2014

ACADEMIC HOT-SPOT ANALYSIS ON INFORMATION SYSTEM BASED ON THE CO-TERM NETWORK

Zezhong Zhang
Renmin University of China, joshua_zhang@foxmail.com

Li Yu
Renmin University of China, buaayuli@ruc.edu.cn

ACADEMIC HOT-SPOT ANALYSIS ON INFORMATION SYSTEM BASED ON THE CO-TERM NETWORK

Zezhong Zhang, School of Information, Renmin University of China, Beijing, joshua_zhang@foxmail.com

Li Yu, School of Information, Renmin University of China, Beijing, P.R.China, State Key Laboratory of Digital Publishing Technology, Beijing, P.R.China, buaayuli@ruc.edu.cn (corresponding author)

Abstract

The amount of research literature is increasing so fast that the scholars are hard to clearly know the state of art about a certain research field. For IS scholars, understanding research hot-spots among numerous academic papers on IS field is always a significant and key task. In this paper, taking Information System field as example, an academic hot-spot analysis method is proposed to automatically find the research hot topic. Firstly, based on the key words of literatures, a co-term network is build, then fast greedy clustering method is used to find research topic, and hot degree of research topics are computed. After downloading the literature information about IS field from WEB OF SCIENCE, research hot topic during latest five years is identified. The result show that three general topics, respectively as GIS, Health-Care IS, Management & Internet, are important research direction. Then, the hot-spots analysis method, which decomposes the "Management & Internet" topic into 10 topic communities, generates and discussed the IS trends on top 5 academic topics of each year from 2009 to 2013 and the heat map of the IS hot-spots in 2013.

Keywords: information system, co-term network, hot-spot analysis, academic topic.

1 INTRODUCTION

Every researchers need to go through papers one after one in order to obtain information they need (Yin, S., et al. (2014, January)), such as hot-spots in one field, so understanding research hotspots among numerous academic papers has become significant task. The amount of academic knowledge is increasing so fast that no expert can capture the entire knowledge structure of a specific knowledge domain, which means it is even not a easy task for IS scholars to fetch the suitable fresh hot topics from thousands of papers related to "information system" that generated each year. Especially, for fresh researchers, they really need the information on trends and hot-spots.

To analysis huge amounts of scientific articles in various research field, Network-based approach, includes co-term network, direct-citation network, co-citation network, bibliographic coupling network and co-author network, are always in use. In this paper, a co-term network are mainly used to fetch the hot-spots and identify trends in research output during the past five years and attempt to generate a wide view of IS discipline with author and citation information.

The rest of the paper is organized as follows. First, we review the IS discipline and topic analysis methods. Then, the co-term network that defined and community discovery method that used to detect topics in this paper are explained, and the hot-spots analysis indicators and methods, which includes hot-spots identification and evolution, are discussed. The results and discussion of this analysis are then presented, focusing on the past-five-year hot-spots in the IS field. We conclude with a summary of findings, limitations, and directions for future research.

2 LITERATURE REVIEW

Information system is a multi-discipline that connects business and computer science. Several IS scholars, such as Culnan, M. J. (1987), Keen, P. G. (1980), Lee, A. S. (1991), Mingers, J., & Stowell, F. (1997), have debated the nature and foundations of Information Systems which has its roots in other reference disciplines such as Computer Science, Engineering, Management Science, and others.

Over the years, researchers have been looking for the intellectual core of IS (Sidorova, A. et al. (2008)), but as the development of information system, thousands of IS papers published each year, IS scholars, especially fresh researchers, have to use topic analysis methods to explore hot-spots.

Among all article analysis methods, co-term analysis using literature focus on vocabulary or common noun phrase, to determine the instance represents the relationship between the subject of each topic. Co-term analysis may be a useful way of describing the patterns and networks embedded in a number of types of text (Jacobs, N. (2002)). A pair of vocabularies to appear in the same paper, the more represents that the closer the relationship between the two subjects or themes.

Based on the principle to construct co-term network, co-term analysis uses the clustering analysis and other statistical analysis methods, to show the complex networks relationship between the numerous analysis objects. Jacobs, N. (2002) pointed that in passing that methodologies closely related to co-term analysis are being used increasingly in a number of disciplines.

3 DATA AND CO-TERM NETWORK

In this paper, we uses the past-five-year, from 2009 to 2013, english jurnal articles under the theme "information system" or "information systems" on the Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI) and Arts and Humanities Citation Index(AHCI). The Scientific paper database that used during the research is the famous Web of Science.

The features of article that needed are title, DOI, Year, authors, keywords, cited frequency. The DOI and title are both used to distinguish different papers, and keywords are used to construct the co-term network, then the authors, year and cited frequency are used in hot-spots analysis.

Totally, 155,207 valid papers are obtain in this research. These papers are generated from related fields, such as computer science, management science, medical, and so on. The mainly journals that involved are Information Systems Frontiers, Decision Support Systems, MIS Quarterly, Expert Systems With Applications, and so on. Totally, there are over 1000 journals involved in.

In this paper, co-term network is constructed by using keywords from each paper. The network of co-term analysis are matrices of co-occurrence. During the research, the keywords are vertices in co-term network,especially the frequency of each keywords represents the weight of the vertex, and the edge between two vertices means the two keywords appear at the same time at least in one same paper. Then, the weight of edge will be the number of papers that the two keywords appear at the same time.

4 ACADEMIC HOT-SPOTS DETECTION

In research from Blei, D. M., & Lafferty, J. D. (2007), A topic can be defined as a probability distribution over terms in a vocabulary, while in this research, one topic is described by a series of keywords. Then, keywords for each paper are derived from a mixture of topics (Blei et al. (2003)).

For topic detection, co-term network cluster analysis and representative keywords selection for each cluster are two important tasks. A hierarchical algorithm should be used to deal with the first task, because there are hundreds of communities to merge after community discovery. Many community structure discovery methods are considered, such as walktrap that proposed by Pons, P., & Latapy, M. (2005), fast greedy that proposed by Clauset, A., Newman, M. E., & Moore, C. (2004).

This paper uses the fast greedy algorithm for inferring community structure from network topology which works by greedily optimizing the modularity(Clauset, A, et al.(2004)). This fast greedy algorithm can easily deal with our co-term network with tens of thousands of vertices, and another advantage is that fast greedy discovers a hierarchical communities. Also, Clauset, A, et al.(2004) show that fast greedy algorithm can extract meaningful communities from this network.

Maximal connected component of each network is used to do the cluster analysis, because even the second largest component only includes 3 to 5 papers.Then, the fast greedy algorithm is used to detect different communities which represent the different topics.

Here, an example shows below, which is a small topic in 2009. In Figure 1, it is obviously that "food", "nutrient", "data management" are special part in the field of health-care IS. According to hierarchical algorithm, this "food and nutrient data management" community is of course a part of "health care,information technology,epidemiology" community, after merging communities to generate only 10 communities.

Another task is the representative keywords selection, after community detection. The representative keywords of one community should be a summary of the set of keywords. To fetch the representative keywords of one community, the degree centrality shows the dominated position of the vertex, while the betweenness centrality presents the importance of vetex in maitaining the connectivity of the community, which means the vertex is absolutely necessary.

In this research, keywords that both exist in the top 5 of degree and betweenness centrality set are fetched as the representative keywords of one community. According the example that provided above, both of the centrality shows that "food composition","nutrient assessment","dietary assessment","food data management" are the key of the example community.

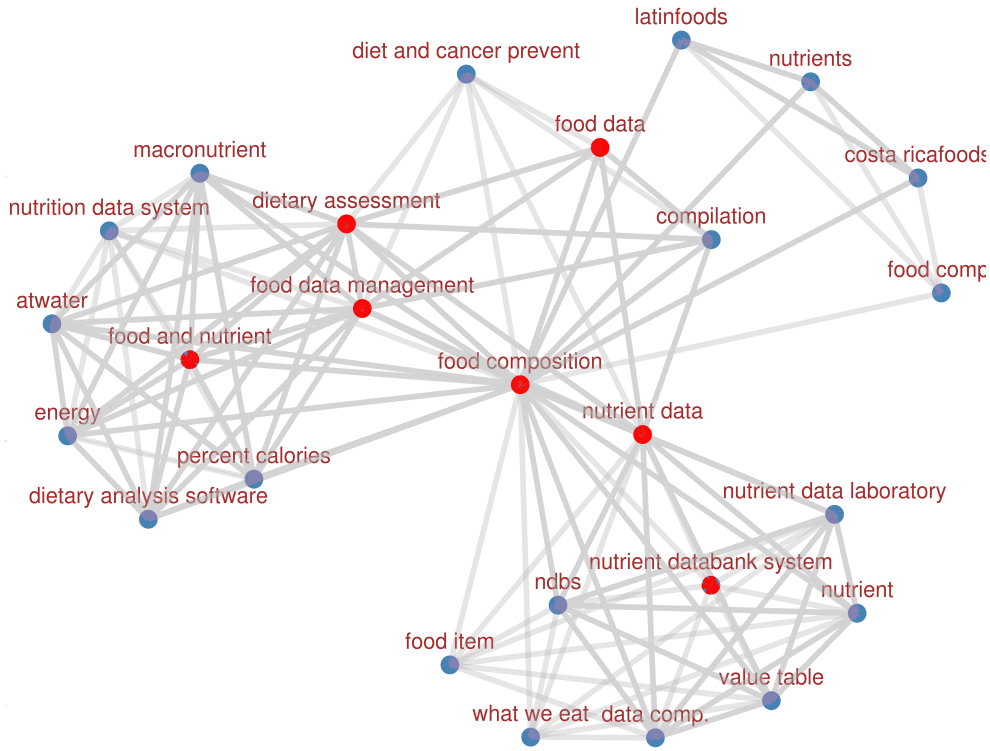

Figure 1. *Example Topic: one of the special field in information system, which related to food and nutrient data management. From these networks, not only typical and mainstream research topics but also special field related to information system can be detected.*

Using the example community that showed above, the list of top 5 keywords, which sorted by the degree centrality, are presented below.

Keywords in Community	Degree Centrality	Betweenness Centrality
food composition	25	169.333333
nutrient data	14	20.000000
dietary assessment	13	9.333333
food data management	13	9.333333
energy factors	9	Not in the top 5, the value is 0

Table 2. *Top 5 representative keywords of the example community: degree centrality and betweenness centrality are both presented here. But for short description, only the top 1 keyword is mostly used.*

Though this example seems a special field of minority, it is a "information system" related topic on the field of health care. To understand all the topics like Figure 1 and Figure 3, the merging of the community according to hierarchical structure is neccessay.

To make the community be more meaningful, communities are merged according to the hierarchical structure. Through this approach, we firstly generated 5 communities of each year, the next section provides a seiries of indocators and methods to fetch the hot-spots.

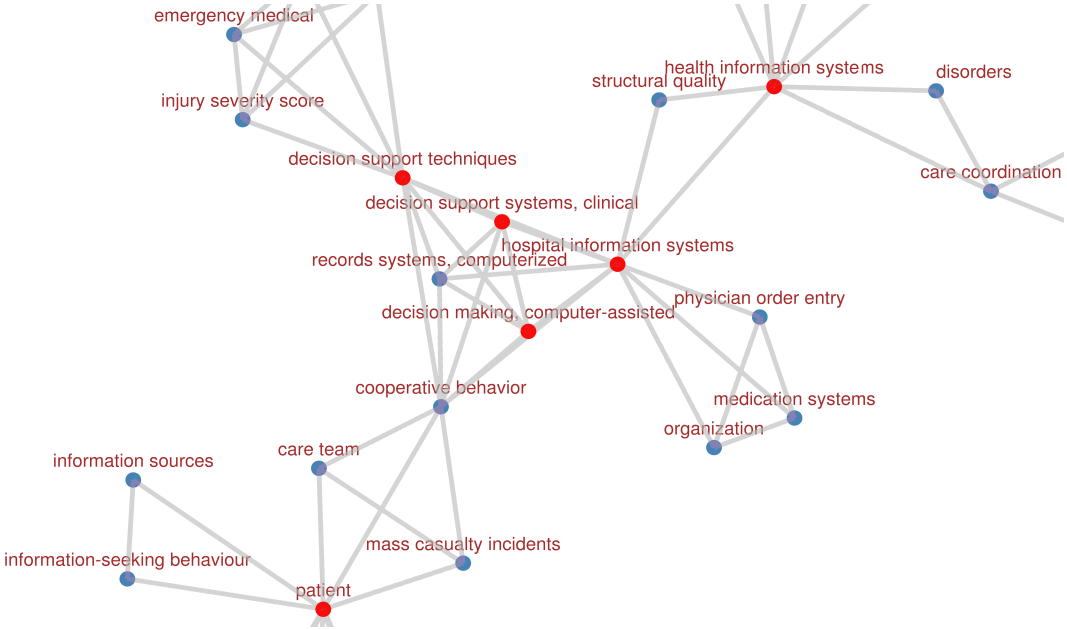

Figure 3. *Part of another Example Topic: the red dots shows the keywords with high degree inside the community, the top 5 of red dots are accepted as representative keywords. Example of another part related to "health care" community, which has the same level with the "food and nutrient data management" community.*

5 ACADEMIC HOTSPOTS ANALYSIS

Hotspot identification is to detect the hot-spots and visualize different hot topics. All this indicators are using the basic information of the community, such as the number of papers related to the i^{th} community P_i, the number of keywords in the i^{th} community K_i, the number of authors of the j^{th} paper P_j, the weight of j^{th} keywords W_j, and the cited number of j^{th} paper C_j.

The Research Attention Degree (RAD), is the indicator that the research outputs degree of one topic. For the i^{th} community, the more papers are published, the more attention are paid.

$$RAD_i = \frac{P_i}{K_i} \quad (1)$$

Research Contribution Degree (RCD), which RCD uses the cited number of each paper, focus on the quality of the research outputs in one topic.

$$RCD_i = \frac{\sum_{j=1}^{P_i} C_j}{P_i} \quad (2)$$

Second Research Attention Degree (SRAD), which is a plus indicator of RAD, indicates the research outputs that be paid attention by others. SRAD indirectly shows the research attention on one topic.

$$SRAD_i = \frac{\sum_{j=1}^{P_i} C_j}{K_i} \quad (3)$$

Author Attention Degree (AAD), is a author-considered indicator, which considers the situation of co-author. It indicates the author's attention in one topic.

$$AAD_i = \frac{\sum_{j=1}^{P_i} A_j}{K_i} \quad (4)$$

Author Contribution Degree (ACD), is the same idea like the RCD, which uses the cited number of each paper. ACD indicates the contribution degree of authors in each communities.

$$ACD_i = \frac{\sum_{j=1}^{P_i} C_j}{A_i} \quad (5)$$

Then, the topic density (TD), which is the average weight of each keywords, indicates the number of the important keywords that community has. Each topic includes many sub-network of keywords, the more important keywords that each sub-network contains, the hotter the topic is.

$$TD_i = \frac{\sum_{j=1}^{K_i} W_j}{K_i} \quad (6)$$

The last part is to simply analyse the trends of topics. For two generalized sets X and Y, the jaccard similarity is $|X \text{ intersect } Y| / |X \cup Y|$ where $|.|$ denotes the cardinality for generalized sets (sum of memberships). In this paper, after calculating the jaccard similarity of each topics from each year, the highest similarity are selected as the same topic of the next year. Because the papers are only from the recent five years, which is not quite long time, the phenomenon that merging of the topics are rare.

6 RESULT AND DISCUSSION

After a series of operations on data, we obtained totally 155,207 valid papers and 233,643 valid vertices of keywords. According the academic topic detection method, we list the top 5 representative keywords of communities after merging. According to the following Table, topics about GIS, Health-Care IS, Internet and Management are clear and detected from 2009 to 2013.

Topics in 2009					
No	keywords1	keywords2	keywords3	keywords4	keywords5
1	networks	electronic data processing	multimedia	scheduling	diffusion
2	education	computer-assisted instruction	learning	visualization	user involvement
3	management	supply chain management	data mining	internet	optimization
4	GIS	geographic information system	remote sensing	simulation	modelling
5	health care	information technology	epidemiology	integration	Performance
Topics in 2010					
No	keywords1	keywords2	keywords3	keywords4	keywords5
1	internet	information technology	performance	systems	design
2	networks	control	communication	simulation	security
3	estimation	virtual reality	visualization	diffusion	entropy
4	epidemiology	parkinson's disease	children	recognition	surveillance
5	GIS	geographic information system	remote sensing	uncertainty	climate change

Table 4. Each topics in 2009 and 2010: communities are merged into 5 of each year.

Using the similarity to calculate the 10 topics from 2009 to 2010, we generate the trends of the topics. In 2009, the "web service" are the hottest, then the "optimization","supply chain management (SCM)"; in 2010, the "knowledge management" went first; in 2011, the "data ming"have high average hot degree; in 2012, "decision support" went first; in 2013, "data ming" and "visualization" are the hottest. These are showed below.

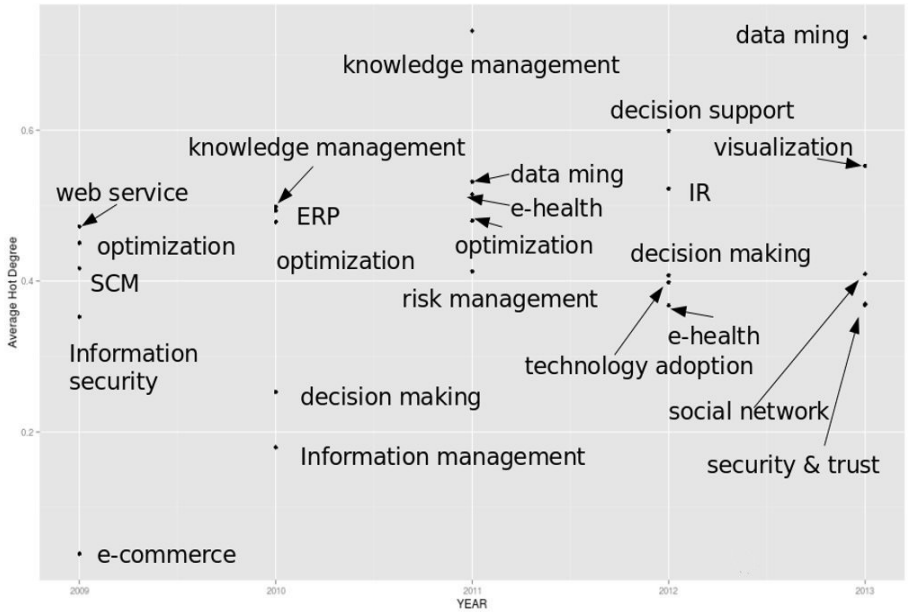

Figure 5. *trends on top 5 topics of each year from 2009 to 2013. the average hot degree is regularized of average attention degree (RAD, AAD, SRAD).*

Hot-spots in 2013 are discussed, and academic topic detection methods are used on the community "Management & Internet". So, the heat map of topics in 2013 are plotted as an example. From the Figure 6, ten hot topics are listed, and "data ming", "visualization", "social network", "security and trust" are the hottest cluster. Topics are clusted due to the same hot reasons.

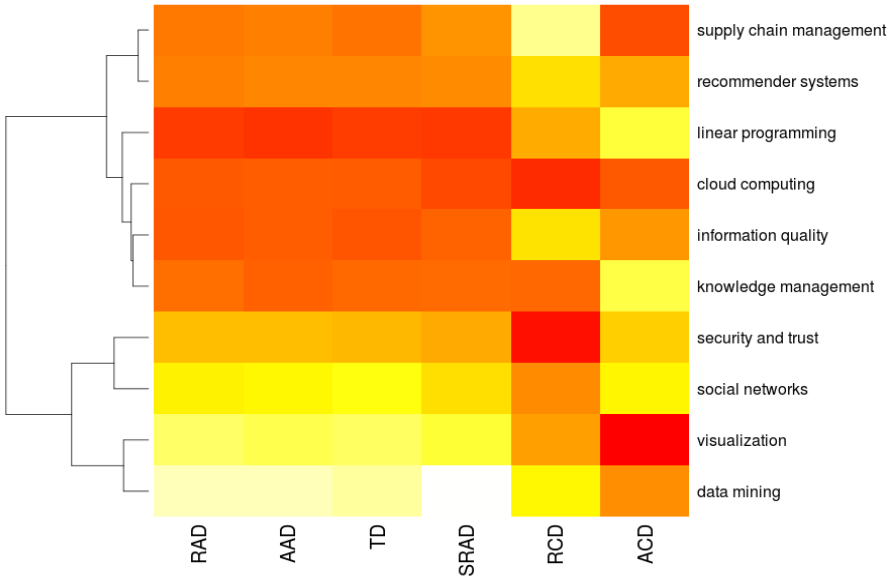

Figure 6. *heat map of ten topics in 2013: after community discovery of the "management and internet", the indicators about hot-spots analysis are used for clustering.*

For verification of this method, the google trend is used to analyse the keywords of hot-spots. An example of the two keywords "social network" and "knowledge management" on the google trends shows the attention degree on "social network" is higher than "knowledge management".

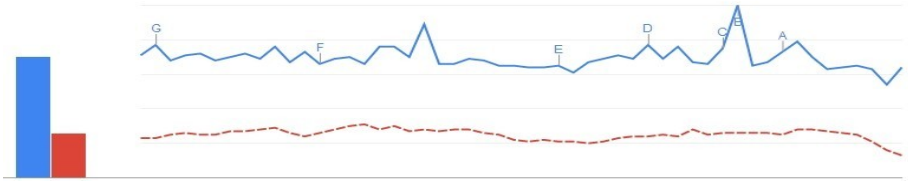

Figure 7. verification of topics in 2013: the red line is "knowledge management" and the blue line represents the "social network". It is clear that in 2013, the "social network" is hotter than "knowledge management".

Of course, it is not enough to verify the hot-spots in 2013 by only two keywords. In order to verify the whole 2013 topics, we compare any two representative keywords to check the heat map result. However, google trends only represents social attention.

7 CONCLUSION

In order to summarize and understand the hot-spots and trend of the field of information systems, this article provides an academic hot-spots detection and analysis method. Based on the co-term network, academic hot-spots indicators in paper can be divided in to three types: attention degree, contribution degree. Through these, general topics in the field of information system are summarized, which are GIS, Health-Care IS, Management & Internet. Then, Management & Internet is analyzed as an example, and the trends from 2009 to 2013 are clearly obtained and the heat map of 10 IS topics in 2013 is generated to assist on the hotspots identification and the hot reasons analysis.

In the future, the verification of the hot-spots is still a problem, though the google trends is an easy way. The google trends is the attentions from the public, not only the academic. Also,the indicators can use network features, which can be considered for further study. And this framework of method can also be realized as a system plugin for hot-spots analysis.

ACKNOWLEDGEMENTS

This work was supported by the grants from National Natural Science Foundation of China (No.71271209, No.71331007), Beijing Municipal Natural Science Foundation (No.4132067), Humanity and Social Science Youth Foundation of Ministry of Education of China (No.11YJC630268), Special Fund for Fast Sharing of Science Paper in Net Era by CSTD (No.2013112), Opening Project of State Key Laboratory of Digital Publishing Technology, Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China.

References

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. Annals of Applied Statistics, 1(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of machine Learning research, 3, 993–1022.

B. R. Martin. (1995). Foresight in science and technology. Technology analysis & strategic management, 7(2),139–168.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

Culnan, M. J. (1987). Mapping the intellectual structure of MIS, 1980-1985: a co-citation analysis. *Mis Quarterly*, 341-353.

Jacobs, N. (2002). Co-term network analysis as a means of describing the information landscapes of knowledge communities across sectors. *Journal of Documentation*, *58*(5), 548-562.

Keen, P. G. (1980). *MIS research: reference disciplines and a cumulative tradition* (pp. 9-18). Center for Information Systems Research, Alfred P. Sloan School of Management.

Lee, A. S. (1991). *Architecture as a Reference Discipline for MIS*. Management Information Systems Research Center, Curtis L. Carlson School of Management, University of Minnesota.

Mingers, J., & Stowell, F. (1997). *Information Systems: An Emerging Discipline?*. McGraw-Hill.

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005* (pp. 284-293). Springer Berlin Heidelberg.

Silver, M. S., Markus, M. L., & Beath, C. M. (1995). The Information Technology Interaction Model: A Foundation for the MBA Core Course. *MIS quarterly*, *19*(3).

Sidorova, A., Evangelopoulos, N., Valacich, J. S., & Ramakrishnan, T. (2008). Uncovering the Intellectual Core of the Information Systems Discipline. *Mis Quarterly*, *32*(3).

Yin, S., Li, C., Zhou, Y., & Huang, J. (2014, January). Detecting Hotspots in Insulin-Like Growth Factors 1 Research through MetaMap and Data Mining Technologies. In *Web Information Systems Engineering–WISE 2013 Workshops* (pp. 359-372). Springer Berlin Heidelberg.