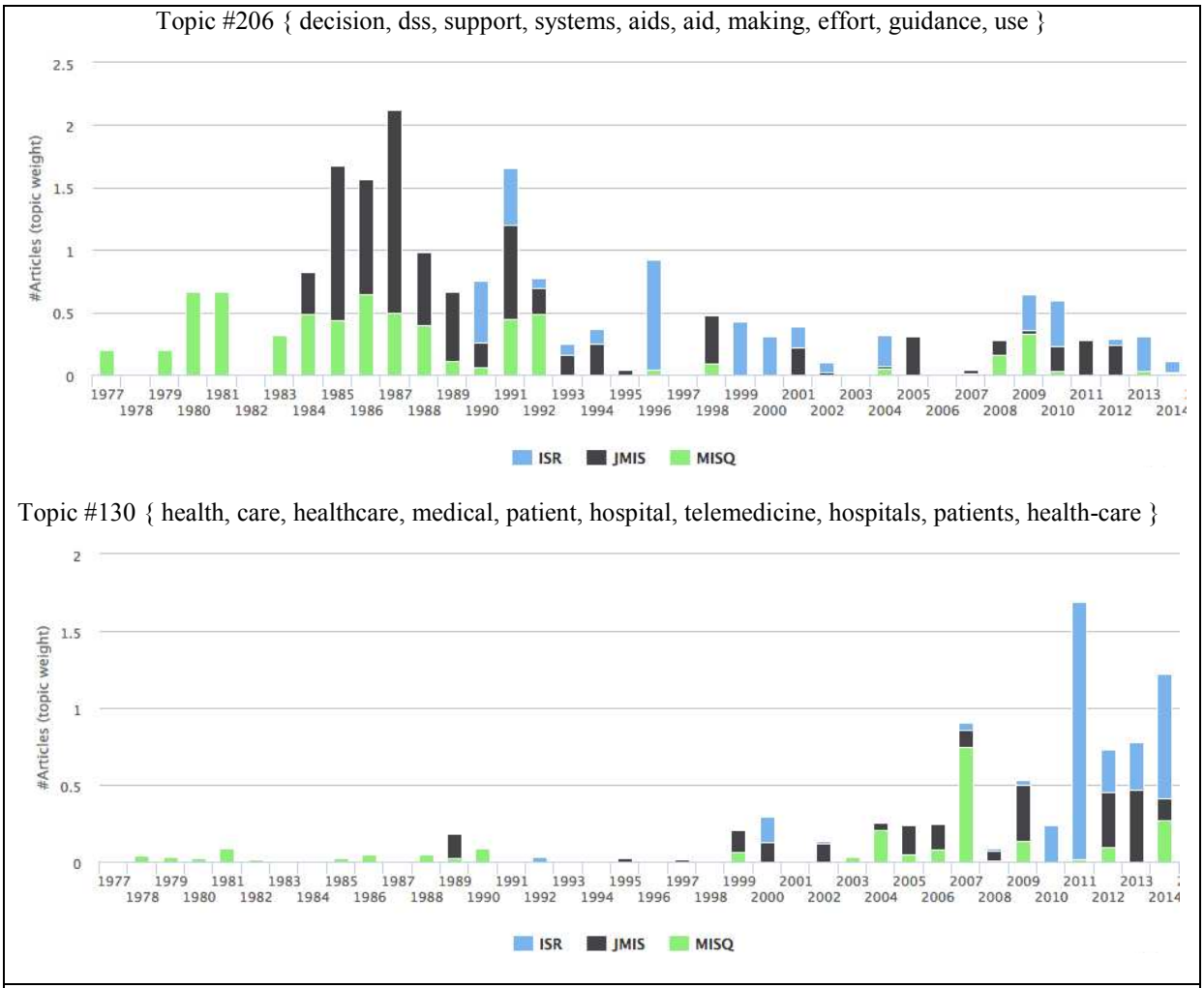

Figure 3. "Fashion Waves" in Research Topics

Evolution of IS Research

While the IS Topic Graph is helpful in giving us an overall picture of the intellectual structure of the field, it does not allow us to explore various IS topics at a deeper level. For this purpose, we resort to the Latent Dirichlet Allocation (LDA) model as an alternative. We use the Gibbs sampling method introduced in Griffiths and Steyvers (2004) for training and inference. To evaluate the number of topics, we fit the LDA model with different number of topics ranging from 30 to 400 and compute the perplexity using 10-fold cross validation. Perplexity is a standard performance measure of different models in natural language processing; a lower value of perplexity indicates a better model performance. As the number of topics increases, the perplexity first follows a downward trend from 30 to 100 topics and then increases a little bit from 100 to 150 topics. However, the perplexity continues to decrease from 150 to 300 and then stabilizes between 300 and 400. In the topic modeling literature, large datasets may use up to 300 topics to train models (Wei and Croft 2006), so we refrain from specifying a larger number of topics. Blei (2012) suggests that held-out accuracy should not be the only criteria to select models and that the interpretability of the results is probably more important in topic modeling. After careful interpretations and comparisons, we find that both 100 and 300 topics produce reasonable results and differ primarily in the level of topic granularity. Thus, we present our results in 300 topics. Interested readers can visit the ISTopic website for results from other models.

After model training, we can make inferences on individual abstracts to describe their topic coverage. Following Baskerville and Myers (2009), we can show the rise and fall of research topics in IS by aggregating the topic proportions within all abstracts published in each year for each topic. Figure 3 presents the trend of two topics (decision support and healthcare) as examples. Visualizations for all topics are available at http://www.istopic.org.

To further examine the evolution of research topics over the history of the information systems discipline, we compile a list of top words extracted from the top 10 most popular IS research topics every year since 1977 in Table 2. Our results remain qualitatively similar if we sample a different number of top topics. In order to focus on the words that are meaningful, we exclude the following words and their plural forms: management, information, system, technology, data, article, paper, study, research, result, finding, approach, method, model, analysis, literature, number, based, use, used, and using.

Year	Top 7 words from the top 10 most popular topics each year
1977	organization, assessment, executive, development, need, effect, design
1978	effect, organization, assessment, development, need, planning, process
1979	organization, planning, design, corporate, reports, strategic, environments
1980	organization, assessment, relationship, framework, corporate, developing, auditing
1981	organization, assessment, making, planning, support, relationship, decision
1982	organization, executives, support, design, corporate, mail, strategic
1983	organization, development, effect, support, administration, graphical, experimental
1984	organization, making, support, design, decision, environments, developing
1985	organization, relationship, development, making, effect, support, design
1986	support, decision, making, computing, design, process, organization
1987	support, design, making, decision, framework, expert, environments
1988	effect, strategic, development, planning, task, support, design
1989	support, organization, computing, design, strategic, environments, gss
1990	development, effect, task, support, organization, applications, strategic
1991	effect, development, investment, task, support, software, strategic
1992	support, organization, relationship, strategic, models, task, effect
1993	development, task, effect, support, group, investment, knowledge-based
1994	support, organization, models, users, electronic, process, user
1995	support, task, effect, organization, measuring, expert, gss
1996	organization, support, investment, effect, group, task, gss
1997	relationship, development, effect, organization, group, task, support
1998	effect, group, investment, markets, task, firms, support
1999	electronic, group, effect, accurate, mail, gss, empirical
2000	investment, organization, empirical, effect, value, factors, projects
2001	organization, market, effect, support, seller, optimal, usage
2002	outcomes, development, work, field, effect, support, electronic
2003	work, field, performance, organization, firm, teams, investment
2004	intention, field, perceived, science, effect, principles, role
2005	organization, market, effect, performance, seller, measuring, optimal
2006	effect, online, role, e-commerce, relationship, business, standard
2007	product, effect, competition, market, relationship, consumer, organization
2008	online, effect, relationship, experience, role, e-commerce, consumers
2009	effect, online, bidders, modeling, controls, software, bidding
2010	empirical, market, effect, competitive, firms, product, seller
2011	investment, effect, firms, hospital, product, medical, modeling
2012	product, effect, empirical, investment, relationship, software, evidence
2013	effect, important, theory, behavior, consumer, product, organization
2014	effect, product, behavior, social, experiment, theory, relationship

Table 2. Evolution of IS Research Topics

From Table 2, we see several major chronological changes in the top 7 words. Below we summarize our observations.

1) We see that organization is the top word in the early years, and in recent years, other words such as support, relationship, product, effect, and investment have taken the top spot alternately.

2) IS topics are more geared towards internal corporate operations in the earlier years, but gradually IS topics have moved to external factors such as markets, customers, and relationships.

3) The word "support" showed up frequently before 2003 but has disappeared among the top 7 words after 2002. This might indicate that IS research has moved from data room concerns to boardroom concerns.

4) IS topics now include forefront business issues such as electronic commerce, consumer relationship, and market competition as shown in the 7 top words in recent years.

ISTopic: An Intelligent Literature Search Tool

Prior studies that aim to map the intellectual structure of the information systems discipline have not investigated the application value of their research results. This study utilizes the research topics identified by topic models to develop an intelligent literature search tool called ISTopic for the IS research community. At present, most commonly used literature search tools include general search engines (e.g., Google) and scholarly business databases (e.g., Business Source Complete). These tools offer full-text indexing and searching features and find articles mainly by keyword matching. Although they are easy to use, it is difficult to find all the relevant articles for a given research topic. The reason is that the completeness of the search results largely depends on how informed of the topic the user is. A user needs to perform multiple searches using different keywords associated with the topic. Across the search results for different keywords, there are often some duplicate articles. The user has to go through these search results and merge all relevant articles to form the final article list. If the user is not familiar with the topic and misses some important keywords, it is possible that the user would also miss some relevant articles. Another drawback with this traditional approach is that many irrelevant articles enter the search results and it takes a lot of time for the user to filter them out. Using ISTopic, a user only needs to conduct one search with a keyword relevant to the topic instead of searching multiple times. The keyword or query term supplied by the user can be a single word or multiple words. Since each research topic is represented by a list of words in topic models, as long as the keyword specified by the user appears in a topic's list of words, that topic is identified as a potential match. This way it does not require the user to be familiar with the topic and know all the different but pertinent keywords. In addition, the number of topics (40 for CTM and 300 for LDA) is much smaller than the number of articles, so the search result for any keyword on ISTopic is often very short (i.e., one to three potential matches in most cases). The relevancy of a topic can be determined by the sequence of the matched keyword appearing in the topic's list of words, because the words in a topic are already sorted by their importance to the topic. After the target topic is found, the final article list is then shown to the user. In short, ISTopic can serve the IS research community at least in the following three practical aspects: (1) researchers can perform their literature search tasks using ISTopic's topic search and article search features; (2) the trending of different research topics is helpful for research students to select a dissertation topic; and (3) researchers can find potential reviewers and collaborators for their research projects.

Conclusion

Driven by the need to understand the structure and evolution of IS research, this study utilizes topic models to conduct disciplinary analysis as well as researcher analysis and develops an intelligent literature search tool. By extending a topic analysis method called Correlated Topic Model (CTM), this study reveals the connections among major research areas and researchers in IS at an intuitive level. With a chronological analysis of topics via Latent Dirichlet Allocation (LDA), we shed light on the evolution of IS research topics based on three premier IS journals, showing a trend of topic diversification and externalization. We also build an intelligent literature search tool called ISTopic for the IS community to explore the value of our research for their own needs. ISTopic is available online for public access at http://www.istopic.org and supports visualization of the "fashion waves" in IS research. For future research, we plan to include other IS journals in our analyses and also expand the coverage to other business disciplines.

References

Alavi, M., and Carlson, P. 1992. "A Review of MIS Research and Disciplinary Development," *Journal of Management Information Systems* (8:4), pp. 45-62.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. 2009. "On Smoothing and Inference for Topic Models," *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27-34.

Baskerville, R. L., and Myers, M. D. 2009. "Fashion Waves in Information Systems Research and Practice," *MIS Quarterly* (33:4), pp. 647-662.

Berthon, P., Pitt, L., Ewing, M., and Carr, C. L. 2002. "Potential Research Space in MIS: A Framework for Envisioning and Evaluating Research Replication, Extension, and Generation," *Information Systems Research* (13:4), pp. 416-427.

Blei, D. M. 2012. "Probabilistic Topic Models," *Communications of the ACM* (55:4), April, pp. 77-84.

Blei, D. M., and Lafferty, J. D. 2007. "A Correlated Topic Model of *Science*," *The Annals of Applied Statistics* (1:1), pp. 17-35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993-1022.

Culnan, M. J. 1986. "The Intellectual Development of Management Information Systems 1972-1982: A Co-Citation Analysis," *Management Science* (32:2), pp. 156-172.

Culnan, M. J. 1987. "Mapping the Intellectual Structure of MIS, 1980-1985: A Co-Citation Analysis," *MIS Quarterly* (11:3), pp. 341-353.

Griffiths, T. L., and Steyvers, M. 2004. "Finding Scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America* (101:1), pp. 5228-5235.

Meinshausen, N., and Bühlmann, P. 2006. "High Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics* (34:3), pp. 1436-1462.

Rainer Jr., R. K., and Miller, M. D. 2005. "Examining References Across Journal Rankings," *Communications of the ACM* (48:2), February, pp. 91-94.

Sidorova, A., Evangelopoulos, N., Valacich, J. S., and Ramakrishnan, T. 2008. "Uncovering the Intellectual Core of the Information Systems Discipline," *MIS Quarterly* (32:3), pp. 467-482.

Taylor, H., Dillon, S., and Van Wingen, Melinda. 2010. "Focus and Diversity in Information Systems Research: Meeting the Dual Demands of a Healthy Applied Discipline," *MIS Quarterly* (34:4), pp. 647-667.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society Series B (Methodological)* (58:1), pp. 267-288.

Vessey, I., Ramesh, V., and Glass, R. L. 2002. "Research in Information Systems: An Empirical Study of Diversity in the Discipline and its Journals," *Journal of Management Information Systems* (19:2), pp. 129-174.

Wei, X., and Croft, W. B. 2006. "LDA-Based Document Models for Ad-hoc Retrieval," *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178-185.