Using Latent Semantic Analysis to Identify Themes in IS Healthcare Research

Emergent Research Forum Papers

Arun Aryal
Robinson Coll. of Business
Georgia State University
Atlanta, GA, USA
arun.aryal@gmail.com

Mike Gallivan
Robinson Coll. of Business
Georgia State University
Atlanta, GA, USA
mikegallivan@yahoo.com

Youyou Tao
Robinson Coll. of Business
Georgia State University
Atlanta, GA, USA
tao.you.you@hotmail.com

Abstract
Latent Semantic Analysis (LSA) is a new text mining approach that is increasingly being adopted by IS scholars. In this paper, we provide an overview of various research contexts in which IS and other business scholars have applied this research method. We first identity the diverse body of scholarly and field-based contexts in which LSA has been applied. Next, we propose an empirical analysis of published research on healthcare information technology (HIT) research, to identify different themes in the IS literature from 1990 to the present date. Our empirical analysis will identify key research trends in IS journals topics for three time periods, based on an analysis of papers published in 15 leading IS journals. In addition to providing more awareness of this research approach, we seek to identify important trends in healthcare IT research over times.

Keywords
Health, healthcare, hospital, latent semantic analysis, lexical analysis, medicine, nursing, text mining.

Introduction
In recent years, text-mining methods have been adopted by business scholars to identify patterns and themes in documents. Such methods are useful for identifying themes in company documents, in web-sites and advertising text, as well as in published academic research. Originating in computer science over two decades ago, text-mining methods have increasingly been employed by behavioral researchers in IS. Our study first reviews the application of these methods in various business disciplines, including IS. When results from text mining are combined with other empirical analysis methods – such as Principal Components Analysis – it is possible to identify different groups of documents, such as specific research streams within a body of published work. In the IS field, the best-known example employing text mining was an empirical study published by Sidorova and colleagues (2008) in *MIS Quarterly*. When the authors combined Latent Semantic Analysis (LSA), a specific text-mining approach, with Principal Components Analysis, they identified 13 distinct research topics in three IS journals over a 30-year period, as well as five levels of analysis for IS research.[1] Other empirical studies of interest to general IS scholars that used LSA identified the relationship between IS and other business disciplines (Hovorka, Larsen & Monarchi 2008), as well as changes in coauthorship networks among IS scholars over time (Xu, Chan & Tan 2013).

Beyond these studies that have used text mining methods on the words appearing in the abstracts of published IS papers, other scholars have applied LSA to map the content within corporate documents (such as annual reports) or to transcripts of interactions between customers and service specialists. In IS, for example, researchers have analyzed firms' annual reports (Kloptchenko & Eklund 2002), customers' social networks (García-Crespo et al. 2008) and interactions between customers and call center staff (Coussement & van den Poel 2008). Other studies analyzed website content to identify employees'

[1] These authors analyzed papers published in *MIS Quarterly, Information Systems Research* and *Journal of Management Information Systems*, since 1990. Their analysis also yielded a more fine-grained analysis of 100 topics in these journals, in addition to the factors corresponding to 13 major research themes and five levels of analysis.

Internet abuse at work (Chou et al. 2010) as well as words appearing in patients' electronic health records in order to identify the causes of fall-related injuries (Chiarini-Tremblay et al 2005).

While the IS field has been an early adopter of LSA and other text-mining tools – given our proximity to computer science, other business disciplines have recently shown interest in text mining. Marketing scholars have used LSA to understand customers' online product reviews (Lee & Bradlow 2011), while accountants analyzed companies' annual reports to evaluate their initiatives related to environmental sustainability (Cho, Roberts, & Patten 2010) and social responsibility (DeWolf, Mejri & Lamouchi 2013). Of course, many studies in other business disciplines outside of IS have also analyzed the contents of published research papers' abstract to identify underlying themes in management journals (Evangelopoulos 2011; Huang et al. 2013) and conferences (Nag & Hambrick 2005; Zupic & Carter 2013).

Since LSA is an increasingly-common tool in IS research, we first provide an overview of the underlying logic of this method, followed by an illustration of its use to analyze a longstanding and increasingly dominant topic in IS research: healthcare IT. We describe a 30-year longitudinal analysis of the main themes in healthcare research published in IS journals, divided into three time periods. While most of the studies that we retrieved from leading IS journals focus on traditional medical contexts (i.e., hospitals and clinics), we also found papers that examine other health contexts, such as online patient communities, IT for the design and production of medical devices or pharmaceuticals, and IT for health insurance.

This research-in-progress paper will explain the specific steps that we followed to analyze 30 years of published health-related research in 20 leading IS journals. At AMCIS 2015, we will display various data visualizations of results (consisting of dominant research streams) for each of the three time periods.

Description of the Method: Latent Semantic Analysis

LSA is a method for extracting the contextual meanings and concepts from text documents. The first step is to read and input file of text. In doing so, the researcher will typically transform words that contain many spelling variants (e.g., organize, organization, organizing, etc.) into "word stems" – so that various grammatical and spelling variations are recognized as having the same meaning. The second step is the creation of a document matrix-vector – which is comprised of two elements: words and documents being analyzed (see Table 1). Documents are anything with a "semantic structure" that an analyst seeks to interpret. For example, documents may be company websites, annual reports, blog posts, or advertising copy.

Word frequency		Documents			
		D1	D2	D3 . . .	D_N
	Word 1				
	Word 2				
	Word $_N$				

Table 1: Document Matrix

The third step in LSA is dimension reduction. The document matrix yields a large vector that needs to be reduced to smaller sets of meaningful concepts. One of the simplest and powerful dimension reduction approaches is Singular Value Decomposition (SVD). SVD is based on linear algebra, details of which are explained in earlier studies (Landauer et al. 1998; Martin & Berry 2007; Meltic & Marcus 2000). At the risk of oversimplifying, SVD finds the obvious patterns and trends within the document matrix. It does so by analyzing which words frequently appear in specific documents (frequency count), as well as other words that often appear nearby (known as *co-occurrences*). These patterns are presented as concepts.

While there are many software tools for performing LSA, one that has is popular in the IS field is *Leximancer*. Several IS studies using this tool have recently appeared (Aryal, El Amrani & Truex 2014; Crawford & Hasan 2006; Debuse, & Lawley 2009; Ridley & Young 2012). In Leximancer, concepts are defined as collection of keywords that "generally travel together throughout the text, for example a concept *building* may contain the keywords mill ...., tower, collapsed, etc. (Leximancer, 2011 p. 10). At the beginning of the process, Leximancer

uses the words that most frequently appear in the text as the keywords. Then, keywords are weighted according to how frequently these words occur within a two-sentence "chunks" of text containing the focal concept, compared to how frequently the same words occur elsewhere. Leximancer then clusters these concepts into higher-level themes. Themes are the highest-level construct, with each theme comprised of concepts that often appear together often in the same text "chunks." Leximancer provides results in the form of "overall" visual maps, where the user can view the themes, concepts, and sub-concepts (keywords that comprise each concept). Once the initial map appears, the user can change the theme size to adjust the grouping of concepts to select fewer but broader themes; conversely, he can drill down to more detail. The user thus selects the desired level of granularity of results.

Most studies that have analyzed published research studies, in order to identify different topics and research streams (Sidorova et al. 2008; Larsen, Monarchi et al 2008), have generally analyzed just the words appearing in the abstracts of published studies. While it is also possible to analyze full-text papers, we did not identify any such studies in our review. If the Leximancer analyst wants to analyze full-text papers, common words that appear in the text of *all* papers should be ignored (e.g., terms like research, study, model, concept, hypothesis, predict, collect, analyze, test, results, table, discussion, conclusion).

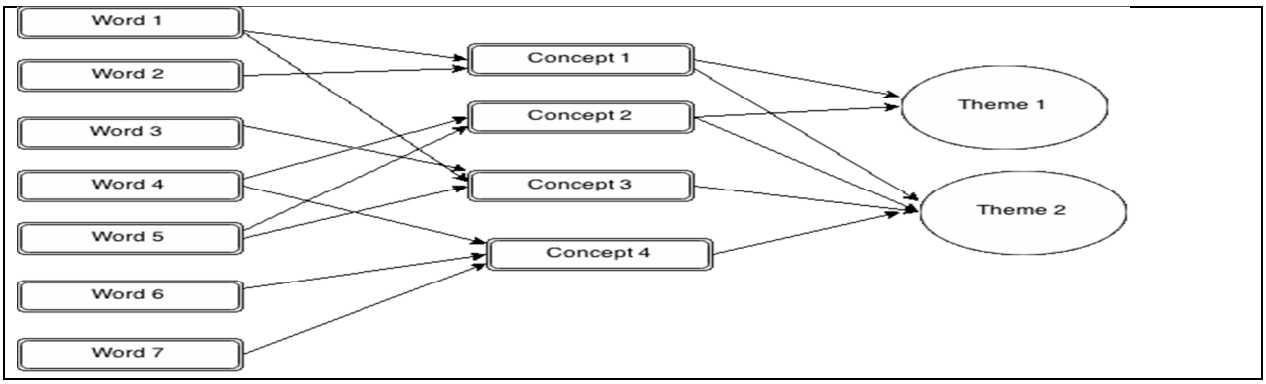


Figure 1: Leximancer processing: transforming words to themes

Leximancer produces visual diagrams, with certain key terms appearing in different-sized circles. Not only is the font size of the key term important, but the color of the circle encasing the word is important, as well. Specifically, the "hot" colors (hues including red, orange, and yellow) indicate that the theme has a stronger relationship with the underlying concepts – meaning that these concepts co-occur frequently with the high-level theme. In contrast, "cold" color circles (blue or purple) reflect a lower co-occurrence.

Research Questions

Question 1: What topics and themes can we identify in research on HIT published in IS journals?

Question 2: How do these research topics change over time, according to discrete time periods (1989-1998, 1999-2006, and 2007-2014)?

Question 3: How does the relationship between different research topics change over time?

Research Methods

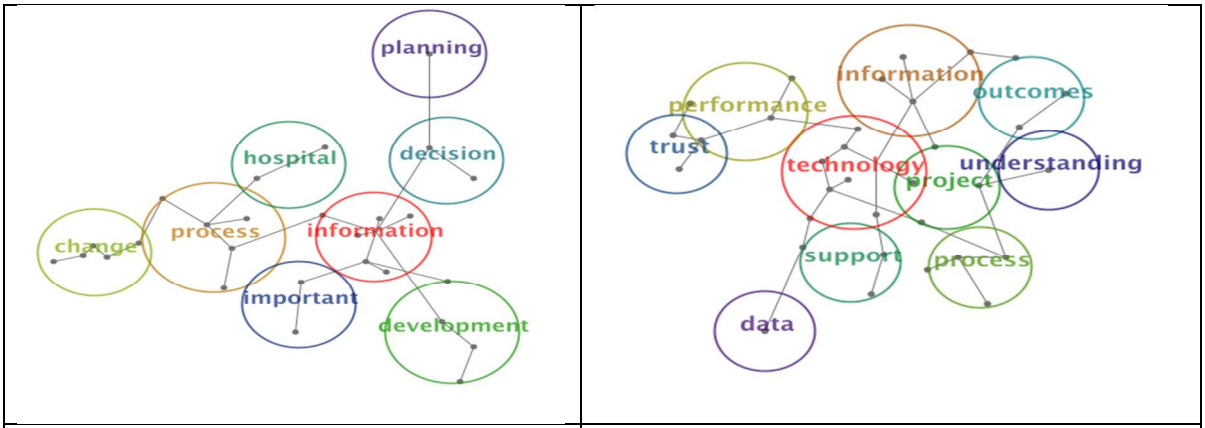
We conducted extensive searches, using multiple online databases, to identify papers in IS journals that focus on healthcare, including health insurance, pharmaceuticals and medical devices in 20 leading IS journals, starting in 1990 (or the first year of publication, if later).[2] In order to identify these papers, we searched a dozen terms in the full-text of papers appearing in these journals: clinic, doctor, health,

[2] ACM Computing Surveys, Communications of the ACM, Communications of the AIS, Decision Support Systems, European Journal of Information Systems, Information & Management, Information and Organization, Information Systems Journal, Information Systems Research, Information Technology & People, Journal of Information Technology, IEEE Transactions on Engineering Management, Journal of AIS, Journal of MIS, Journal of End User Computing, Journal of Organizational Computing, Journal of Strategic IS, Journal of Systems & Software, and MIS Quarterly.

hospital, medical, medicine, nurse, nursing, pharmacy, pharmaceutical, physician, surgery and surgical. We used nine databases to search for papers in these journals (ACM Digital Library, EBSCo Business Source, Emerald, ProQuest ABI/Inform, Sage, Science Direct, Palgrave, and Web of Science), in order to ensure complete coverage of these 20 IS journals. We also compared our results to studies listed in two reviews about healthcare IT (Chiasson & Davidson 2004; Romanow et al 2012) to ensure that we did not omit any relevant papers. After retrieving papers with these terms, we screened all papers to ensure that they focused on health-related contexts, thus excluding papers that mention these terms in passing, and papers that used these terms in another context (e.g., "doctoral" as a synonym for "PhD," health in the context of health club, etc.). This process to screen non-healthcare papers was labor intensive. On the one hand, we found that many papers that *were* set in healthcare contexts did *not* mention one of the search terms in their title, abstract, or keywords – so it was necessary for us to identify these terms in the full-text of the paper, and then retain these papers for our study. Conversely, many papers featuring such terms in their abstracts were not related to health or medical contexts. Thus, we carefully searched for the relevant terms in each paper's full text, in order to determine that it was indeed related to healthcare.

We created a spreadsheet showing basic bibliographic details for each of the 550 papers we identified. We coded each paper's title, publication year, as well as author names and journal title. We also created a master file for each of three time periods (1985-1998, 1999-2006, and 2007-2014), containing the title and abstracts of all papers published during that time period. Next, we excluded common "stop words" (and, not, with, or) and we instructed Leximancer to merge word variants (organize, organized, organizing, organization, etc.). Once these parameters for stop words and merge words were set, we allowed Leximancer to analyze the master file consisting of all remaining words appearing in paper abstracts for that time period.

Next we produced visual diagrams for each time period displaying the high-level themes. We also drilled down into each theme, in order to view the concepts that comprise each one. Next, we used SPSS to perform a Principal Components Analysis (Sidorova et al. 2008) to identify different sets of research papers, based on the themes and concepts that characterize them. We labeled each resulting factor with a "research stream title" that reflects the attributes shared by papers loading on that factor. The research stream title may identify the theoretical framework of the study (e.g., TAM, Actor-Network Theory, or Business Value of IT), as well as the research methods, or type of setting (e.g., online communities). For illustration purposes, we show preliminary results, based on a partial subset of papers in Time Periods 1 and 2. Figure 2 shows the results from a 20% sample of papers in each of the first two time periods.



Time Period 1 (1985-1998)	Time Period 2 (1999-2006)

Figure 2: Sample results with key terms for Time Period 1 (1985-1998) and 2 (1999-2006)

For Time Period 1, the themes are: *information, process, change* (the hot colors), as well as *hospital, development, decision*, and *planning* (the cool colors). For Time Period 2, the themes are *technology, information, performance* (the hot colors), and *support, process, outcomes, data, trust, understanding* (the cool colors). In our full analysis, we will show similar data visualizations, as well as the results of the Principal Components Analysis, from which we will identify "research stream" titles for each similar set of papers (based on papers sharing common theories, research methods, technologies, or research settings).

REFERENCES

Aryal, A., El Amrani, R., and Truex, D. P. 2014. Understanding the Emergent Structure of Competency Centers in Post-implementation Enterprise Systems. In Information Systems and Global Assemblages. (Re) Configuring Actors, Artefacts, Organizations. pp. 95-114. Springer Berlin Heidelberg.

Blake, R. 2010. "Identifying the Core Topics and Themes of Data and Information Quality Research," *Proceedings of AMCIS*.

Chiarini-Tremblay, M., Berndt, D. J., Foulis, P., and Luther, S. 2005. "Utilizing Text Mining Techniques to Identify Fall Related Injuries," *Proceedings of AMCIS,* 2005, 109.

Chiasson, M. W., and Davidson, E. 2004. "Pushing the Contextual Envelope: Developing and Diffusing IS Theory for Health Information Systems Research," *Information and Organization* (14:3), pp. 155-188.

Cho, C. H., Roberts, R. W., and Patten, D. M. 2010. "The Language of U.S. Corporate Environmental Disclosure," *Accounting, Organizations and Society* (35:4), pp. 431-443.

Chou, C. H., Sinha, A. P., and Zhao, H. 2010. "A Hybrid Attribute Selection Approach for Text Classification," *Journal of the Association for Information Systems* (11:9), pp. 491-518.

Coussement, K. and Van den Poel, D. 2008. "Integrating the Voice of Customers through Call Center Emails into a DSS for Churn Prediction," *Information & Management* (45:3), pp. 164-174.

Crawford, K., and Hasan, H. 2006. "Demonstrations of the Activity Theory Framework for Research in Information Systems," *Australasian Journal of Information Systems*, 13(2).

DeWolf, D., Mejri, M., and Lamouchi, R. 2012. "How do Multinational Corporations CEOs Perceive and Communicate about Social Responsibility?" *Research Journal of Finance & Accounting* (3), pp. 18-34.

Debuse, J., and Lawley, M. 2009. "Desirable ICT Graduate Attributes: Theory vs. Practice," *Journal of Information Systems Education*, 20(3), 313.sc

Evangelopoulos, N. 2011. "Citing Taylor: Tracing Taylorism's Technical and Sociotechnical Duality through Latent Semantic Analysis," *Journal of Business & Management* (17:1), pp. 57-72.

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. 2010. Management's Tone Change, Post-earnings Announcement Drift and Accruals. *Review of Accounting Studies*, (15:4), pp. 915-953.

García-Crespo, Á., Colomo-Palacios, R., Gómez-Berbís, J. M., and Ruiz-Mezcua, B. 2010. "SEMO: A Framework for Customer Social Networks Analysis based on Semantics," *Journal of Information Technology* (25:2), pp. 178-188.

He, W., Zha, S., and Li, L. 2013. "Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry," *International Journal of Information Management*, 33(3), 464-472.

Hovorka, D., Larsen, K. and Monarchi, D. 2008. "Conceptual Convergences: Positioning Information Systems among the Business Disciplines," *Proceedings of ECIS*.

Huang, K. P., Tung, J., Lo, S. C., and Chou, M. J. 2013. "A Review and Critical Analysis of the Principles of Scientific Management," *International Journal of Organizational Innovation* (5:4), pp. 78-85.

Indulska, M., Hovorka, D., and Recker, J. 2012. "Quantitative Approaches to Content Analysis: Identifying Conceptual Drift across Publication Outlets," *European Journal of Information Systems* (21:1), pp. 49-69.

Kloptchenko, A. and T. Eklund. 2002. "Combining Data and Text Mining Techniques for Analyzing Financial Reports," *Proceedings of AMCIS*.

Larsen, K., Li, J., Lee, J. and Bong, C.H. 2012. "Transdisciplinary Approach to Construct Search and Integration," *Proceedings of AMCIS*.

Larsen, K., Monarchi, D., Hovorka, D. and Bailey, C. 2008. "Analyzing Unstructured Text Data: Using Latent Categorization to Identify Intellectual Communities in Information Systems," *Decision Support Systems* (45:4), pp. 884–896.

Landauer, T. K., Foltz, P. W., and Laham, D. 1998. "An Introduction to Latent Semantic Analysis," *Discourse Processes* (25:2), pp. 259-284.

Lee, S., Song, J., Baker, J., Kim, Y., and Wetherbe, J. 2011. "The Commoditization of IT: Evidence from a Longitudinal Text Mining Study," *Communications of Association for Information Systems* (29:12), pp. 221-242.

Lee, T. Y., and Bradlow, E. T. 2011. "Automated Marketing Research Using Online Customer Reviews," *Journal of Marketing Research,* (48:5), pp. 881-894.

Leximancer. 2011. Leximancer Manual, Version 4.

Ma, T.R., and Cantu, F. 2008. "Knowledge Discovery in Academic Registrar Data Bases using Source Mining: Data and Text," *Proceedings of AMCIS*, 2008.

Maletic, J. I., and Marcus, A. 2000. "Using Latent Semantic Analysis to Identify Similarities in Source Code to Support Program Understanding," *Tools with Artificial Intelligence*, pp. 46-53

Martin, D. I., and Berry, M. W. 2007. Mathematical Foundations Behind Latent Semantic Analysis. Handbook of Latent Semantic Analysis, pp. 35-56.

Nag, R., and Hambrick, D. C. 2005. "What Is Strategic Management, Really? A Consensus View on the Essence of the Field," *Proceedings of the Academy of Management Conference*.

Ridley, G., and Young, J. 2012. "Theoretical Approaches to Gender and IT: Examining Some Australian Evidence," *Information Systems Journal*, 22(5), 355-373.

Romanow, D., Cho, S., and Straub, D. 2012. "Editor's Comments: Riding the Wave: Past Trends and Future Directions for Health IT Research," *MIS Quarterly*, 36(3), iii-xvi.

Sasson, E., Ravid, G. and Pliskin, N. 2014. "Text Mining and Temporal Trend Detection on the Internet for Technology Assessment: Model and Tool," *Proceedings of ECIS Conference*.

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

Xu, J., Chau, M., and Tan, F. 2014. "The Development of Social Capital in the Collaboration Network of Information Systems Scholars," *Journal of the Association for Information Systems* (15), pp. 835-859.

Yu, Y., and Li, Y. 2012. "The Research Trend on Global Software Development from 1999 to 2011: The Latent Semantic Analysis Approach," *Proceedings of Southeast Decision Sciences Institute* (SEDSI).

Zupic, I., and Cater, T. 2013. "Bibliometric Methods in Management and Organization: A Review," *Proceedings of the Academy of Management Conference*, pp. 13426.