FINDING OUT SUBJECT-MATTER EXPERTS and RESEARCH TRENDS USING BIBLIOGRAPHIC DATA

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Computer Engineering

by
Arzum KARATAŞ

September 2015
İZMİR

We approve the thesis of **Arzum KARATAŞ**

Examining Committee Members:

Assoc.Prof. Dr. Hürevren KILIÇ
Department of Computer Engineering, Gediz University

Assist. Prof. Dr. Selma TEKİR
Department of Computer Engineering, İzmir Institute of Technology

Assist. Prof. Dr. Tuğkan TUĞLULAR
Department of Computer Engineering, İzmir Institute of Technology

07 September 2015

Assist. Prof. Dr. Selma TEKİR
Supervisor, Department of Computer Engineering
İzmir Institute of Technology

Prof. Dr. Halis PÜSKÜLCÜ
Head of the Department of Computer Engineering

Prof. Dr. Bilge KARAÇALI
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGMENTS

I would like to thank to my thesis adviser Asst. Prof. Dr. Selma TEKİR, for her supervision, her respect to my individual opinions all along the work and for providing me the opportunity to perform such a joyful study. I appreciate her sincerity, patience and attentive style.

Further, I had the pleasure of working with Assoc.Prof.Dr. Hürevren KILIÇ. He was the one who uplifted me when I was in trouble with critical decisions. His precious support, and confidence have been the driving force of my courage and enthusiasm.

Next, I should thank to Asst. Prof. Dr. Serap ŞAHİN to step me up to complete my thesis and courage me to make my dreams come true.

Moreover, I would like to thank to Asst.Prof.Dr. Md. Haidar SHARIF and my dear colleagues in Department of Computer Engineering and friends working at the different departments at Gediz University. It has been a true pleasure to work in such a friendly environment.

Furthermore, I would like to thank to Asst.Prof.Dr. Tuğkan TUĞLULAR for his valuable evaluations for this work.

Finally, I give thanks to the ones who I missed to mention whilst they were willing to stay with me in my good and bad days. It's an immense blessing to know they are with me and stay.

ABSTRACT

FINDING OUT SUBJECT-MATTER EXPERTS and
RESEARCH TRENDS USING BIBLIOGRAPHIC DATA

With the prevalent use of information technology, it is very easy to reach nearly any information. However, if it is desired to be specialized in an area, the first thing to do is to know who are the experts in that area. Since experts have valuable knowledge, it is important to find these experts. Also, it is vital to be aware of trends for researchers who want to be expert in a topic or who want to enter into a new area. This work includes an empirical study for finding experts and research trends in academic world. We created a citation network from KDD proceedings and an author-keyword bipartite graph from bibliographic data of the same set of proceedings. Then, we applied link analysis algorithms HITS and PageRank, respectively. The results show that it is possible to detect two expert types (one that works intensively on a single subject and another having high level knowledge of various subtopics of a subject-matter). Moreover, topical trends are identified as doing peak, periodic, and having the same shape rather than showing absolute increase, decrease or stationary pose.

ÖZET

KONU UZMANLARININ ve ARAŞTIRMA EĞİLİMLERİNİN BİBLİYOGRAFİK VERİLER KULLANILARAK BULUNMASI

Bilişim teknolojilerin gittikçe yayılmasıyla her türlü bilgiye erişmek mümkün hale gelmiştir. Eğer bir kişi bir konuda uzmanlaşmak isterse, o konudaki uzmanları bilmek yapacağı işlerin başında gelmelidir; çünkü, uzman kişiler o konuda en değerli bilgiye sahip olan kişilerdir. Benzer şekilde bir konuda uzmanlaşmak ya da yeni bir araştırma alanına giriş yapmak isteyenlerin araştırma konularının eğilimlerinden haberdar olması gerekir. Bu tez akademik dünyada konu uzmanlarının ve araştırma eğilimlerinin bulunması için yapılan bir deneysel çalışmayı içermektedir. Veri olarak KDD bildirilerinden bir atıf ağı ve bu bildirilerin bibliyografik verisinden bir yazar-anahtar kelime çizgesi oluşturulmuştur. Bunlara sırası ile HITS ve PageRank link analiz algoritmaları uygulanmıştır. Çalışmanın sonucunda hem özel alanlarda çalışan yazarlar hem de bir ana konunun alt konularında yüksek seviyede bilgi sahibi olan yazarlar konu uzmanı olarak tespit edilmiştir. Ayrıca veri içinde bir örüntü anlamına gelen eğilimlerin; sadece düzenli bir artış, azalış yada sabit bir duruş olmadığı çizgede tepe yapma, periodik özellik gösterme ve benzer bir grafiğe sahip olma anlamına da gelebileceği görülmüştür.

To my dear Uncle,
I miss you so much ...

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Acronym	Definition
CN	Citation Network
AKG	Author-Keyword Bipartite Graph
dblp	Digital Bibliography & Library Project

CHAPTER 1

INTRODUCTION

1.1. Problem Statement

The prevalence of the Internet and the birth of web 3.0 created an environment where nearly everybody contributes to web of information. Due to the resultant diversity and information depth, in potential information varying from recipes to scientific experiments has become easily reachable. On the other hand, time is limited and generally people want to reach the information they need or want directly.

Information/knowledge has been one of the most valuable meta. In today's information era, data are everywhere, but knowledge is relatively quite rare. Experts are the people or systems that have the knowledge on a subject-matter. Actually systems are trained by the people. Thus, experts can be regarded as only people that have valuable knowledge on a subject-matter.

Information technology brings about fast changes in science and technology. Interests of people change fast as well. Trends are sprang out of these interests. Trend can have different meanings in different areas. For example, it is the currently preferred clothes in terms of fashion. It can mean most popular scents in terms of perfumes. It can mean red stilettos as for shoes. In general, trend means the most popular or the things that draw the highest attention. In particular, topical trend means how a topic changes over time.

The purpose of this thesis is to find out subject-matter experts and trends on scientific literature. The specific subject area chosen is Knowledge Discovery and Data Mining. The academic papers published in KDD proceedings between 2000 and 2014 are used. The reason for selecting KDD proceedings is twofold: First, it's one of the top conferences in this area and second, the published papers are easily accessible via ACM digital library[1].

1.2. Motivation

This thesis is within the area of the Knowledge Discovery and Data Mining. Knowing the leading researchers and trend topics is essential, motivating and joyful for a new researcher like the owner of the thesis. This information provides the new researchers with whom to follow and what to read. In fact, being aware of the state-of-the-art work is a crucial step of making a scientific contribution in any research area. In addition, the existence of powerful web-based systems like Google Scholar[2], Microsoft Academic Search[3], ArnetMiner[4], CiteSeerX[5] and Rexa[6],is inspirational to search academic world and reveal the academic experts and research trends.

1.3. Solution Approach

Citation network constructed out of bibliographic data is shown to indicate what publications are the most authoritative ones through the use of the link analysis algorithms like HITS. In a similar manner, alternative author-keyword bipartite graphs generated from bibliographic data should have potential to tell us who is expert in a subject area and how a research topic popularity changes over time. Link analysis done on the keyword-author bipartite graphs in combination with the results from citation graph analysis should provide new insights.

In the context of link analysis, it's vital to distinguish the concept of expert from that of authority as authority measurement is supported by the existing algorithms. In general, *expert* is someone having a special skill or knowledge obtained from training or experience and *authority* means an accepted source of information. They are so close to each other in meaning. Sometimes they can be used interchangeably in daily language. However, when the experimental setup is considered, experts have the same definition but authorities are the publications in the citation network, not authors directly but authority papers can imply that they are written by authoritative authors. The authority

publications can be directly measured by using link analysis algorithms like HITS; however, there is no direct measurement for being expert. Since research trends and subject-matter experts are high level concepts, it is not possible to directly measure these concepts from the network. Therefore, being expert as a high level concept is addressed by using directly measurable concepts like authority and importance via link analysis algorithms like HITS and PageRank respectively.

Kleinberg[7] says that links contain a hidden human judgment and this type of judgment is obtained via link analysis. This latent judgment is very useful for determining the importance, impact or authority inside of the global structure of the network. That is, let's assume that page p has a link to page q and it can be commented as creator of page p has in some measure conferred authority on q. For example, this latent conferred authority approach is used to reveal most authoritative documents in the citation network constructed. Similarly, in author-keyword bipartite graphs this latent judgment can give clues about expert nominees or the subject-matters that are worked by them.

As a data set the papers from KDD proceedings, dblp bibliographic data on KDD, ArnetMiner citation network data[9] are collected and used. More detailed information on the experimental work can be found in Chapter 4.

1.4. Thesis Organization

This thesis is organized into five chapters. The summary for each chapter can be found in the paragraphs below:

First, Chapter 2 gives the background information about the problem. In Section 2.1 citation networks/graphs are explained. In Section 2.2 bipartite graphs are introduced. Later, in Section 2.3 link analysis along with the fundamental link analysis algorithms HITS and PageRank is explained.

Next, Chapter 3 presents the review of literature. The history of the solution development starting from link analysis to expert finding models, how it is evolved and

what kind of solutions are proposed in the past are stated. In determining the methodology of this work, the inspired/adapted related work is discussed. In summary, first a big picture is drawn for the problem giving a start and then milestones for evolving solutions.

Later, Chapter 4 is dedicated to experimental part. This chapter contains every detail related to the problem solution approach. First of all, proposed system overview or big picture is presented to the reader. Then, this system is zoomed in by dividing it into some smaller functional pieces. These pieces can be ordered as data collection, data preparation or process, Author-Keyword graph construction for author names-keywords input data, citation network construction for the input citation network data, performing PageRank on the Author-Keyword graph, performing HITS algorithm on the citation network built, graph application of visualization and filtering mechanisms, and finally determining the experts and trends. Design and implementation issues of the proposed solution are discussed in detail. Also, the experimental results are obtained and evaluated from the perspective of expert and trend finding.

Finally, Chapter 5 summarizes the key points, and presents the results and future work .

CHAPTER 2

BACKGROUND

2.1. Citation Graphs

A citation graph is a directed graph in which each vertex represents a publication like paper or journal article and each edge represents the citation relation. In Figure 1, A gives a reference or citation to the B. It can be understood that the heads of the arrows show the publication(s) referenced.

Citation graphs are used in bibliometrics and information science. Bibliometrics is defined as “the statistical analysis of books, articles, or other publications” in the Glossary of Statistical Terms[10]. The citation graph of the web is an important source for web search engines. For example, Google uses citation graph of the web for the calculation of PageRank of a web page[8]. Also, Kleinberg[7] focuses the use of citation graph of web/documents for analyzing the collection of related pages to a search topic or query and finds out the most authoritative and hub pages related to that query.

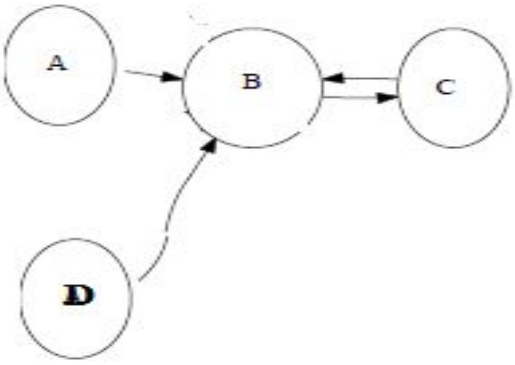


Figure 1. A Simplified Citation Network Example

A sub component of the citation network of the International Symposium on Graph Drawing and Network Visualization 2014[11] can be seen in the Figure 2. The vertices represent the papers published in the conference. The size of the vertices represents the number of the citations of a paper.

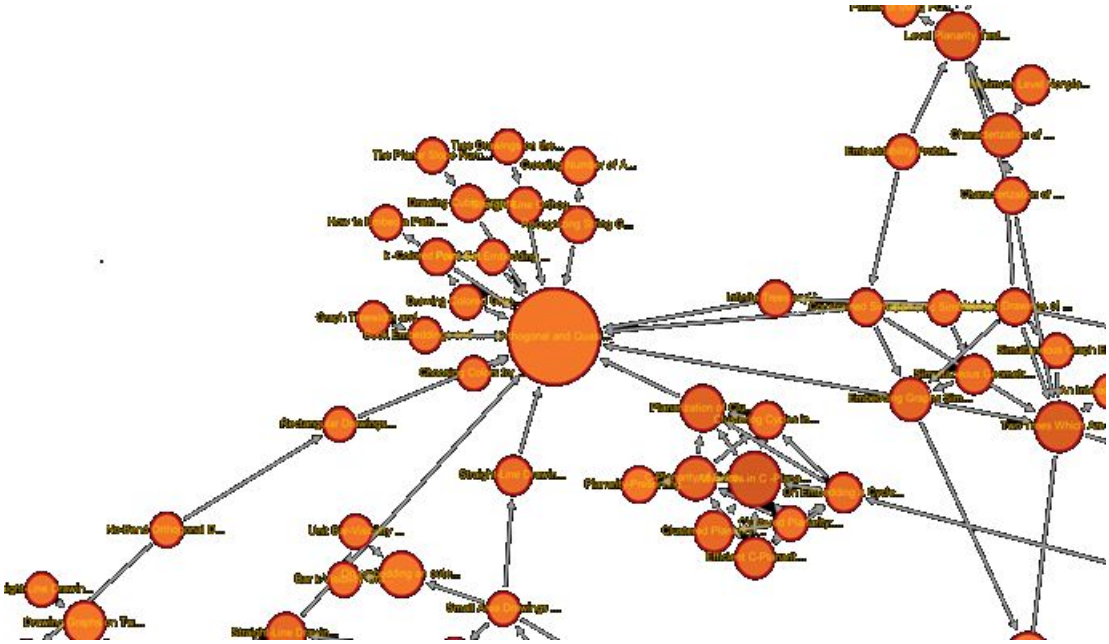


Figure 2. A Citation Network Example [12]

2.2. Bipartite Graphs

Graph theory glossary[13] prepared by Chris Caldwell defines the bipartite graph as "a graph whose vertices can be partitioned into two disjoint subsets U and V such that each edge connects a vertex from U to one from V." Actually, Wolfram's Web Resources[14] define a bipartite graph as a special case of k-partite graph. If k is two, then the graph is called as bipartite graph. Some illustrations showing bipartite graph can be seen in the Figure 3.

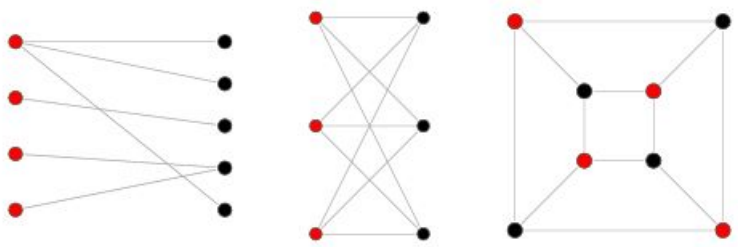


Figure 3. Three Sample Illustration of Bipartite Graphs [14]

Article entitled as “Bipartite Graph” resides on the Wikipedia[15] mentions that if the all vertices of a graph can be colored by only two colors such that vertices have the same colors in a set, this graph is called as bipartite graph. When modeling relationships between two different set of objects, bipartite graphs are used naturally. For instance, a graph of couples, with an edge between a man and a woman if the man or woman has a relationship between them, is a natural example of an affiliation network by using bipartite graphs. In this example, women are a set while men are another set.

An example given in the article[16] for the usage area of the bipartite graphs is railway optimization problem. This problem takes the schedule of trains and their stops as input, then finds a set of train stations as small as possible such that every train visits at least one of them. Actually, this problem is finding a minimum cost dominating set in the bipartite graph. In the graph, vertex sets are trains and stations while edges represent the visiting relationship between a train and a station.

2.3. Link Analysis and Link Analysis Algorithms

Barry and Linoff define the link analysis as “the process of building up networks of interconnected objects in order to explore pattern and trends. Link-analysis is based on a branch of mathematics called "graph theory"” [17].Link analysis is one process of knowledge discovery to identify, analyze, find out and visualize patterns in data. It has many usage areas like examining intelligence, computer security analysis, search engine optimization, ranking people, papers or objects, measuring influence, popularity or prestige and so on.

In the article entitled as “Link analysis” in Wikipedia[18], it is explained that there are four types of proposed link analysis solutions: heuristic-based solutions, template-based solutions, similarity-based solutions and statistical solutions. Heuristic-based solutions use decision rules created by expert knowledge and they work on structured data. Template-based solutions use Natural Language Processing to extract information from unstructured data. Similarity-based solutions apply a weighted scoring mechanism for objects to identify possible links. Statistical solutions use lexical analysis to identify potential links.

Kleinberg[7] says that links contain a hidden human judgment and this type of judgment is very useful for the formulation of the notion of authority. For example, page p has a link to page q. It can be commented as creator of page p has in some measure conferred authority on q. Link structure is used for defining notions of importance or standing, impact and influence with the same motivation as Kleinberg’s notion of authority.

Citations are links between the objects. The purpose of citations are grouped into four groups in the wiki page of University of British Colombia[19] and those are the followings:

- attribution of ideas or research for explaining a point, questioning or use of data, tools, definitions or methods
- providing proof that a thing is well-researched by providing review of literature or historical resources
- help disseminate useful knowledge with additional information or showing other ideas
- give formal credit for research to find funding.

Citation analysis is a kind of link analysis, as well. It is defined at the library web page of Illinois university as counting citations. It is a measure showing the number of citations done by other scholars for a publication or paper. This measure is very useful when a researcher doing a literature review in a field because she can first start reading papers that have high citation counts. The metrics are produced by citation analysis is used to evaluate individual impact of a scholar/researcher or a

department/university or a journal or a journal article or book etc.

There are two most popular link analysis algorithms: PageRank and HITS. The detailed information for them can be found just below.

2.3.1 PageRank

Brin and Page[8] present in detail the first version of Google, a prototype of a large scale search engine, in their paper. They say that Google is designed to crawl, index efficiently the web without any dependency to a query and produce search results that improved the results of the existing search systems. They explain how to crawl and index inside the system. The further information can be found at [8]. To obtain the results that have high precision, they use two features. First feature is calculating a quality ranking score for each web page and utilize the links for improving the search results. For this quality ranking score, they use the link structure of the Web independently from a query and call this scoring method as PageRank. They define the PageRank as follows:

"We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is the damping factor which can be set between 0 and 1. We usually set d to 0.85. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as in the (2.1) :

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)}\right), \quad \textbf{(2.1)}$$

The sum of the PageRanks of the web pages are one because PageRank algorithm build a probability distribution over web pages.

PageRank is modeled as a model of user behavior. They assume that there is a random surfer. The surfer starts the surf from a random page and he keeps clicking on the links without going back. However, the surfer gets bored and jumps into another

random page. By using this assumption, they define the PageRank of a page as the probability of the page being visited by the surfer. As for the d, damping factor, it is the probability that at each page the surfer gets bored and jump another page.

Franceschet[20] summarizes the notion of the PageRank as “a web Page is important if it is pointed to by other important pages”. He explains three factors for determining the PageRank score for a web page. First factor is the number of in-links . The second factor is outgoing links and the third one is the PageRank of the linking pages. Then he states the two main problems that lead to the birth of the version of the PageRank formula that contains the damping factor: presence of dangling nodes and buckets. In Figure 1, a dangling node and a bucket can be seen:

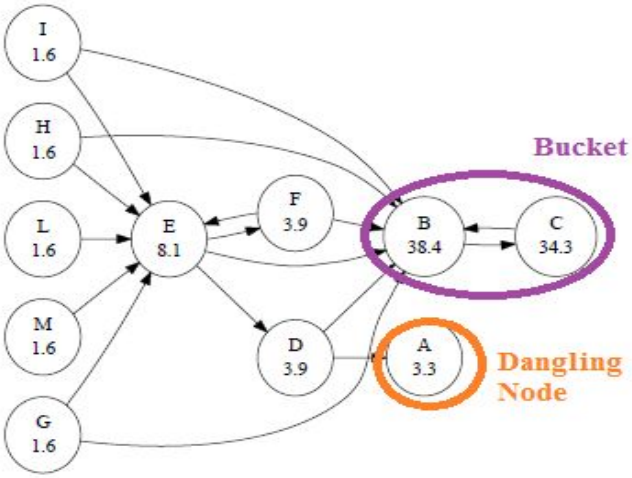


Figure 4. A PageRank Instance with Solution [20]

Dangling nodes mean the pages with no outgoing links. Dangling pages capture the random surfer indefinitely. Inside of the calculation, there is a matrix holding the probability distribution that the surfer moves from page i to page j by clicking one of the outgoing links of i. To handle the dangling node problem, the probability of dangling node is changed with 1/number of web pages value. Bucket means strongly connected components without outgoing links to the rest of the graph. The damping factor already

represents this condition. That's why, damping factor, d, is added to the calculation of PageRank score.

In Figure 4, page I, H, L, M and G have no incoming links, that's why, they take the lowest scores. The score of the page E is lower than page C even if page E takes many endorsements while C takes only one endorsement. The reason behind is C takes the links from the most important page B.

Franceschet[20] states that PageRank stands on the shoulders of many giants works starting from 1906 to 1998 like Markov Theory, Perron-Frobenius Theorem, Power method, Leontief's Econometric model, Seeley's, Katz's and Hubbel's Sociometric model, Wei's sport ranking model, Pinski and Narin's Bibliometric model and Kleinberg's HITS algorithm. He also remarks that those giants are used in many different areas including Web information retrieval, bibliometrics, sociometry and econometrics.

2.3.2. HITS

HITS is short for hyperlink-induced topic search and it is a web page ranking method proposed by Kleinberg. It is used as a part of Ask search engine[21]. Kleinberg's HITS algorithm works in the same nature with Brin & Page's PageRank algorithm. These two algorithms use the link structure of the web graph to discover the relevance of the web pages. Although PageRank works on the whole web graph, HITS works on a subset of the web graph according to a query. The following idea is at the heart of HITS algorithm: A good hub increases the weight of pages that it pointed while a good authority increases the weight of pages that point because there is a circular relationship in the nature of authoritativeness and hubness. This relationship can be seen in the updating of authority and hub weights described below. [22]

Kleinberg states that there is a "mutually reinforcing" relationship between hubs and authorities in the network. This relationship implies that a good authority is a page pointed by many good hubs and a good hub is a page that points to many good authority

pages. Also, he defines the relationship as the hub weight is the sum of the authority nodes pointed by this hub, and the authority weight is the sum of the hub nodes that point to this authority.[7]

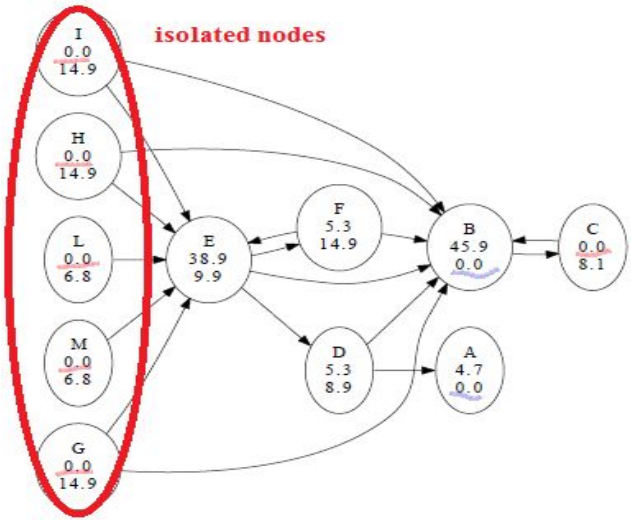


Figure 5. A HITS Instance with Solution [20]

In the Figure 5, page I, H, L, M, and G pages are not important and authoritative but they are the best hubs in the network because they point the good authority pages E and B. Page B is both important and authoritative while it is not a good hub while page C is important but not authoritative. Hub score of page B is zero since authority score of the page C is zero. As can be seen from the graph, isolated nodes have zero authority score.

Kleinberg introduces the HITS algorithm in his paper[7]. Devi et al.[23] restate the algorithm and the steps following are restated their HITS pseudocode.

Step 1: Determine a base set S.

- Take most related pages returned by a search engine for a given query and call them as root set, R.
- Initialize S with R.

Step 2: Expand S by using links of the root set

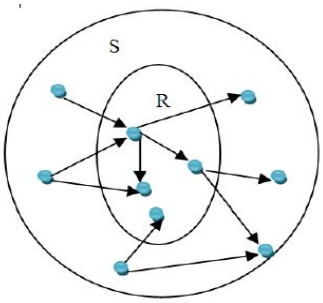


Figure 6. Expanding Set S [23]

- Add the pages referenced by a page resides in the root set R to the set S.
- Add the pages that reference to a page resides in R.
- For each node page p initialize the authority weight of p, a(p), and the hub weight of p, h(p), to 1.

Step 3: Update authority and hub weights

If the n pages are obtained for a query by the search engine, then HITS algorithm creates n by n adjacency matrix and matrix[i,j] element is regarded as 0 if there is no links from i to j and 1 otherwise. Then it continues with the update of weights.

- Update authority and hub weights for each node in the S by using the following formula, (2.2).

$$a_i^{(t+1)} = \sum_{j:j\to i} h_j^{(t)}$$
$$h_i^{(t+1)} = \sum_{j:i\to j} a_j^{(t+1)} \quad \textbf{(2.2)}$$

a_i represents the authority score of i^{th} page and similarly h_i represents hub score of page i.

Step 4: Normalize the scores

Normalize authority scores by dividing by the square root of the sum of squares of all the authority scores. In similar to authority score normalization, normalize hub scores.

CHAPTER 3

REVIEW OF LITERATURE

Many works can be found in the literature on the link analysis to ranking papers, pages, people or objects, to measure prestige or importance and expert finding issue. In this section, most relevant works towards the problem defined in Chapter1 is presented.

Pinski and Narin[24] developed a self-consistent influence weighting methodology for scientific journals. An eigenvalue problem is identified by creating cross citing matrix between journals or aggregates of journals. This formulation leads to a size independent influence weight for each journal.Also, they define influence per publication and the total influence measure. Moreover, they use the hierarchical influence diagrams to visualize journal relationships. Actually they use 103 physics journals as data set.

Palacios-Huerta and Volij[25] research the measuring influence based on data contained in the communication network between scholarly publications, patents, web pages, judicial decisions. They propose to use the data obtained from the network to address measure of influence like prestige, diffusion of knowledge, the productivity of academicians, the ranking algorithms employed in search engines in the web. They apply an axiomatic methodology for handling ranking problem and present an axiomatic model for intellectual influence. Then they find a unique ranking method of journals can be characterized by five axioms like anonymity, invariance to citation intensity, weak homogeneity, weak consistency and in-variance to splitting of journals. They call the method as the Invariant method and the method is proposed first by Pinski and Narin[24]. Palacios-Huerta and Volij say this method is different from the other methods like the Counting Method, the Modified Counting Method and the Liebowitz-Palmer method because of the applied axiomatic approach. Actually, this invariant method is at the core of PageRank that is used by Google to rank the web pages.

Then, one of the most popular link analysis algorithms, PageRank, is used to calculate random walks on a graph. Guo and Barbosa[26] present an approach guided by natural notion of semantic similarity for entity linking.They build an entity graph, and represent each candidate entity with a stationary probability distribution . This probability distribution is obtained with a random walk on that graph. These random walks are produced by a personalized PageRank algorithm. The algorithm produce a score between connected nodes. The scores or probability values can be regarded as relatedness between each entity and target entity. They called these probability distributions as semantic signatures of the entities. They uses MSNBC, AQUAINT and ACE2004 well-known public benchmarks to compare their systems to the other entity linking systems like PriorProb, Local, Cucerzan, M&W, AIDA, GLOW and RI. Then, they show the superiority of their methods REL-RW (Robust Entity Linking with Random Walks) to others mentioned just above.

Shahaf and Guestrin[27] examines some methods for automatically create a coherent chain to link two news articles and also they provide an algorithm with theoretical guarantees for this linking work. The authors create a bipartite graph used to calculate influence of a document_i on document_j with respect to a word w like in Figure 7. The square ones show documents while circular ones show keywords. The authors explain that the edge weights in the graph indicate the strength of the correlation between a document and a word and the weights can be interpreted as random walk probabilities. Their intuition is that if the two documents are highly connected and word w plays an important role in this link, the influence between document_i and document_j can be calculated via word w. First, they calculates part of the time the walker spends on each node as stationary distribution for random walks starting from document_i. Then, they investigate the effect of the word w on these walks by turning it into a sink node. Next, they calculates the stationary distribution one more time and find the difference between these two distribution. As conclusion, they define the influence as this difference.

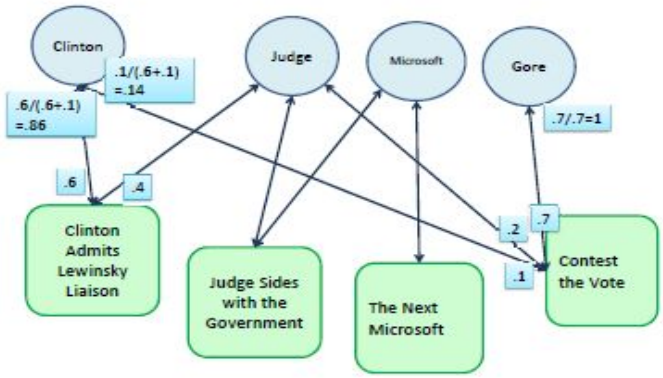


Figure 7. Keyword - Title Graph [27]

The thesis is not directly interested in linking two papers, however the idea that proposed by the authors are interesting for the thesis: creating a bipartite graph between documents and keywords, interpreting the edge weights as random walk probabilities and calculating these probabilities by using PageRank algorithm.

After gauging influence, people develop formal models to put one step forward this influence by searching experts. Balog et al.[28] present two general strategies formalized by using generative probabilistic models for expert searching in a set of given documents or papers. First strategy is directly modeling an expert's knowledge based on the papers. On the other hand, the second strategy is locating papers on topics, then finding the associated experts to these topics. They use 2005 edition of the TREC test collection to evaluate and compare the models in the two strategies. TREC is short for The Text REtrieval Conference[29]. They find that the second strategy produces better results and better time response from first strategy.

Besides the formal models, a new model category is constructed : Author - Topic models. Rosen-Zvi et al.[30] introduce an author-topic model for documents that extends Latent Dirichlet Allocation to add authorship information to the documents. Actually, it is a simple probabilistic model to find out the relationship between authors, papers, topics and words.There is a multinomial distribution over topics for each author and there is another multinomial distribution over words for each topic. If a document has multi-author, it is modeled as a distribution over the topics. Otherwise, a document

has only one author, it is modeled as a distribution over words instead of topics. They show topics recovered by the model they propose.

Then, this expert finding task is specialized for academia. There are a few systems for academic search like Microsoft Academic search,Google scholar, Rexa academic search engine, CiteSeerX digital library and search engine in computer and information science and ArnetMiner.

Tang et al.[31] introduce in detail an academic search system called ArnetMiner that extracts and mines academic social networks. This system extracts and retrieves profiles of researchers from the web, retrieves and integrates the publication data from digital libraries like ACM, DBLP, CiteSeerX and so on. Then it models the academic network that it creates. After, it provides expertise search, author interest finding, academic suggestion like paper suggestion and citation suggestion and people association search based on the modeling results. Also, Tang et al. propose a unified tagging approach to researcher profile extraction, a framework for name disambiguation, three generative models called as ACT (Author-Conference-Topic) for modeling topical aspects of papers, authors and publication venues.

Tang et al.[32] investigate and formalize the extraction of an academic researcher social network. They find and extract profile information of researchers and then they combine the information via the semantic-based profiling from the web. Their paper is the first paper gives the formalization for this extraction work. Tang et al.‘s system first obtained the related documents for a researcher from web by using a classifier. Then they extract basic information like research interest, affiliation, position, person photo, contact information and educational history of the researcher by using CRF, Conditional Random Fields. They use CRF as tagging model. For illustrate, for <image> token, two tags are assigned : Photo and Email since an e-mail is possibly shown as an image. After, they extract publication information of the researchers from DBLP bibliographic data set. To integrate extracted personal and publication information of the researcher, they propose a constrained based probabilistic model using Hidden Markov Random Fields to name disambiguation on the publication dataset. Those constraints are CoOrg, CoAuthor, Citation, CoEmail, FeedBack and

$\tau - CoAuthor$ for two papers. CoOrg means principal authors of two papers are from the same organization. CoAuthor means two papers have a secondary author with the same name. Citation means a paper cites the another. CoEmail means principal authors of the two papers have same e-mail address. Feedback means showing user interaction. $\tau - CoAuthor$ means there are one common author in τ extension.They also put some experimental results showing that their system produces better results from the systems using classification. Also, they are first to make research profiling in a unified approach and applying a constrained-based probabilistic model for name disambiguation. They also apply these methods for expert finding and discuss the results by comparing some measures like P@5, P@10, R-prec, MAP, bpre and MRR.

Next, multiview graphs are introduced.Wang et al.[33] propose a framework for generating pictorial and temporal story lines with the idea they can give a more enjoyable summaries for the reader for a given topic from text and images collected from the internet. Their system takes a topic and documents containing images and text related to this topic as input. Then they construct a multi-view graph object where each vertex is an image associated with a text describing it. The graph includes two types of edges: directed and undirected edges. While undirected edges represent a certain level of similarity between the vertices or objects, directed ones represent a certain type of pairwise temporal relationship. After construction of the multi-view graph, their system chooses a set of representative objects by using minimum-weight dominating set approximation algorithm. After obtaining this dominating set, they apply Steiner tree algorithm to this set to create a story line capturing the temporal and structural information when connecting this set. They collect manually 355 images and text from Flickr, ABC News, Reuters, AOL News and National Geographic and use them as generic data set.

Zhang et al.[34] propose a unified framework to find out dominating patents on a multi-view patent graph that contains both patent content and patent citations. Their proposed framework produces three main linkage of patents: PatentLine, PatentTrace and PatentLink. PatentLine shows the technology evaluation tree of a specific field. PatentTrace trace a given patent document to its roots. As for PatentLink, it explores the

possibly relations between two patent document. They use 16,518 US granted patents in physics from the State Intellectual Property Office of the P.R.C. Make PatentLine realize, after construction of the multi-view graph that each vertex is a patent document, they discover dominating patents and apply Steiner tree algorithm to create the patent line. There are two types of edges in the multi-view object graph: undirected and directed edges.Undirected edges have weights showing the content similarity of connected vertices. The directed edges shows the citation relationship between two vertices.

In the context of this work, multi-view object graphs can be used to look at the CN with different point of view beyond only citation relationship and to obtain the dominating papers in the network to strengthen the results coming from AKG.

CHAPTER 4

EXPERIMENTAL WORK

In this work, there are two principal components: a citation network and author-keyword bipartite graph. The citation network constructed out of bibliographic data indicate what publications are the most authoritative ones through the use of link analysis algorithms like HITS. Similarly, alternative author-keyword bipartite graphs generated from bibliographic data should have potential to tell us who is expert in a subject area and how a research topic popularity changes over time. The combination of the results from citation graph analysis and link analysis done on the keyword- author bipartite graphs should provide new insights.

As for the experimental setup, we constructed an environment that reveals two types of affiliation: First one is the relationship between authors and the keywords (subject descriptors) part of the papers written by these authors of publications because subject-matters are adapted as keywords. Second one is the citation relationship between publications and more detailed information is given in “Construction of Citation Network ” sub-part. By nature of author and keyword relationship, there is a bipartite relationship. Thus, it is appropriate that represent their relations with a bipartite graph whose a set of vertices includes author names when the second set includes keyword labels. The edges between these sets are bi-directed. The author-keyword bipartite graph constructed can be seen in Figure 8.

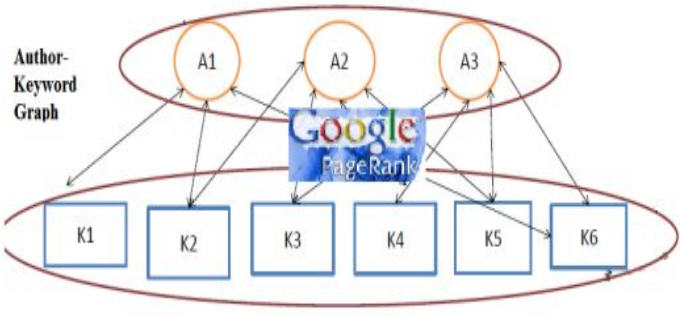


Figure 8. Author - Keyword Network

In Figure 8, {A1, A2, A3} shown in circles is the set for representing name of the authors. Similarly, {K1, K2, K3, K4, K5, K6} shown in rectangles is the set for representing keywords.

This graph can answer the following questions:

1- On which subject-matters does an author write?

2- Who writes about K subject-matter?

3- What is the influence of a keyword on relating two authors ?

4- What is the influence of an author on relating two keywords?

Running a random walk on this author-keyword bipartite graph can say influence of a keyword on relating two authors and influence of an author on relating two keywords. Shahaf and Guestrin[27] make a work that shows the applicability of running random walks on a bipartite graph.The details of the work is in the just following paragraph.

The authors mentioned just above examines some methods for automatically create a coherent chain to link two news articles and also they provide an algorithm with theoretical guarantees for this linking work. The authors create a bipartite graph used to calculate influence of a document_i on document_j with respect to a word w like in Figure 9. The square ones show documents while circular ones show keywords. The authors explain that the edge weights in the graph indicate the strength of the correlation between a document and a word and the weights can be interpreted as random walk probabilities. Their intuition is that if the two documents are highly connected and word w plays an important role in this link, the influence between document_i and document_j can be calculated via word w. First, they calculates part of the time the walker spends on each node as stationary distribution for random walks starting from document_i. Then, they investigate the effect of the word w on these walks by turning it into a sink node by cutting all out-links. Next, they calculates the stationary distribution one more time and find the difference between these two distribution. As conclusion, they define the influence as this difference. For illustration, they can measure the influence of *"Judge"* keyword when relating two news title, *Clinton Admits Lewinsky Liaison* and *Judge Sides with the Government* by using the approach just mentioned.

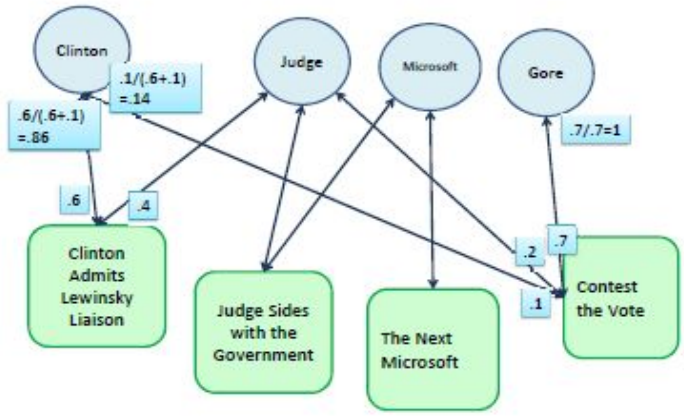


Figure 9. Keyword - Title Graph [27]

On our author-keyword bipartite graph, the common working subject-matter of two authors can be found with a straightforward approach instead of applying link analysis algorithm like PageRank. In the approach, first two separate interest sets for these authors are prepared and then the intersection of these sets are found. These intersection set gives the common working area for specified authors in the network. However, the weight or importance of each keyword in the intersection set is exactly same in this approach. That is, in this approach it is not possible to see the influence of a keyword relating two authors. The approach used by Shahaf and Guestrin gives the influence but its cost is too high to carry out this experiment.

Another usage of running random-walk on a graph is to see the importance of the nodes from the link structure between the nodes by using stationary distribution. As a random-walk, PageRank algorithm is applied in the work because Gleich[40] says that the mathematics of the PageRank are general and can be applied to any graph in any domain. Applicability of the PageRank is directly related to stationary distribution at nodes of a graph. Sarkar[41] states that if a graph is irreducible and aperiodic, then a stationary distribution always exists. Even if a graph does not hold for these two conditions, irreducibility and aperiodicity, the graph can be modified slightly so that the graph has these two properties.

Hence, the system contains two main parts: Author-Keyword graph and Citation Graph respectively. In Figure 10, the overview of the system can be seen.

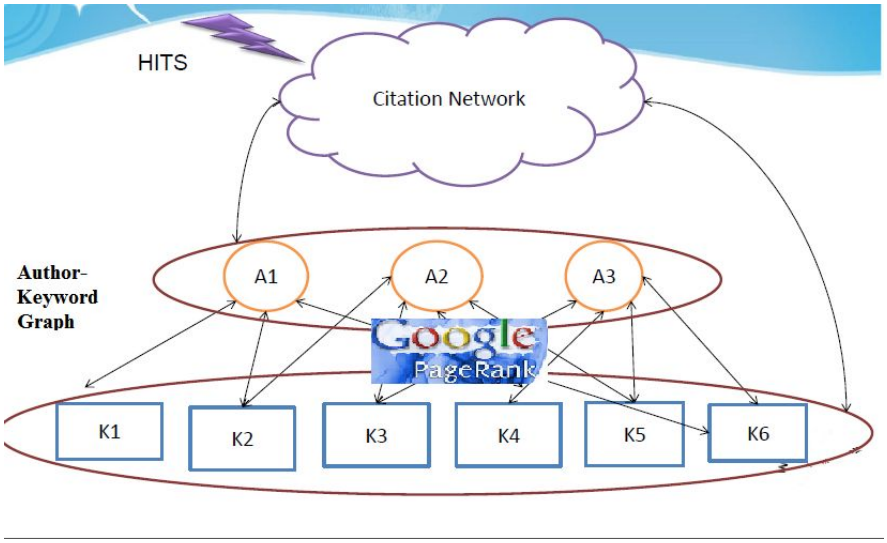


Figure 10. Overview of the Proposed System

As for the citation network or citation graph, it is constructed from bibliographic data of publications. It shows the citation relationships between these publications. By the way, these publications are exactly the same set, out of which Author- Keyword Bipartite Graph is constructed.

To make it clearly, it is time to zoom in to the proposed system. In Figure 11, the detailed version of the system can be seen:

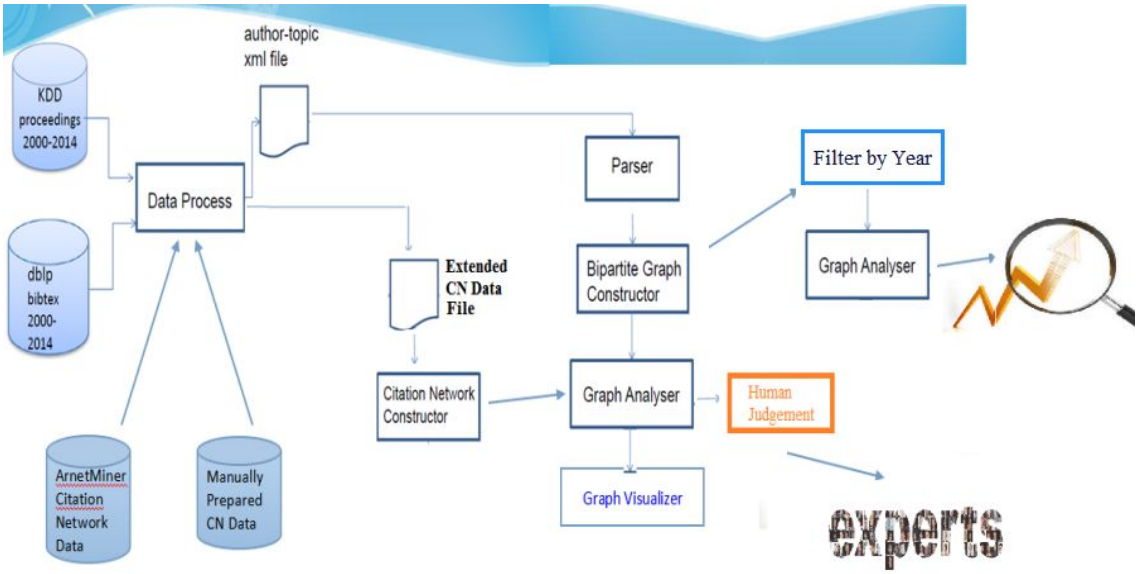


Figure 11. A Detailed View of the Proposed System

The obtained data from KDD proceedings and dblp's KDD bibliographic data is taken by the data process function. After that, two files are produced: one for creating Author-Keyword graph and the other for the construction of the citation network. After the construction operation of these two, graph analyzer applies the ranking algorithms PageRank and HITS respectively. Then the analyzer produces both a curve showing how a research topic popularity changes over time for a selected small set of keywords and PageRank scores for author names and keywords to deduce who are subject-matter experts. Last, the Graph Visualizer visualizes both of the graphs on demand.

4.1. Data Collection

It is clear that data collection is the first step for obtaining knowledge or useful information after the system design. According to the design, two kinds of sources (KDD proceedings and bibliographic data of KDD proceedings) are needed to construct the Author-Keyword bipartite graph. As for the citation network construction, ArnetMiner citation network dataset and some semi-automatically prepared data that is not covered by the dataset are used.

4.1.1. KDD Proceedings

As part of this thesis, the papers published in the KDD proceedings from 2000 to 2014 are used for constructing the graphs. The reason behind the selection of KDD proceedings is that it is one of the top conferences in computer science/engineering with an acceptance rate of 18% on average. Thus, it can be plausible to select the experts on the Data Mining and Knowledge Discovery domain from the authors who publishes in these proceedings. The second reason for choosing KDD proceedings is, it is easy to access to proceedings via ACM membership.

All published KDD proceedings that are between 2000 and 2014 are downloaded as pdf documents and they are stored in a separate folder according to their years. To sum up, there are 1982 papers whose format is pdf to work with.

4.1.2. Bibliographic KDD Records

In the thesis, it is aimed to deduce subject-matter experts by using bibliographic data. Dblp computer science bibliography is chosen as the bibliographic data source. The acronym for the dblp computer science bibliography is "Digital Bibliography & Library Project". Dblp is a service that provides free bibliographic information on substantial computer science journals and proceedings those published by ACM, Elsevier, Emerald, Springer, Wiley, MIT Press and many more. It is a joint service of the University of the Trier[35] and Schloss Dagstuhl[36]. There are some statistics related to the dblp computer science bibliography given below:

The number of publications is 2,982,835.

The number of authors is 1,576,902.

The number of conferences is 4,295.

The number of journals is 1,414.

The data served by dblp can be obtained either online as the XML, JSON, RDF and BibTeX formats, or offline as the XML format. It is preferred to use offline XML file for the nature of the thesis. This file contains some attributes for each publication. The format of a publication is similar to the following:

```
<inproceedings mdate=" " key=" ">
  <author>    </author>
  <author>    </author>
  <title>    </title>
  <pages>    </pages>
  <year>    </year>
  <booktitle>KDD</booktitle>
  <crossref></crossref>
  <url></url>
</inproceedings>
```

Inside this XML file, there are 1689 publications. This file will be used on the construction of Author-Keyword graph after some data preparation.

4.1.3. ArnetMiner Citation Network Dataset

The data set contains citation data from DBLP, ACM and some other source . It is a constructed for only research purposes. It has seven versions. Its last version, DBLP-Citation-network V7, is used for this thesis. This version includes 2,224,021 articles/papers and 4,354,534 citation relationship and it is released by May 25, 2014.

DBLP-Citation-network V7 is a special formatted file that has a separate structured block for each paper. Each line of a block represents an attribute of a paper. The structured block for a paper can be seen below:

#* ----> paper title
#@ ----> author(s)
#t ----> year
#c ----> publication venue
#index00 ----> the id number of this paper
#% ----> the id number of the paper to which this paper gives reference
#! ----> the abstract text for this paper

In this block, only the line starting with #% can repeat because a paper can give reference to many other articles.

4.2. Data Preparation Process

In the data collection part, how to retrieve and store of a KDD publication is explained. However, in this part data preparation for the construction of Citation Graph and Author-Keyword Graph and some parts of GraphAnalyser component are stated.

First of all, dblp computer science bibliography XML file is a very large file. Its size is nearly 1,5GB . It is not opened by a text editor like notepad directly to read. To solve this problem, this large file is separated into small chunks by using a free software tool. Then, each chunk is opened and the publications whose year is not between 2000

and 2014 are detected and removed manually. Next, a new XML file named as **“dblp_mixed.xml”** is created, the following lines are written inside of it.

```
<?xml version="1.0"?>
<start>
</start>
```

After that, the publications reside in other chunks are copied and pasted into the new XML file. However, this data are not sufficient for our purpose as the publications do not contain keywords, categories, subject descriptors and general terms parts. Thus, the parts should be obtained, then they should be added to the XML file in appropriate publications records.

To do this, it is needed to parse the content of these papers whose format is pdf in order to get some specific attributes such as categories and subject descriptors, general terms and keywords for the articles if the papers have those parts. One of the main assumptions in this thesis is the papers that at least have **keywords** part can be used because in the thesis, keywords represent subject-matters. That is why, this assumption is directly related to the purpose of the thesis: to find out subject-matter experts on Data Mining and Knowledge Discovery domain. A free open source Java software tool whose title is “pdfBox”[37] is used to obtain the text contents. By using this tool, the only first pages of these articles are obtained and saved in a separate text file named as the same with that of pdf paper; because the needed parts like keywords are written on the first page of each paper.

After the texts are obtained, the next step is to parse the text files one by one and obtain keywords as a must, categories and subject descriptors and general terms are optional. After examining the whole set of papers, it is seen that there is no unified structure for all papers . For example, some papers have only keywords part, while others can have keywords, general terms and subject descriptors, or a combination of the keywords, general terms and keywords. For this reason, MetaExtractor and MetaParser components inside of the Data Process part handle each case that is mentioned. MetaExtractor extracts keywords, general terms and categories and subject descriptors as a concatenated string. As for the MetaParser, it parses those strings and returns back the list of them separately. Hence, these lists are added to the

dblp_mixed.xml file.

In summary, for the construction of Author-Keyword Graph, the keyword information is a must and the publications must contain this data. With this purpose, keywords,categories and subject descriptors and general terms parts are obtained by some piece of helper codes, then they are inserted into the XML file like in the following format:

```
<inproceedings mdate="2012-12-12" key="conf/kdd/YiS00">
    <author>  </author>
      <author>  </author>
    <title> </title>
    <descriptor> </descriptor>
    <term></term>
    <keyword>  </keyword>
    <keyword>  </keyword>
    <keyword>  </keyword>
    <pages>  </pages>
    <year>  </year>
    <booktitle>KDD</booktitle>
    <crossref>  </crossref>
    <url>  </url>
</inproceedings>
```

After this insertion operation, this XML file is ready for constructing Author-Keyword Graph.

Lastly, DBLP-Citation-network V7 is opened with a free text editor and the blocks whose publication venue are not equal to KDD are removed from the file. Actually, the following publication venues are detected and removed manually.

bigMine, BIOKDD, WISDOM, SNAKDD, BigMine, WebKDD/SNA-KDD,
ADKDD@KDD, KDD, Workshop, MDU/KDD, ADMM, WEBKDD,
PinKDD Software Mining, Revised Papers from MDM/KDD,
Revised Papers from AKDD/RDMCD

They are removed because the papers belonging to these conferences are not inside of KDD proceedings downloaded. After this removal operation, the blocks in different files are merged and saved into another text file named as “arnetminer_updated.txt”.

After the examination of the **arnetminer_updated.txt** file, it is clearly seen that there are only three blocks belonging to 2012, there is no block for 2013 and 2014. To determine the criticality of this missing part, the number of the papers published between 2012 and 2014 are counted. The counts are as the following:

The number of the papers published in 2012 --> 200 papers

The number of the papers published in 2013 --> 185 papers

The number of the papers published in 2014 --> 193 papers

The sum of those papers is 578 papers. It is nearly one third of the whole papers in KDD proceedings. Hence, it is determined to add new blocks for those lost article references. These new blocks have the same structure with the described above. The only difference between them is the newly added blocks has **#indexAK00** instead of #index00. The new index keyword is assigned to avoid collision of the index id because the ids must be unique for each paper. After those additions, the final version of the arnetminer_updated.txt contains 2287 blocks. In other words, the Citation Graph contains 2287 papers.

4.3. Author - Keyword Graph Construction

Preparation of the XML data for Author-Keyword Graph was explained in the Data Preparation Process part. In this part, the methodology and tools used for graph construction and analysis are presented.

Many open source Java graph libraries are examined to construct a bipartite graph. Among them, JUNG is decided to be used. It is short for the Java Network/Graph Framework. It is a free open source software that provides a powerful graph framework. Its architecture supports various representations of entities and the relations between them. For example, it supports multi-modal graphs, hypergraphs, graphs with multiple edges or single edges, directed and undirected graphs. Besides, it permits to annotate the entities such as vertices and edges. Moreover, it includes various built-in graph algorithms from graph theory like shortest path and calculation of network distances,

from data mining like clustering and from social network analysis like centrality, PageRank, HITS, etc. Also, it provides a filtering mechanism so that the users can obtain a specific portion of the graph. For example, if the user annotated edges in the graph data then he can put a filter to filter out the edges whose weight is more than 0,7. Also, it is possible to extend filtering cases. The filtering mechanism that is used in this thesis is explained later. Last but not least, it provides a visualization framework so that users can easily visualize their network data.

There are two ways to create a bipartite graph in JUNG. The first one is to create all edges and vertices one by one. The second way is to prepare a .txt file as in the following format and to delegate the creation of the bipartite graph to the hands of JUNG.

vertexName1 vertexName2

vertexName3 vertexName4

vertexName vertexName5

vertexName6 vertexName7

vertexName8 vertexName2

In here, pre-built *BipartiteGraphReader* component reads this file and labels the first column vertices as a one set and labels the second column vertices as another. Also, it creates an edge between two vertices written on the same row of the text file. For example, vertexName1, vertexname3, vertexname, vertexname6 and vertexname8 are labeled as PART_A while vertexname2, vertexname4, vertexname5 and vertexname7 are labeled as PART_B. Also, an edge is created between vertexName1 and vertexName2. The edges are created similarly. It should be considered not to write the same vertices in both parts.

In the thesis, the second way explained just above is preferred. Therefore, a graph data text file is needed that contains author names on the first column and keyword(s) on the second column and those columns must be separated with spaces or a tabs. In the thesis, a text file is prepared for Author-Keyword Graph construction. The format can be found below:

authorName keyword1 keyword2 .. keywordN

authorName keyword1 keyword2 .. keywordN

authorName keyword1 keyword2 .. keywordN

In this format, all keywords belonging to the same article are written on the same row as separately. That is, each keyword on the row represents a new vertex if it is not created before. Even if time information is not written in this file, it is stored in another text file because the time information is needed to create some sub author-keyword bipartite graphs according to the year. These are explained in 4.8 Determining Experts and Trends part.

On the other hand, an Author-Keyword Graph that have 17448 vertices and 45704 edges is created if individuals format is used for the construction. On the construction of the graph, each vertex is labeled with a type label as author or keyword and with a label as author name or keyword. In addition to this, each edge is weighted as 1.0/ number of total edges between the same two vertices.

4.4. Citation Graph Construction

The input data preparation for citation network is explained in the Data Preparation part. As a reminder, the nodes of the CN is publications, edges indicate the citation affiliation between these publications and the publications are exactly the same publications that contain the keywords and authors used in the creation of AKG. In this part, the construction of a citation network for the **arnetminer_updated.txt** is explained.

First of all, the *arnetminer_updated.txt* file is read and stored article and citation information stored on the memory for each publication record in the file. Then by using this information, vertices and edge relationships are created. In JUNG, there is no pre-built implementation for creation and manipulation of a citation network, but there are directed sparse graph, directed sparse edge and directed sparse vertex structures. By using these structures and the information coming from the file prepared before, the

Citation Graph is created. Its number of vertices is 2287 and the number of edges is 3129.

4.5. Applying PageRank on Author-Keyword Graph and Subgraphs

The PageRank algorithm is explained in detail in Chapter 2. Therefore, in this part, only application of the PageRank to the Author-Keyword Graphs is discussed. After the creation of the Author-Keyword Graph, PageRank is aimed to apply to obtain edge weights between keyword and author labels. In addition to this, for each year between 2000 and 2014 a sub author-keyword graph (fifteen sub-graphs in total) is created to track how a keyword popularity is changed over time.

Although JUNG has pre-built PageRank algorithm implementation, it couldn't be used as it needs a directed graph to work. However, the Author-Keyword Graph is a KPartiteGraph. That's why, PageRank of JUNG is modified to reflect that the graph used in the algorithm is KPartiteGraph instead of Directed Graph. Additionally, Author-Keyword Graph edges are made as directed edges starting from keyword(s) to author vertex. Then, PageRank is applied and the rank scores is added to all vertices on the Author-Keyword Graph by "pageRankScore" label. Bias value is chosen as 0,85 because it is default value suggested value by Brin&Page[8].

The elaboration of the page rank results for author-keyword graph is utilized in Section 4.8 Determining Experts and Research Trends.

4.6. Applying HITS on the Citation Graph

HITS is explained in a detailed manner in Chapter 2 Background, and preparation of citation network data is explained in this chapter in Section 4.4- CN construction. Thus, in this part, the application of HITS to Citation Network and the results are presented.

After the construction of the Citation Network, built-in HITS algorithm is applied with a specific parameter for implying that the authority rankings is produced. In fact, it is enough to adjust the second parameter as true when HITS constructor is created. The authority ranks obtained after the run of the HITS algorithm on the Citation Network is given in the Table 6 resides in Appendix A.

After the construction of the Citation Network, built-in HITS algorithm is applied once more to obtain hub rank scores. Actually, it is enough to adjust the second parameter as false when HITS constructor is created. After this, the hub ranks for the Citation Network is given in the Table 7 in the Appendix A.

After applying HITS onto the Citation Network, "HitsAuthorityScore", "HitsHubScore" and "title" labels are added to each vertices in the Citation Network. The vertices/papers that have higher authority scores are regarded as the most influential or dominating papers whereas the vertices/papers that have the higher hub scores are regarded as the most dependable papers to give references to the most influential ones.

4.7. Graph Visualization

The graph visualization is useful to evaluate the PageRank and HITS results. Since the number of vertices is too much, it's difficult to track all the vertices. In order to support this evaluation, filtering mechanism is used to remove seemingly less relevant. The first filter is put onto Author-Keyword Graph. It filters out the vertices whose PageRank is greater than a specified threshold, then the Author-Keyword Graph is visualized by using JUNG visualization framework. In the generated visualizations, vertex size is dependent on the PageRank and HITS scores.

In the Figure 16 resides in Appendix B , the blue colored vertices show the authors and the red ones show the keywords on the Author-Keyword Graph when the lower limit is 0.0005478. Also, it can be seen from the graph, Christos Faloutsos writes at least one paper about clustering, social network, anomaly detection and text mining In Figure 17 in the Appendix C, the most twenty influential papers are shown in the

Citation Network. The direction of the edges are not considered. It only gives the related papers. However, In Figure 18 in the Appendix C, the most influential papers are shown with the reference relationship.

4.8. Determining Experts and Research Trends

In this part, the experimental results are discussed in detail. In the Table 1, the first-20 ranked author names and keywords from the author-keyword bipartite graph can be seen.

Table 1. Top 20 PageRank Scores

Rank	Vertex Label	PageRank Scores
1	data-mining	0.0012732709430581897
2	clustering	0.0010301057575731493
3	classification	8.456338242459594E-4
4	social-networks	7.813305361538582E-4
5	text-mining	6.523049411130134E-4
6	machine-learning	5.938739471634679E-4
7	jieping-ye	5.71188772000735E-4
8	christos-faloutsos	5.696042577902038E-4
9	philip-s.-yu	5.54509969920494E-4
10	anomaly-detection	5.478728197270317E-4
11	jiawei-han	5.318816796897532E-4
12	graph-mining	5.124258275749302E-4
13	ravi-kumar	4.277199847856868E-4
14	collaborative-filtering	4.217804089909378E-4
15	time-series	4.205290476874286E-4
16	information-extraction	4.129103125987376E-4
17	feature-selection	4.093905808395928E-4
18	social-network	3.9058382905162895E-4
19	recommender-systems	3.8706525408146693E-4
20	hui-xiong	3.697588560082048E-4

As seen from the Table 1, *data mining* keyword ranked as first in the network. Also, *clustering*, *classification*, *social networks*, *text mining* and *machine learning* have high page rank scores, too. What is common about all these keywords is their generality.

According to the observations done after running random walks on the bipartite graph, the keywords having high page rank scores implies generality while the author names having high page rank implies those authors works on general keywords. In Table 2, the author names in Top 50 PageRank scored nodes can be seen. These author profiles are checked and many of them write a book on subjects in KDD. Thus, writing a book can be regarded as working on general subject-matters.

Table 2. Ranked Author Names Listed in Top-50

Rank	**Vertex Label**	**Rank**	**Vertex Label**
7	jieping-ye	**28**	deepak-agarwal
8	christos-faloutsos	**30**	chengxiang-zhai
9	philip-s.-yu	**31**	srinivasan-parthasarathy
11	jiawei-han	**32**	thorsten-joachims
13	ravi-kumar	**39**	heikki-mannila
20	hui-xiong	**40**	jian-pei
24	bing-liu-0001	**45**	naren-ramakrishnan
25	huan-liu	**49**	vipin-kumar
26	tao-li	**50**	evimaria-terzi

The following authors can be regarded as examples of possible experts.

- **Jiawei Han** and **Jian Pei** writes *Data Mining: Concepts and Techniques (Han, Kamber, Pei, Morgan Kaufman, 2011)* on **data mining** whose rank is 1.

- **Christos Faloutsos** writes *Graph Mining: Laws, Tools, and Case Studies Synthesis Lectures on Data Mining and Knowledge Discovery (Faloutsos, Chakrabarti, Morgan Claypool, 2012)* on **graph mining** whose rank is 12.

- **Philip S. Yu** writes *Domain Driven Data Mining (Cao, Zhang, Zhao, Springer, 2012)* a book on **data mining** whose rank is 1.

- **Hui Xiong** writes *Clustering and Information Retrieval(Wu, Xiong, Shekhar,Kluwer Academic Publishers, 2003)* on **clustering** whose rank is 2.

- **Bing Liu** writes *Web Data Mining: exploring hyperlinks, contents, and usage data (Liu, Morgan & Claypool Publishers, 2012)* on **web data mining** whose rank is 36.

- **Huan Liu** writes *Social Media Mining: An Introduction(Zafarani, Abbasi, Liu,Cambridge University Press,2014)* on **social media** whose rank is 27.

- **Thorsten Joachims** writes *Learning to Classify Text using Support Vector Machines, (Joachims, Kluwer/Springer, 2002)* on **support vector machines** whose rank is 33.

- **Heikki Mannila** writes *Principles of Data Mining(Hand, Mannila, Smyth, MIT Press, 2001)* on **data mining** whose rank is 1.

- **Vipin Kumar** writes *Introduction to Data Mining(Tan, Steinbach, Kumar, Addison-Wesley, 2005)* on **data mining** whose rank is 1.

- **Evimaria Terzi** writes *Privacy and Online Social Networks(Zheleva,Terzi, Getoor, Morgan and Claypool Publishers,2012)* on **privacy** whose rank is 29.

The results obtained support the initial idea. That is, application of random-walks on the keyword-author network produce meaningful results. The authors having high PageRank scores are good nominees to be experts that have good level knowledge of various subtopics of a subject-matter because many of them write at least one book on subject-matters in KDD (keywords in the graph) that have high PageRank scores. Since writing a book needs specialties, they are good nominees to become subject-matter experts.

The observations obtained from this graph lead to some questions. The first one is that whether the authors working on specific subject-matters can be distinguished with such an experimental setup. The answer is yes because the PageRank implies generality and importance and the PageRank scores for specific keywords are low. After the keywords that have too low PageRank are chosen, their related authors are obtained from author-keyword bipartite graph. Therefore, the authors found are possibly authors working on specific subtopics or subject-matters. The second question is that whether experts working on too specific subject-matters can be distinguished. The answer is yes because too specific keywords are used as subject descriptors by a handful of authors.

As an evidence, some specific keywords shown in the Table 3 are examined and it is seen that they are related to only one author or co-author of the same document. Then, those can be regarded as experts for that subtopics.

Table 3. Some Selected Too Specific Keywords and Their Related Authors

Keyword	**Author(s)**
distance functions	Charu C. Aggarwal
speaker recognition	Charu C. Aggarwal
interactive marketing	Usama M. Fayyad
latent class models	Mark Sandler
selective sampling	Hwanjo Yu
junction trees	Nikolaj Tatti
propositionalization	Claudia Perlich Foster J. Provost
map-reduce	Raghu Ramakrishnan
ordinal regression	Torsten Joachims
robust fitting	Saharon Rosset
cost quantification	George Forman
goodness score	Hanghang Tong Christos Faloutsos

Another question is that whether adviser of the student takes higher scores or not in the network. The answer for this question is yes. Here we provide specific evidence to support our idea that the high PageRank values are indicators of being an expert. As a general matter of fact, advisers are better experts than their students and they appear as co-authors with their students in lots of papers. Then, advisers should have higher PageRank scores than their students.For instance, Christos Faloutsos , Jiawei Han and their students are examined and the following results in the Table 4 are obtained.

Table 4. Sample Advisers(C. Faloutsos & Jiawei Han) and Their Student Rankings

Author Name	**Rank**		**Author Name**	**Rank**
Christos Faloutsos	8		Jiawei Han	11
Lei Li	180		Jialu Liu	1078
Jure Leskovec	88		Brandon Norick	5225
Flip Korn	458		Xiang Ren	5900
Deepayan Chakrabarti	200		Jingbo Shang	4156
Spiros Papadimitriou	2357		Fangbo Tao	5386
Jimeng Sun	126		Jingjing Wang	3774
Hanghang Tong	124		Chao Zhang	2863
Fan Guo	1163			
Leman Akoglu	408			
Evangelos E. Papalexakis	1663			
Alex Beutel	1483			
Jay Yoon Lee	7301			
Zhiqiang Bi	742			
Mary Mcglohon	964			

As it is seen from Table 4, Christos Faloutsos and Jiawei Han have higher rank from their students as in our expectations.

As for the citation network, it is applied HITS algorithm, and authority and hub ranks for each publication in the network are obtained. In the Table 5, authority and hub ranks for Author names listed in Top-50 from author-keyword bipartite graph can be seen. When these ranks are examined, it is not possible to say that there is a direct relationship between these ranks and Page Rank ranks of the author-keyword bipartite graph for that authors. The nodes in the citation network represent publications whereas nodes in the author-keyword bipartite graph represent author and keyword labels. Since the nodes types in the citation network and the author-keyword bipartite graph are different from each other, it is usual not to relate them directly. However authority and hub ranks are high for the authors like in the PageRank ranks. For that reason,these two rank data(PageRank ranks and HITS ranks) have potential to be combined with a new structural work/experiment.

Table 5. Authority&Hub Ranks Author Names Listed in Top-50

Vertex Label	Authority Rank	Hub Rank
jieping-ye	242	210
christos-faloutsos	3	26
philip-s.-yu	23	14
jiawei-han	51	42
ravi-kumar	11	49
hui-xiong	111	201
bing-liu-0001	188	339
huan-liu	42	80
tao-li	107	93
deepak-agarwal	44	54
chengxiang-zhai	34	157
srinivasan-parthasarathy	36	33
thorsten-joachims	127	132
heikki-mannila	16	18
jian-pei	113	173
naren-ramakrishnan	258	56
vipin-kumar	231	255
evimaria-terzi	16	11

When it comes to the meaning of the concept of trend, the Oxford web dictionary[39] defines the trend as “A general direction in which something is developing or changing”. Therefore, trend word evokes absolute increase, decrease or stationary pose as general way of changing of something in our minds. However, they are not the only cases to express a trend. In the context of this work, we define the *topical trend* as how a topic (keyword) popularity changed over time.

In order to determine trends, how a keyword popularity changed over time should be examined according to our trend definition. That’s why, fifteen sub author-keyword bipartite graphs are prepared for each year between 2000 and 2014. Then, PageRank is applied to all sub-graphs to see the popularity/importance of keyword for a specific year. After that, PageRank scores of each keyword according to the years are collected. Then, trend graphics are drawn for each keyword. On the x-axis,

there is year information from 2000 to 2014 while on the y-axis, there are PageRank scores of the keyword with respected to these years.

As for detecting trends, we divided the keywords into three groups by using human judgment as having high PageRank score, having moderate level PageRank score and having low page rank score. The keywords ranked in top-190 are regarded as the keyword group that has high PageRank score while the keywords ranked after top-6630 are regarded as the keyword group that has low PageRank score. The keyword group that has moderate level PageRank score are considered as the keywords between these two groups. We spared them into groups with the insight that there is possibly common behavior in each group. According to the plots for selected words from each category, we see a common pattern in each group; therefore the results supported our idea.

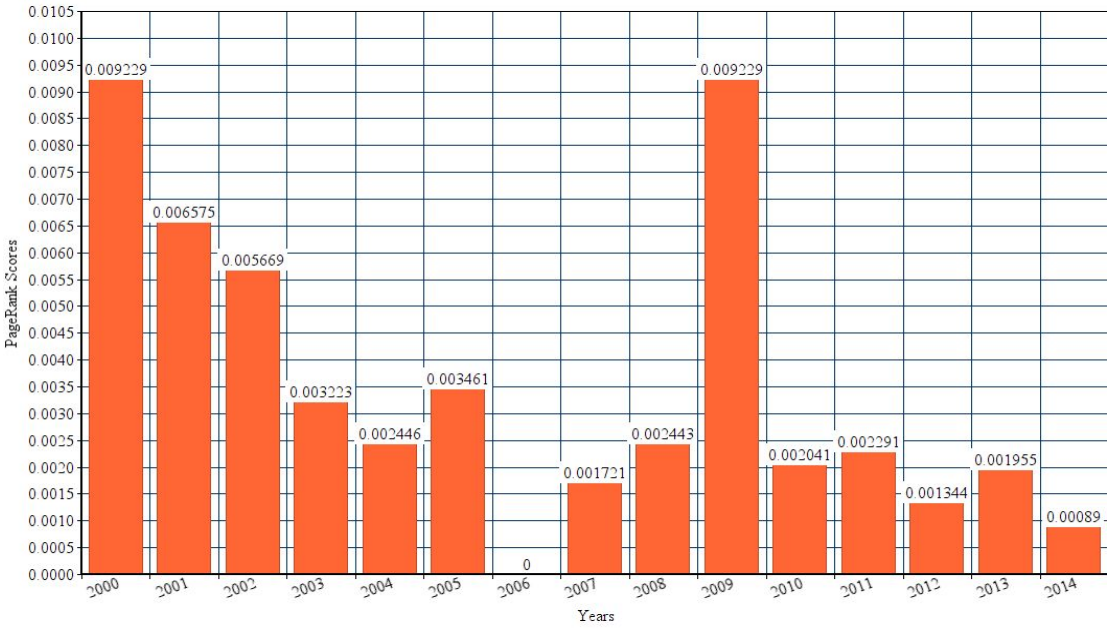


Figure 12. Trend Plot for “data-mining” Keyword

As for the groups having high PageRank score like *data-mining, clustering, classification, social-networks, machine-learning* and *text-mining* have the almostly same shape like in Figure 12 that makes a deep in 2006 and a top in 2009. Since those words are general and they are always worked, their plots contain many bars. In Figure

12, trend plot for data mining keyword is seen as an example of the trend plot of this group. However, only plots of *social-networks* and *machine-learning* poses different because the keyword set we constructed out of our author-keyword bipartite graph contains variations of those keywords like s*ocial-network,-social-network, -machine-learning.* The plots belonged to the selected keywords except data-mining for this group are shown in the figures in Appendix D.

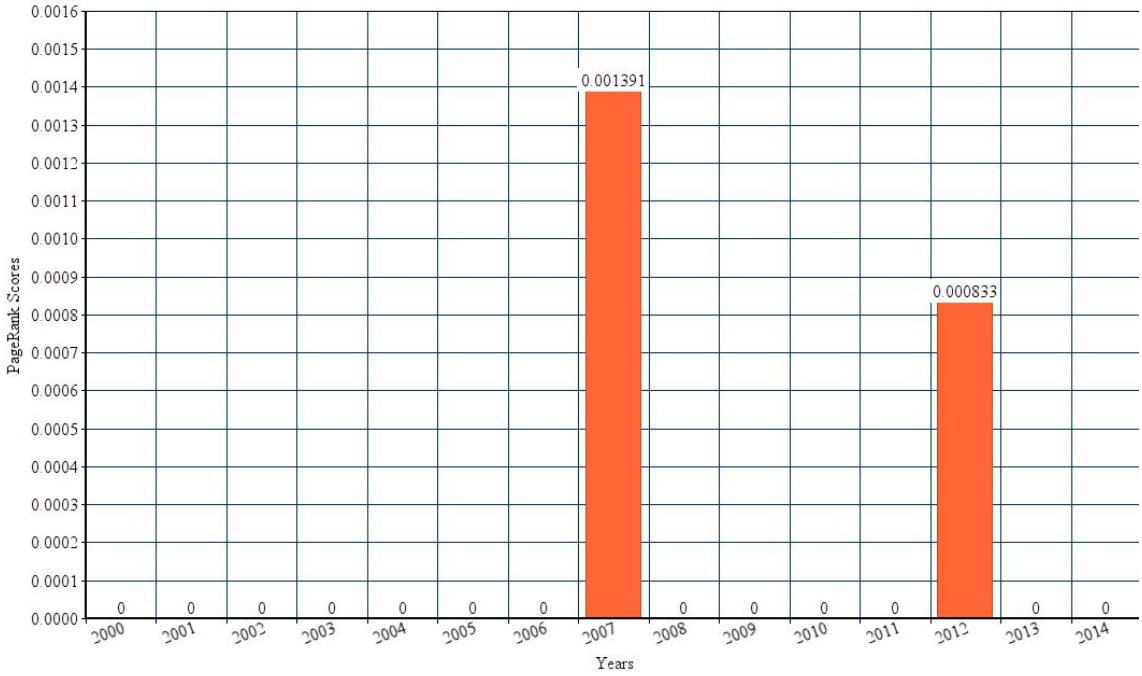


Figure 13. Trend Plot for “algorithmic-advertising” Keyword

As for second group that have moderate level page rank score like s*parcity, sports-analytics, online-communities, real-time-bidding, polarity-analysis,* and *algorithmic-advertising,* we see that they have a pattern to peak and tend to repeat in doing peak behavior. In Figure 13, trend plot for “algorithmic-advertising” keyword is seen. In the figure, there is a peak and repetition of making a peak. Probably, it will make another peak in some year because that group contains moderate level of keywords that most probably will be worked from time to time.The plots belonged to the selected keywords except algorithmic-advertising for this group are shown in Appendix D.

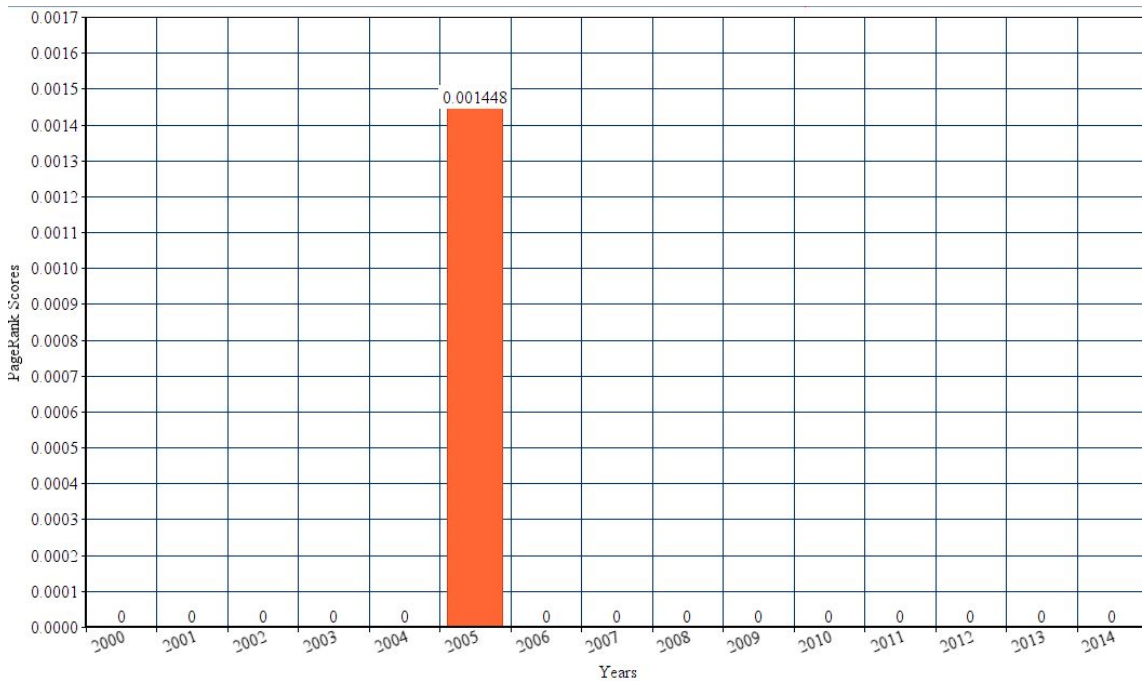


Figure 14. Trend Plot for "general-model" Keyword

When it comes to the groups having too low page rank score like *general- model, robust-fitting, knowledge-acquisition, natural-language-computing, privacy-in-data mining, cost-quantification* and *prior-knowledge,* have a pattern to make at most one peak. In Figure 14, trend plot for "generic-model" keyword is seen. Since the keywords in those group are quite specific, their PageRank scores are too low because of their less popularity. Thus, they tend to make at most one peak or no peak at all. The plots belonging to the selected keywords except generic-model for this group are shown in Appendix D.

Such an approach is convenient for determining trends because popularity measure is taken from these subgraphs by applying PageRank algorithm for a keyword for each year to track how its popularity changes over time. In keeping with our trend definition, we track the changes in popularity of keywords over time with this approach.

Similar to us, Newman et al.[42] interpret topical trends in their work. They propose a framework to extract topics, topic trends, topics relating entities in a large set of documents to find out an appropriate topic for each document inside the collection. After they. find the topics, they divide the topics into two sub-groups: *four seasonal topics* and *event* topics. Four seasonal topics are related to general news while event

topics are related to breaking news like September 11 Attacks, the news containing runaway successes like box-office return of Harry Potter and the Half-blood Prince or containing too specific details like DC Sniper that John Muhammad and Lee Malvo that they were in a white van. They obtained the results in Figure 15. The detailed explanation can be found in the paragraph below.

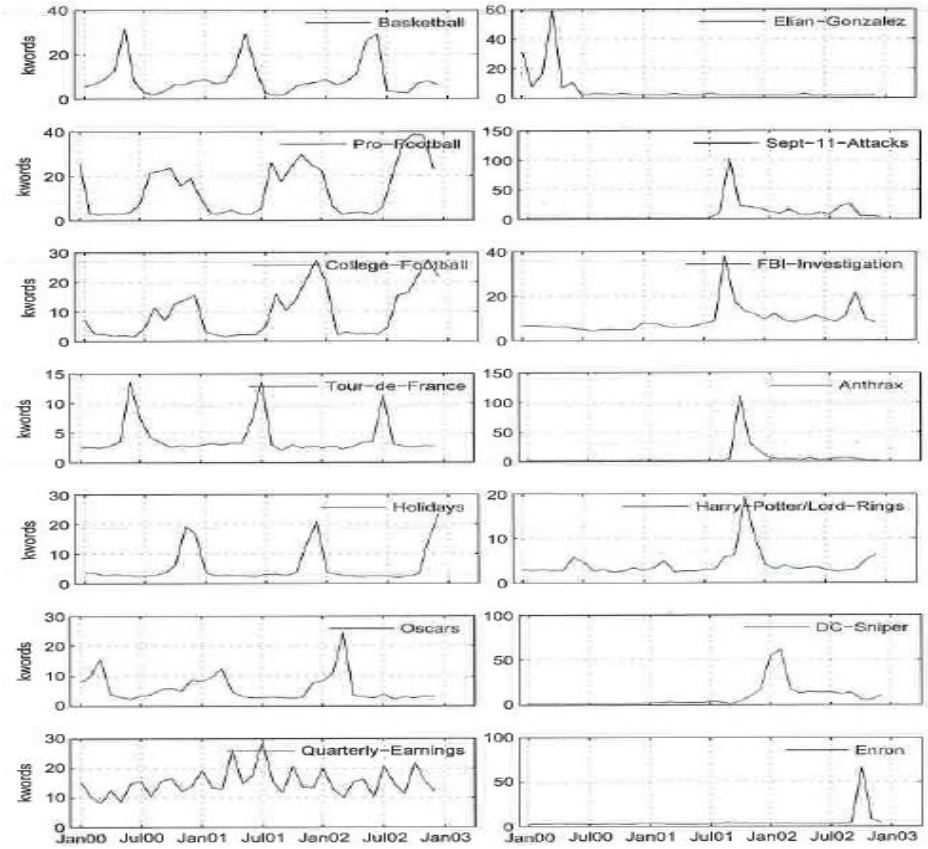


Figure 15. Selected Topic-trends by Newman et al.

They say that there is difference in the curves on left hand side and the ones on the right hand side. The ones on the left hand side are four seasonal topics and have periodicity while the ones on the right hand side are event topics and have peaks. The periodicity of the curve on the left represents trends of four seasonal topics while the peaks of the curve on the right represent trends of event topics. Hence, trend means a pattern in the data, it can show different patterns like making a peak, having a periodicity like we assert.

CHAPTER 5

CONCLUSION

The aim of this thesis is to deduce the subject-matter experts and find topical research trends on Data Mining and Knowledge Discovery proceedings and corresponding dblp bibliographic data of these proceedings by using link analysis algorithms like PageRank and HITS.

Expert is someone having a special skill or knowledge obtained from training or experience. In the context of this work, two types of authors are regarded as experts. First type is the authors working on intensively of a single subject and the second type is authors having high level knowledge of various subtopics of a subject-matter [43]. By using this expert definition, we run random walk (PageRank algorithm) on our author-keyword bipartite graph. After running random walk on the bipartite graph, we see that the keywords having high page rank scores imply generality while the author names having high page rank imply those authors work on general keywords. These author profiles that have high PageRank are checked and eleven of them found to write a book on subjects in KDD. Since writing a book needs specialties, they are good nominees to become subject-matter experts.

In addition to this, we see that experts working on very specific subject-matters can be distinguished because too specific keywords are used as subject descriptors by a handful of authors. Thus, those can be regarded as experts for that subtopics. Hence, we cover both becoming experts on specific matters and becoming expert on a general subject-matter condition.

In general, trend word evokes absolute increase, decrease or stationary pose as general way of changing of something in our minds. However, they are not the only cases to express a trend. In the context of this work, we define the *topical trend* as change of popularity of a topic(keyword) over time. That's why, we prepare fifteen sub author-keyword bipartite graphs for each year between 2000 and 2014. Then, we apply PageRank to all sub-graphs to see the popularity/importance of keyword for a specific

year. After that, we collect the PageRank scores of each keyword according to the years and we obtain trend graphics for each keyword. For detecting trends, we divided the keywords into three groups as having high PageRank score, having moderate level page rank score and having low page rank score. According to the curves for selected words from each category, we observe a common pattern inside of each group. We see that the keywords that have high PageRank score have almost the same shape and they show continuity in the curve. In addition, the keywords having moderate level PageRank score have a pattern to peak and tend to repeat in doing peak behavior. Moreover, the keywords having low PageRank score tend to make at most one peak or no peak at all.

In finding subject-matter experts and research trends, the results from the author-keyword bipartite graph can be validated or strengthened using the data coming from the citation network. Integrating these two knowledge sources has the potential to say more. One such integration method is the transformation of the citation graph into a multi-view graph by the addition of alternative edges. These alternative edges can be created through the use of keyword set/or author(s) similarity between the articles. Thus, doing analysis on this multi-view graph can produce useful insights regarding the high-level concept of subject-matter expert.

REFERENCES

[1] ACM Digital Library, [Internet] http://dl.acm.org/ (accessed date: March,2015)

[2] Google Scholar[Internet], https://scholar.google.com (accessed date : January 2015)

[3] Microsoft Academic Search [Internet], http://academic.research.microsoft.com/ (accessed date : 2 May 2015).

[4] ArnetMiner[Internet], http://www.arnetminer.org (accessed date: March,2015)

[5] CiteSeerX Digital Library and Search Engine [Internet], http://citeseerx.ist.psu.edu/index (accessed date : January 2015)

[6] Rexa Academic Search Engine[Internet] http://rexa.info/ (accessed date: May,2015)

[7] J. Kleinberg. Authoritative sources in a hyperlinked environment, 1999.

[8] S. Brin, and L. Page. The anatomy of a large scale hypertextual Web search engine. In Proc. of WWW7, Brisbane, Australia, April, 1998.

[9] ArnetMiner Citation Network DataSet“DBLP_citation_2014_May” [Internet] https://aminer.org/lab-datasets/citation/ (accessed date: March,2015)

[10] OECD Glossary of Statistical Terms. [Internet], http://stats.oecd.org/glossary/ (accessed date: June 2015)

[11] 22nd International Symposium On Graph Drawing[Internet] http://gd2014.informatik.uni-wuerzburg.de/ (accessed date: May 2015)

[12] Citation Network Picture[Internet], http://ongraphs.de/blog/wp-content/uploads/2014/01/close-up-gd-network.png (accessed date: May 2015)

[13] Glossary of Graph Terms[Internet], http://primes.utm.edu/graph/glossary.html(accessed date: May 2015)

[14] Online Math Resource[Internet]
http://mathworld.wolfram.com/BipartiteGraph.html,
(accessed date:February 2015)

[15] Internet Encyclopedia[Internet], https://en.wikipedia.org/wiki/Bipartite_graph
(accessed date: February 2015)

[16] A library page[Internet]
http://www.library.illinois.edu/learn/research/citationanalysis.html,
(accessed date : June 2015)

[17] Barry, M. and G. Linoff (1997). "Data Mining Techniques - for marketing, sales and customer support.", Wiley Computer Publishing p. 216-242

[18] Internet Encyclopedia[Internet],
https://en.wikipedia.org/wiki/Link_analysis(accessed date: February 2015)

[19] Wiki page[Internet],
http://wiki.ubc.ca/Library:Citation_Analysis_&_Impact_Factors
(accessed date: February 2015)

[20] M.Franceschet. PageRank : Standing on the shoulders of giants.
In Communications of the ACM,Volume 54 Issue 6, June 2011, pages 92-101

[21] A Search Engine[Internet], http://www.ask.com/ (accessed date : January 2015)

[22] Lecture#4 : HITS algorithm - Hubs ans Authorities on the Internet[Internet],
http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html,
(accessed date : January 2015)

[23] P. Devi , A. Gupta , A. Dixit. Comparative Study of HITS and PageRank Link based Ranking Algorithms. International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014

[24] G.Pinski and F. Narin. Citation influence for journal aggregates of scientific publications. Theory, with application to the literature of the physics.Information Processing & Management, 12(5), pages 297-312, 1976.

[25] I.Palacios-Huerta and O. Volij. The measurement of intellectual influence. Econometrica, pages 963-977, 2004

[26] Z. Guo and D. Barbosa. Robust Entity Linking via Random Walks. In Proc. of CIKM' 14, pages 499-508

[27] D.Shahaf and C. Guestrin, Connecting the Dots Between News Articles. In Proc. of KDD' 10, pages 623-632.

[28] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proc. of SIGIR'06*, pages 43–55, 2006.

[29] Text Retrieval Conference[Internet], http://trec.nist.gov/ (accessed date: July 2015)

[30] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In Proc. of UAI'04, 2004.

[31] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (SIGKDD'2008). p.990-998.

[32] J. Tang, D. Zhang, and L. Yao. Social network extraction of academic researchers. In Proc. of ICDM'07, pages 292–301, 2007.

[33] D. Wang, T. Li, and M. Ogihara. Generating pictorial storylines via minimum weight connected dominating set approximation in multi-view graphs. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pages 683–689. AAAI, 2012.

[34] L. Zhang, L. Li, T. Li, and D. Wang. PatentDom: Analyzing Patent Relationships on Multi-View Patent Graphs.CIKM'14,2014

[35] Iniversitat Trier[Internet],https://www.uni-trier.de/(accessed date: October 2014)

[36] A Non-Profit Organization[Internet], https://www.dagstuhl.de/en/about-dagstuhl/ (accessed date: October 2014)

[37] A Java PDF Library[Internet], https://pdfbox.apache.org/(accessed date: November 2014)

[38] D. Mimno and A. McCallum. Expertise modeling for matching papers with reviewers. In Proc. of KDD'07, pages 500–509, 2007.

[39] Oxford Online Dictionary[Internet]
http://www.oxforddictionaries.com/definition/english/trend,
(accessed date: June 2015)

[40] D.F. Gleich. PageRank beyond the web. Journal paper. SIAM Review, 57(3):321-363, August 2015.

[41] P. Sarkar. Random walks on graphs: an overview [Internet]
http://docslide.us/documents/1-random-walks-on-graphs-an-overview-purnamrita-sarkar.html (accessed date : July 2015)

[42] D.Newman, C. Chemudugunta, P. Smyth,M. Steyvers. Analysing entities and topics in news articles using statistical topic models. In IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23-24, 2006. Proceedings. p.93-104

[43] Tekir, S.(2004). *An implementation model for open sources evaluation* (Master's thesis, Izmir Institute of Technology, Izmir, Turkey).
Retrieved from http://hdl.handle.net/11147/3329

APPENDIX A

EXPERIMENTAL RESULT TABLES

Table 6. HITS Authority Scores for the Citation Network Vertices

Rank 1: 0.14857286650869536	Maximizing the spread of influence through a social network
Rank 2: 0.08759277368395169	Mining the network value of customers
Rank 3: 0.05964563924140573	Cost-effective outbreak detection in networks.
Rank 4: 0.05679954709744769	Mining knowledge-sharing sites for viral marketing
Rank 5: 0.04568242905682805	Efficient influence maximization in social networks
Rank 6: 0.03613468879589923	Scalable influence maximization for prevalent viral marketing in large-scale social networks
Rank 7: 0.02449732689963539	Group formation in large social networks: membership, growth, and evolution
Rank 8: 0.023416401436626415	Graphs over time: densification laws, shrinking diameters and possible explanations
Rank 9: 0.02012191069448858	Inferring networks of diffusion and influence
Rank 10: 0.020020717655905827	Meme-tracking and the dynamics of the news cycle
Rank 11: 0.019656155144042096	Influence and correlation in social networks
Rank 12: 0.01931296470395942	Social influence analysis in large-scale networks
Rank 13: 0.014267345234105173	Feedback effects between similarity and social influence in online communities
Rank 14: 0.013189568818965117	Information diffusion and external influence in networks
Rank 15: 0.012807931622173243	ArnetMiner: extraction and mining of academic social networks
Rank 16: 0.01241545730151184	Finding effectors in social networks
Rank 17: 0.0118039966980966	Fast discovery of connection subgraphs
Rank 18: 0.008879303769879613	Bursty and hierarchical structure in streams
Rank 19: 0.007848326387303785	Microscopic evolution of social networks
Rank 20: 0.007701603783072848	Structure and evolution of online social networks

Table 7. HITS Hub Scores for the Citation Network Vertices

Ranks	Title of the Publication
Rank 1: 0.021352171680167314	Sparsification of influence networks
Rank 2: 0.020451387212884076	Information cascade at group scale
Rank 3: 0.019549352355277626	Confluence: conformity influence in large social networks
Rank 4: 0.019546049806044966	Trial and error in influential social networks
Rank 5: 0.018789196428689946	Scalable influence maximization for prevalent viral marketing in large-scale social networks
Rank 6: 0.017332618195616285	Social action tracking via noise tolerant time-varying factor graphs
Rank 7: 0.017261652166763918	Cascading outbreak prediction in networks: a data-driven approach
Rank 8: 0.016862436446461477	Minimizing seed set selection with probabilistic coverage guarantee in a social network
Rank 9: 0.016314582739303245	Fast influence-based coarsening for large networks
Rank 10: 0.01598868274125987	STRIP: stream learning of influence probabilities
Rank 11: 0.01586488361678608	Repetition-aware content placement in navigational networks
Rank 12: 0.01586488361678608	Efficient influence maximization in social networks
Rank 13: 0.015640146960366866	Community-based greedy algorithm for mining top-K influential nodes in mobile social networks
Rank 14: 0.01541674030597325	Extracting social events for learning better information diffusion models
Rank 15: 0.014998522528618812	Finding trendsetters in information networks
Rank 16: 0.014629984390262528	Challenges in mining social network data: processes, privacy, and paradoxes
Rank 17: 0.01430272849715582	Group formation in large social networks: membership, growth, and evolution
Rank 18: 0.013560266781898254	Finding effectors in social networks
Rank 19: 0.012681082030983318	RecMax: exploiting recommender systems for fun and profit
Rank 20: 0.012402226951158784	Social influence based clustering of heterogeneous information networks

APPENDIX B

VISUALIZATION of PAGERANK SCORES

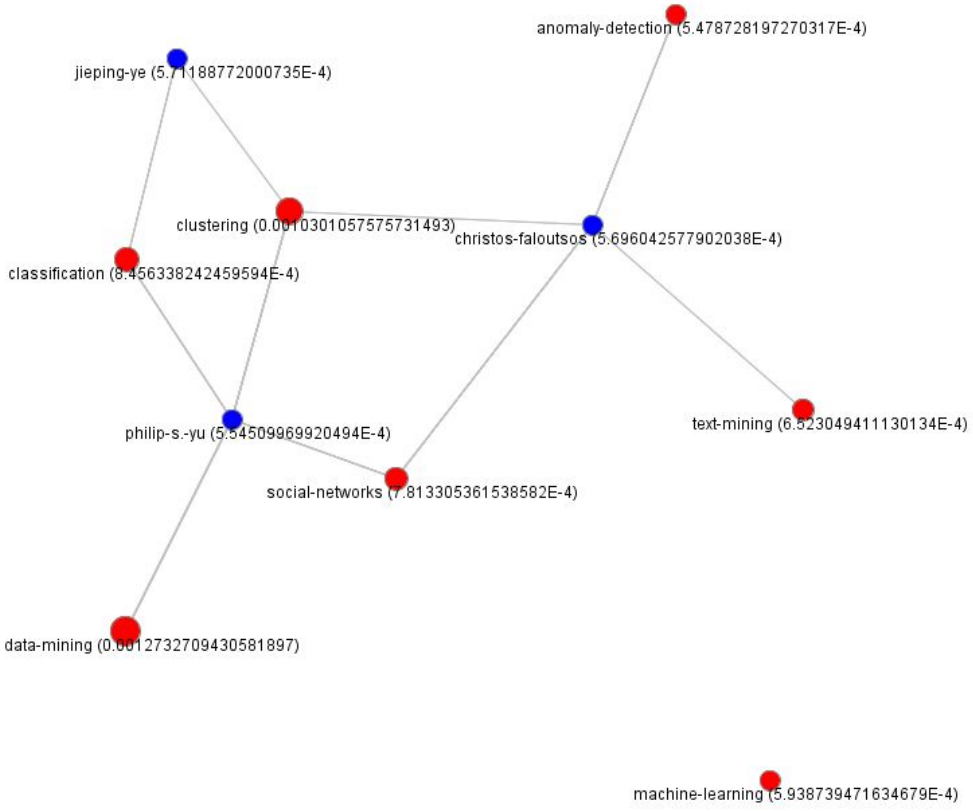


Figure 16.Visualization of The PageRank Scores (lower limit is 0.0005478)

APPENDIX C

VISUALIZATION of CITATION NETWORK

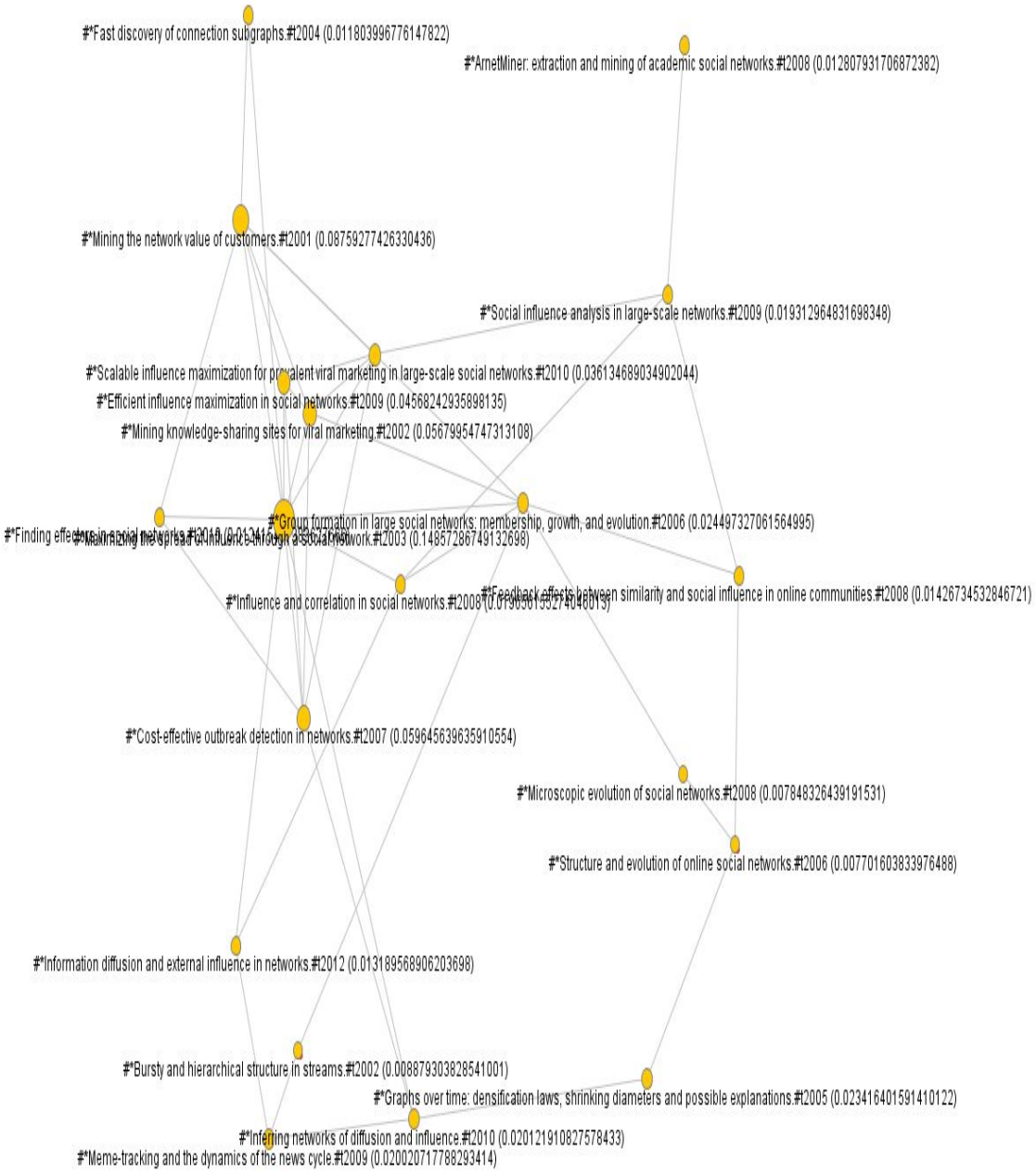


Figure 17. The Vertices That Have Highest 20 Authority Scores in Citation Network Without Reference Consideration (threshold is 0.007)

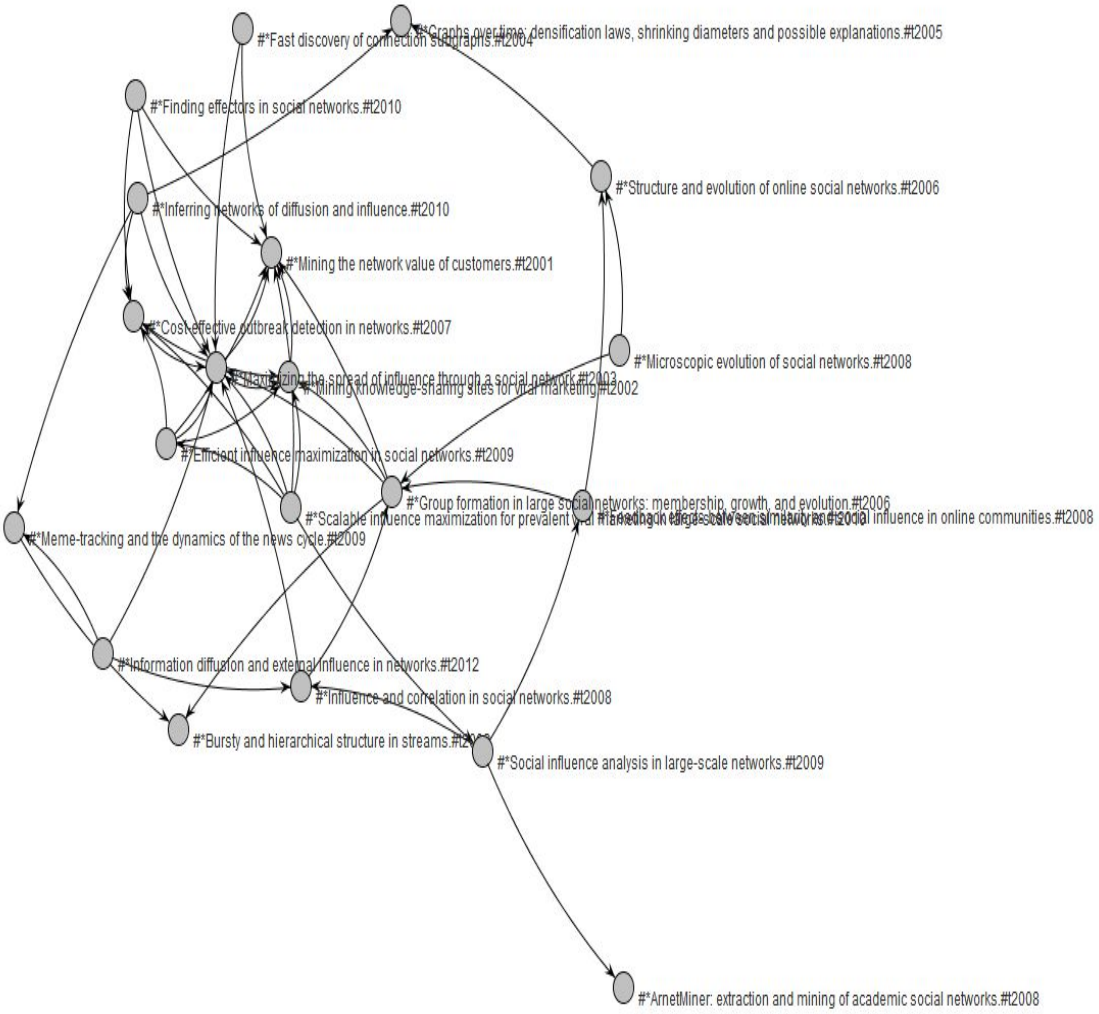
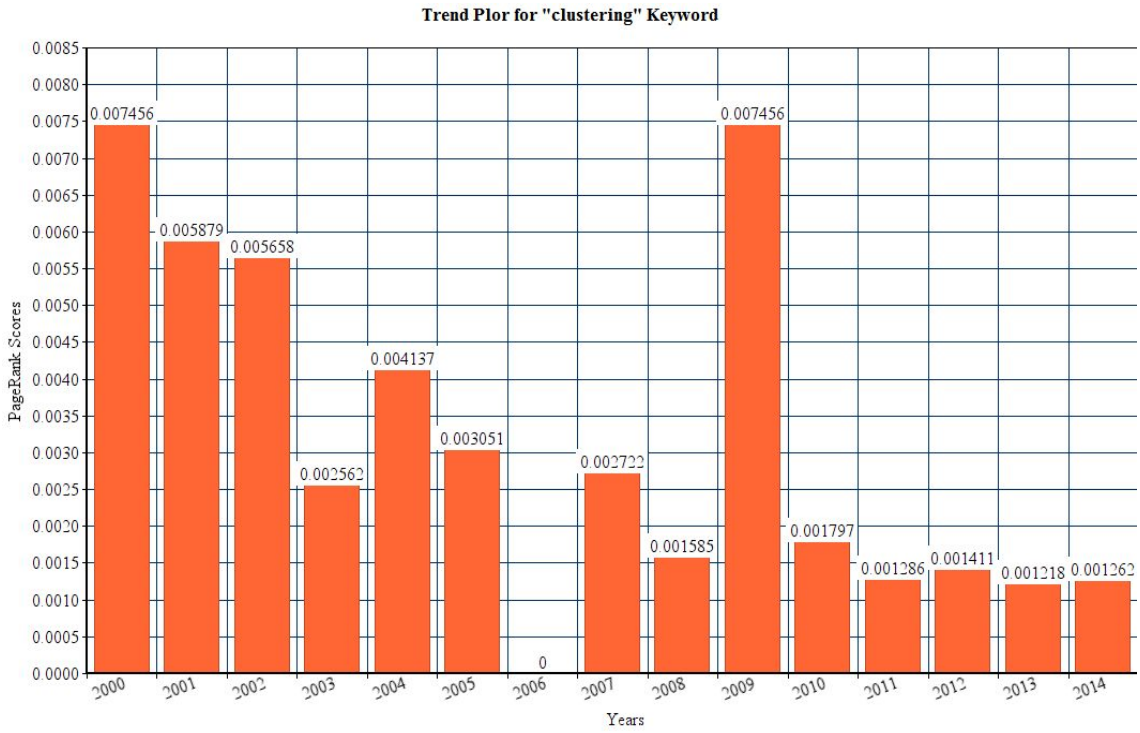


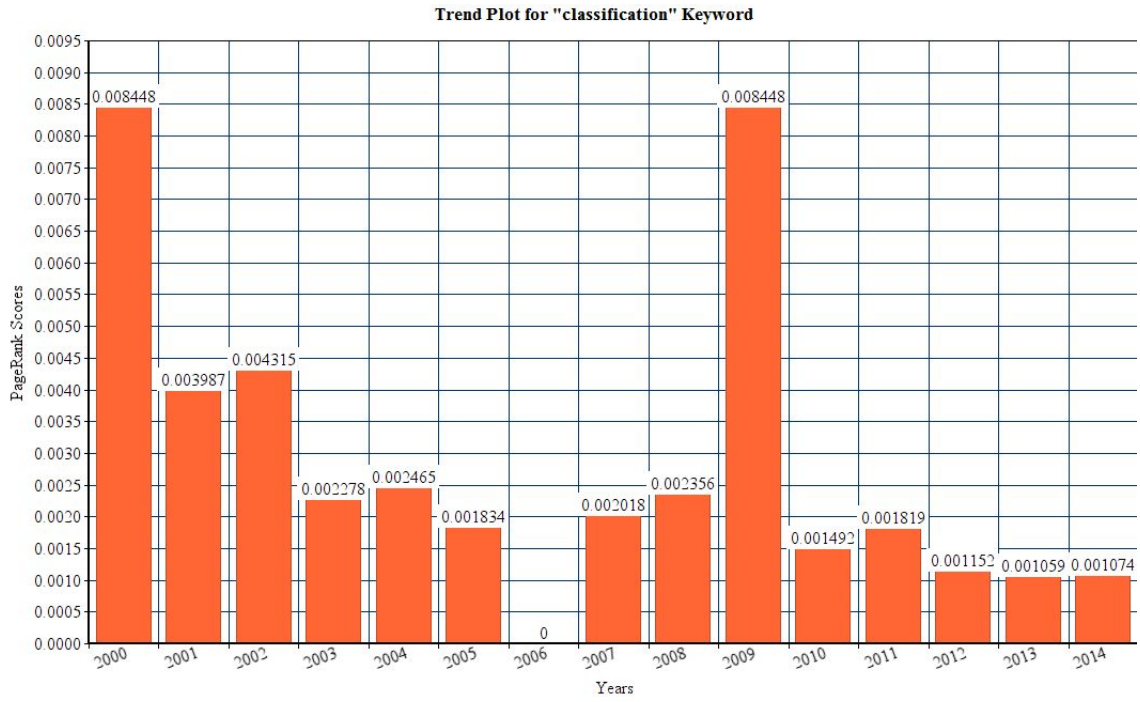
Figure 18. The Vertices That Have Highest 20 Authority Scores in Citation Network With Reference Consideration (threshold is 0.007)

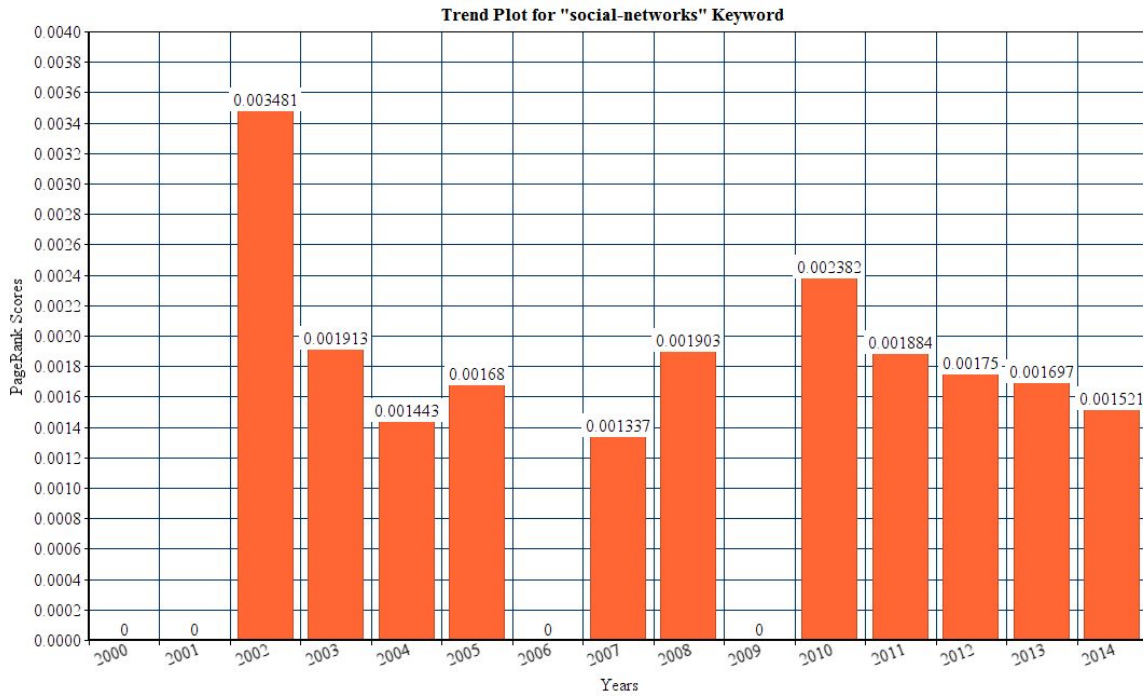
APPENDIX D

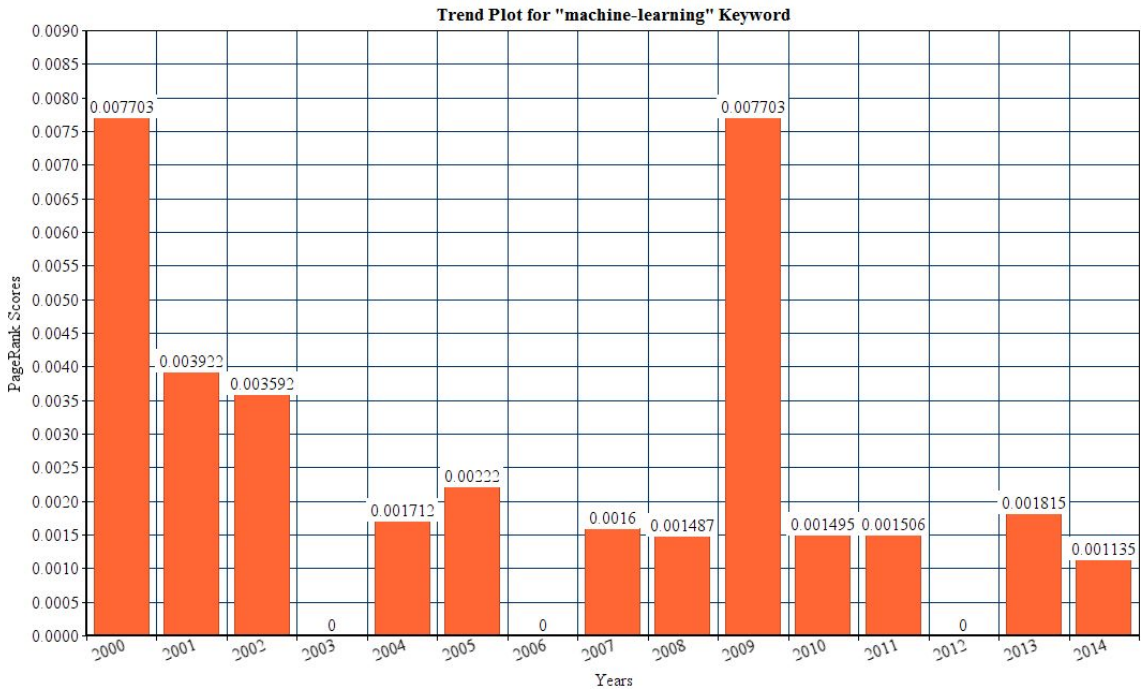
TREND PLOTS of SELECTED KEYWORDS

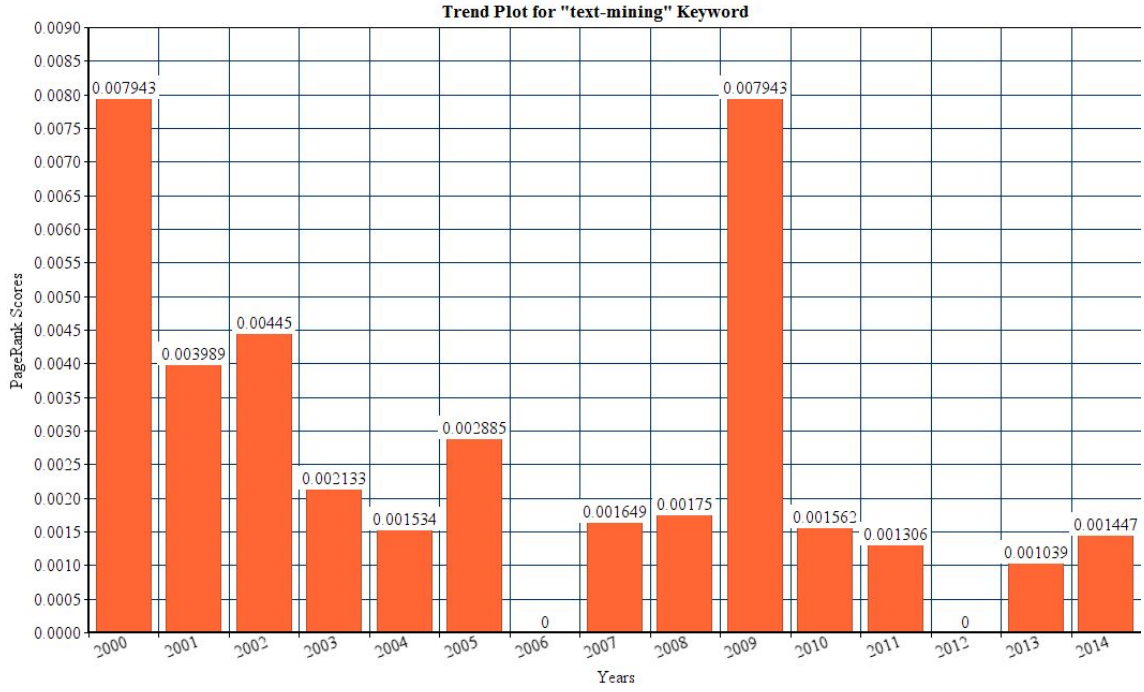
A) Trend Plots for Some Selected Keywords that have high PageRank



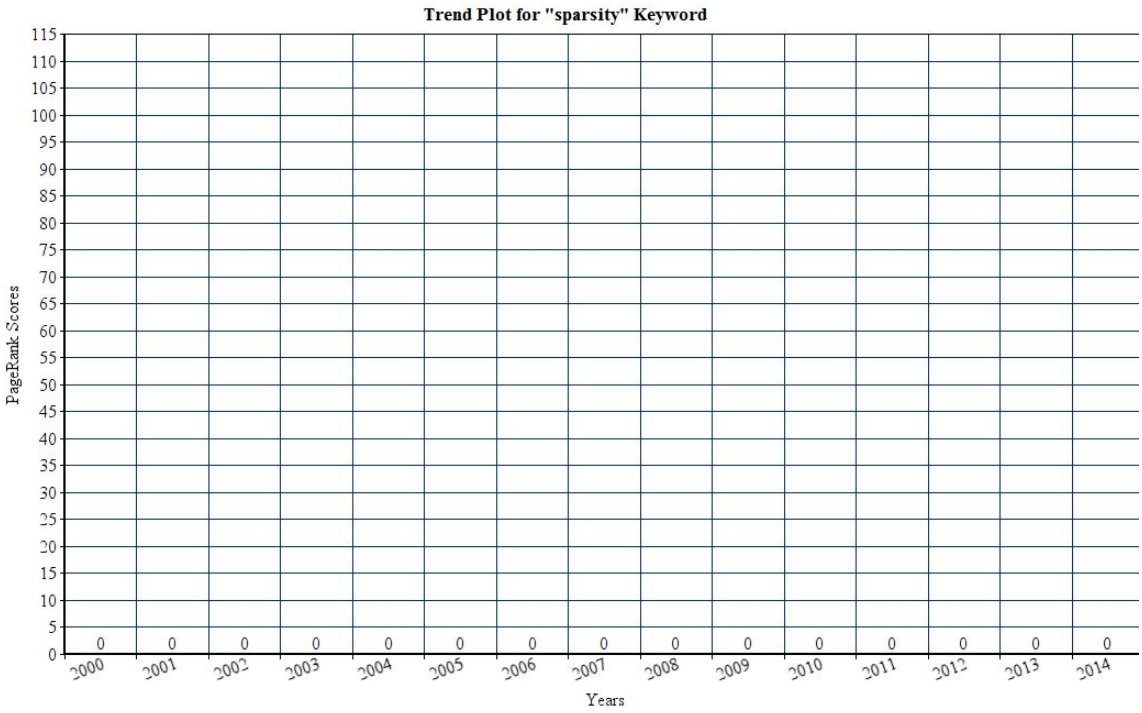

Trend Plot for "classification" Keyword
PageRank Scores
Years
0.008448
0.003987
0.004315
0.002278
0.002465
0.001834
0
0.002018
0.002356
0.008448
0.001492
0.001819
0.001152
0.001059
0.001074
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

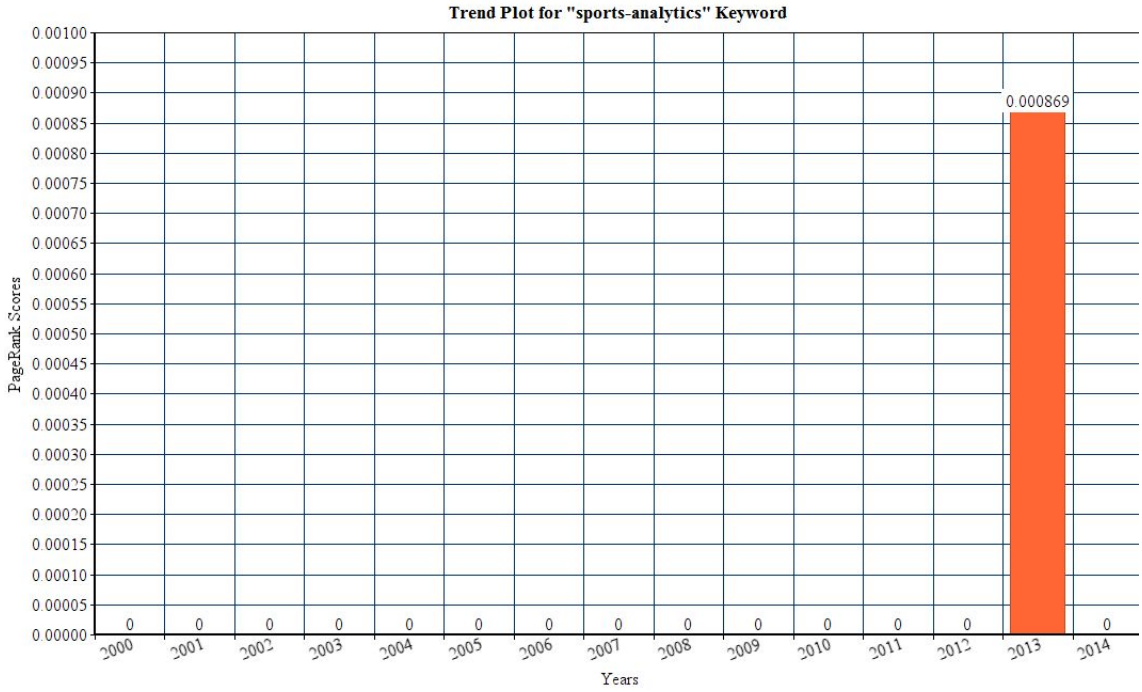

Trend Plot for "social-networks" Keyword
PageRank Scores
Years
0
0
0.003481
0.001913
0.001443
0.00168
0
0.001337
0.001903
0
0.002382
0.001884
0.00175
0.001697
0.001521
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

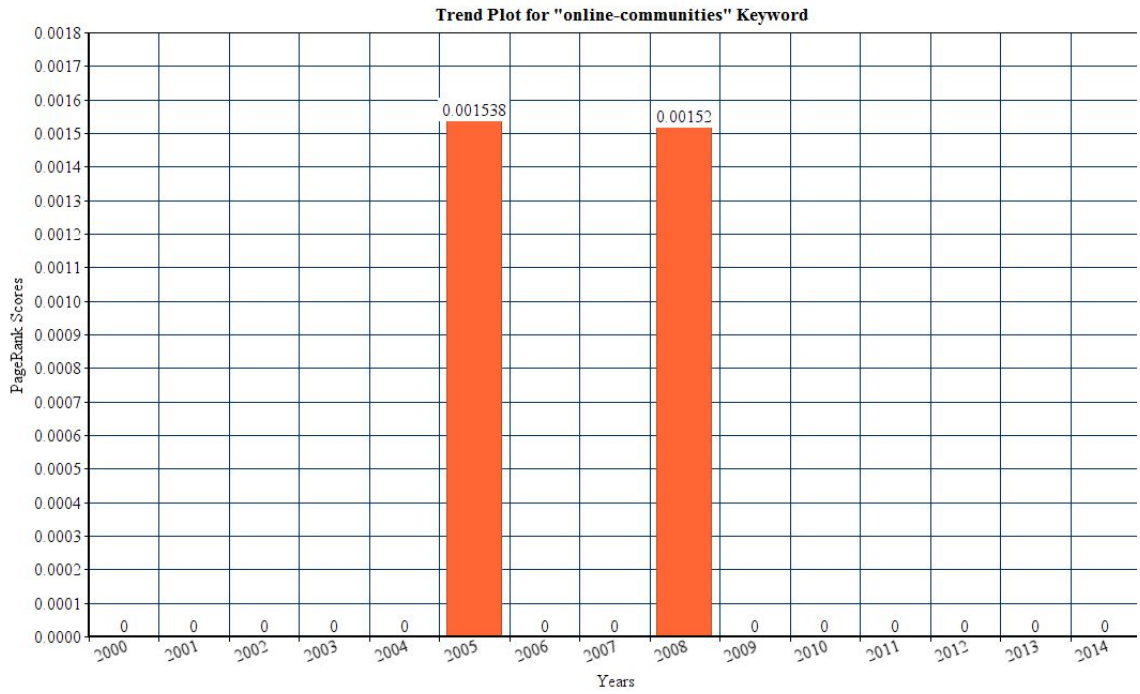

Trend Plot for "machine-learning" Keyword
PageRank Scores
Years
0.007703
0.003922
0.003592
0
0.001712
0.00222
0
0.0016
0.001487
0.007703
0.001495
0.001506
0
0.001815
0.001135
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

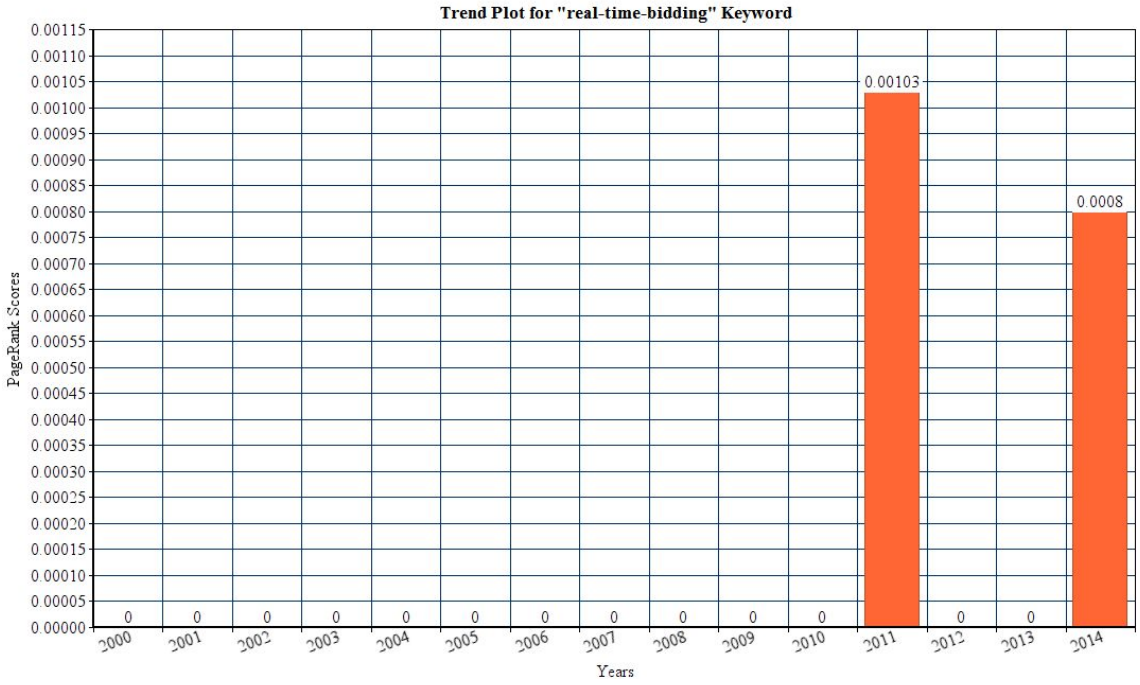

Trend Plot for "text-mining" Keyword
PageRank Scores
Years
0.007943
0.003989
0.00445
0.002133
0.001534
0.002885
0
0.001649
0.00175
0.007943
0.001562
0.001306
0
0.001039
0.001447
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

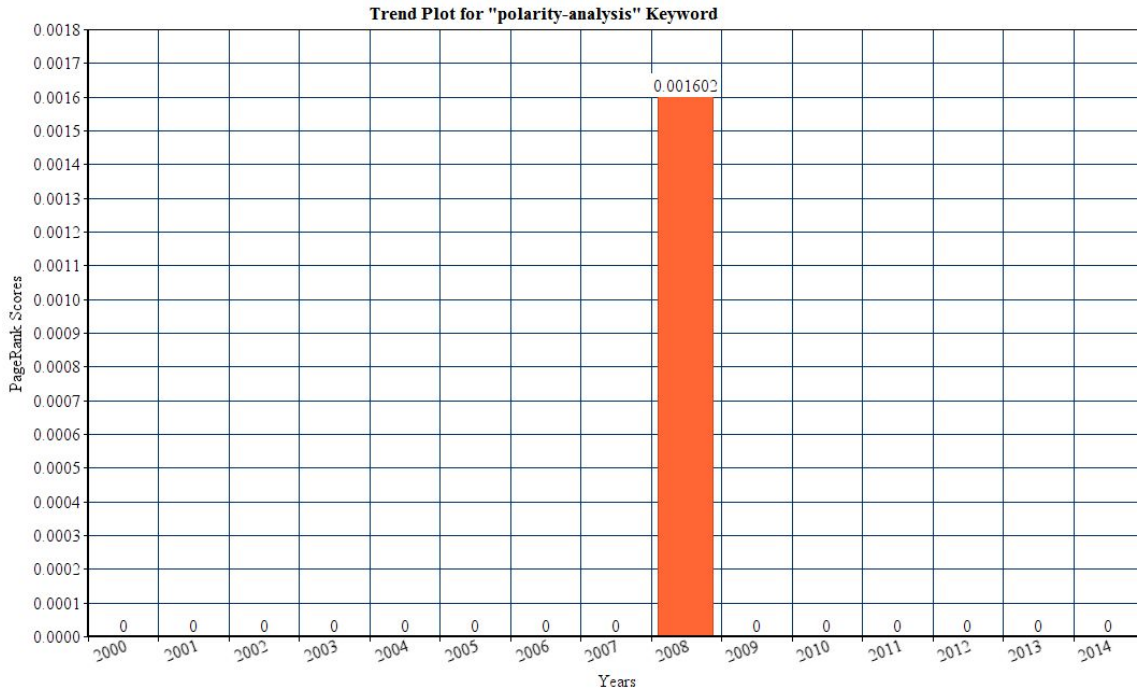
B) Trend Plots for Some Selected Keywords that have moderate level of PageRank

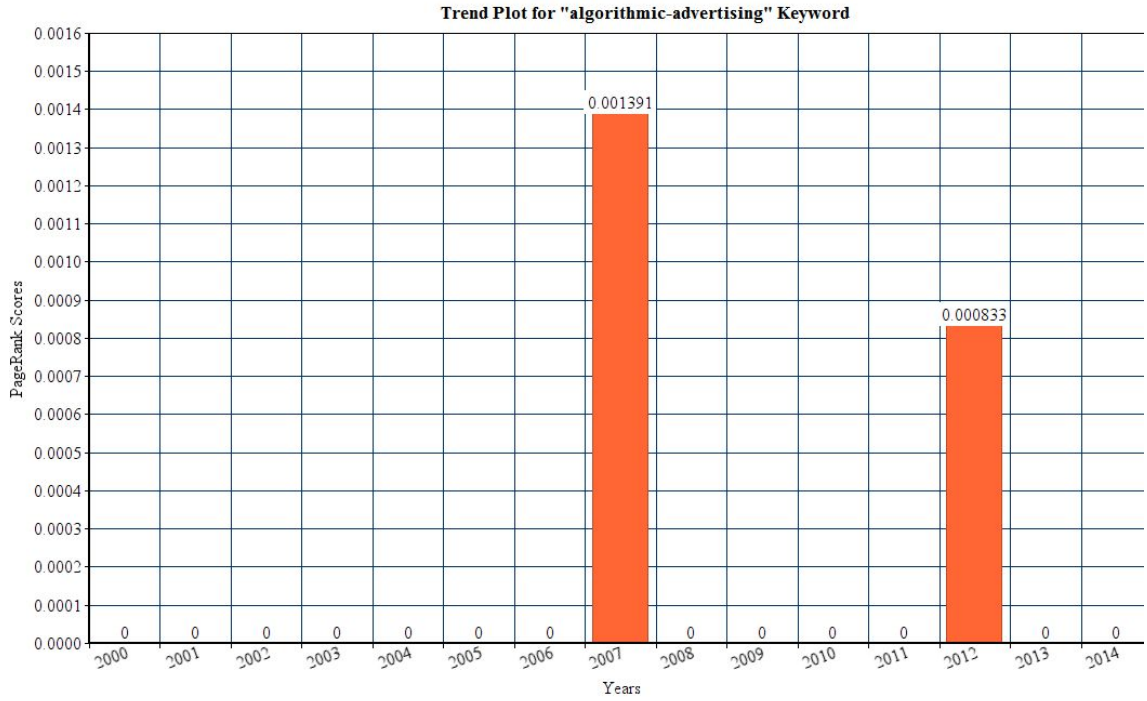




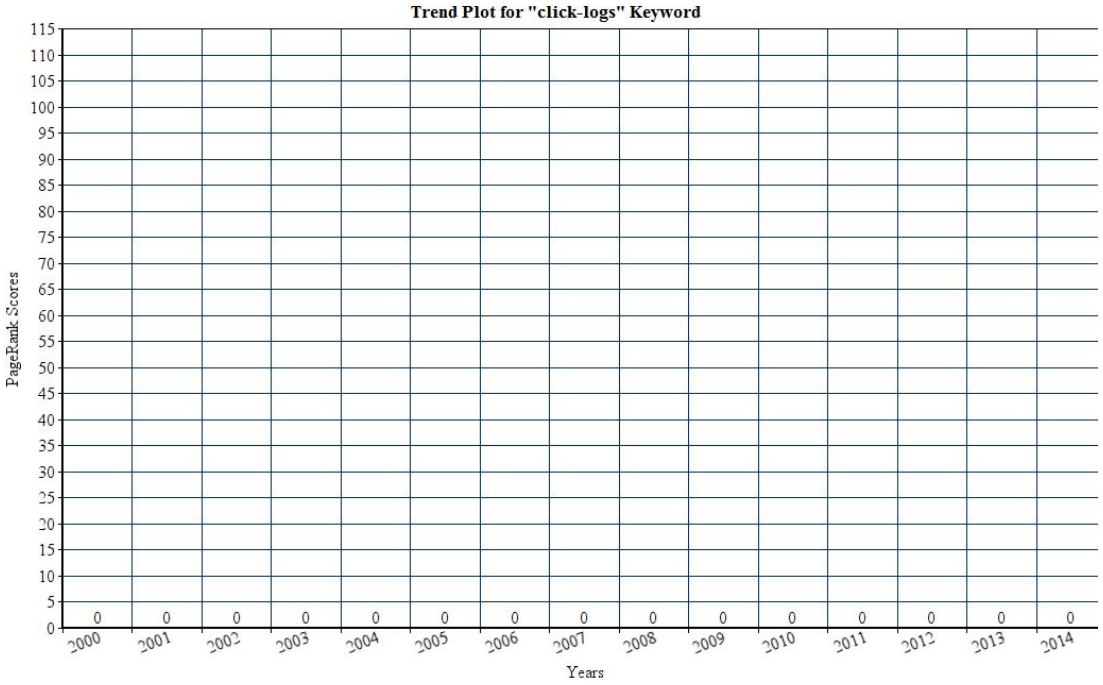

Trend Plot for "online-communities" Keyword
PageRank Scores
Years
0.001538
0.00152
0.0018
0.0017
0.0016
0.0015
0.0014
0.0013
0.0012
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

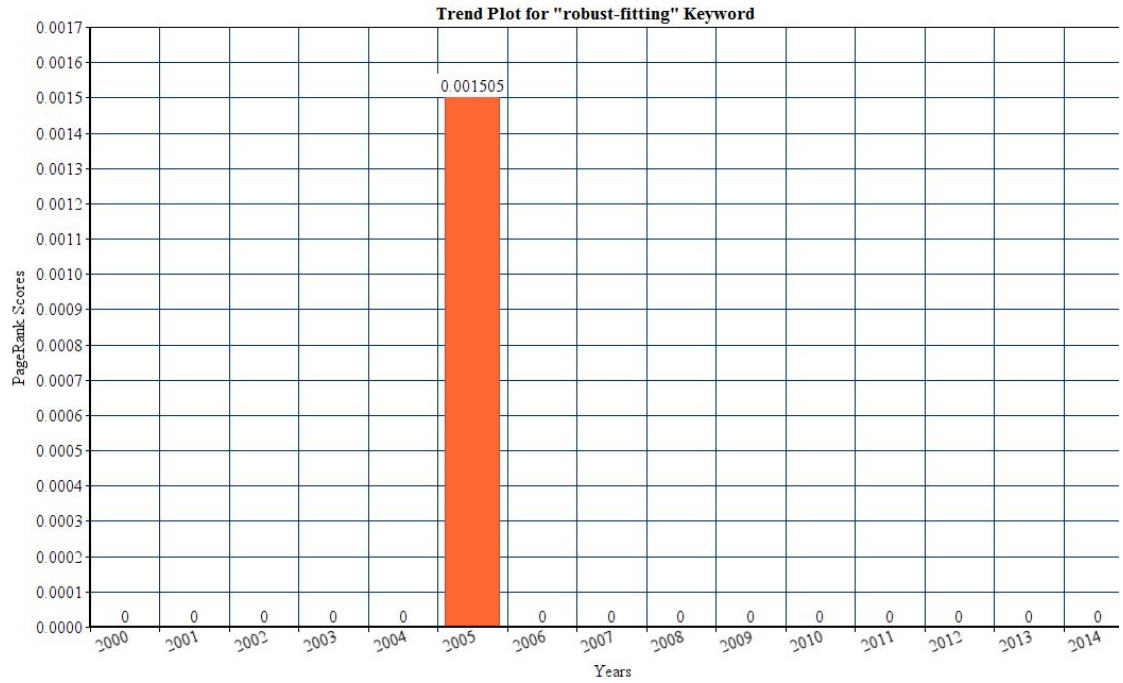

Trend Plot for "real-time-bidding" Keyword
PageRank Scores
Years
0.00103
0.0008
0.00115
0.00110
0.00105
0.00100
0.00095
0.00090
0.00085
0.00080
0.00075
0.00070
0.00065
0.00060
0.00055
0.00050
0.00045
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

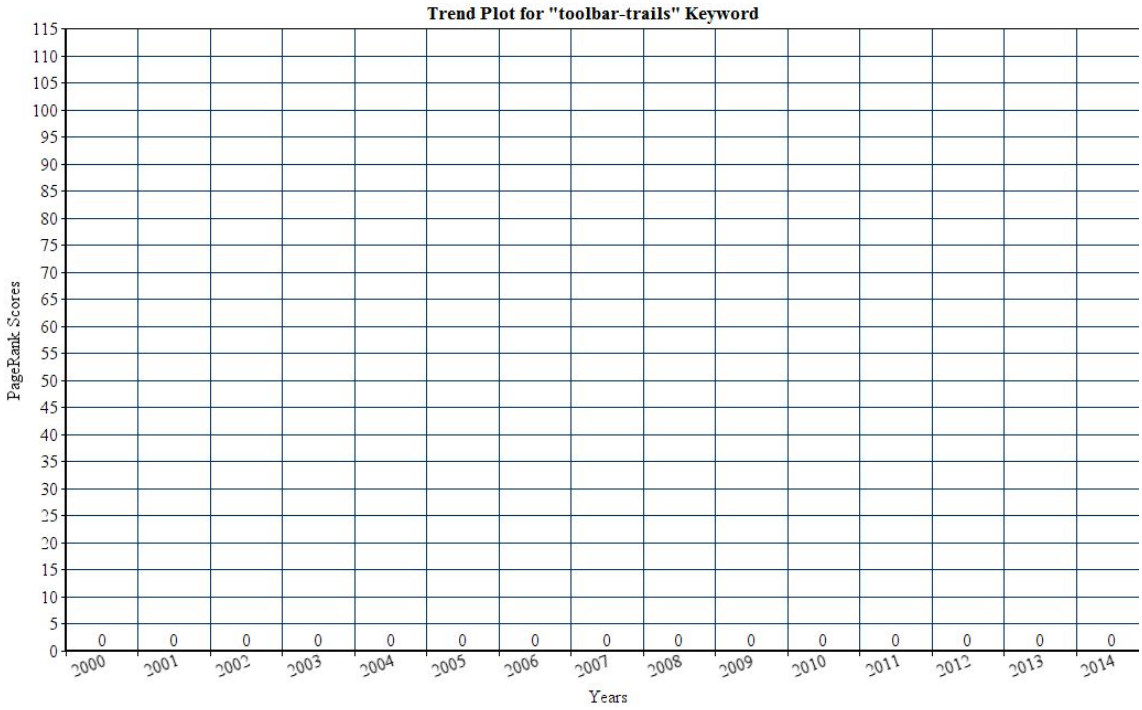

Trend Plot for "polarity-analysis" Keyword
PageRank Scores
Years
0.001602
0.0018
0.0017
0.0016
0.0015
0.0014
0.0013
0.0012
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

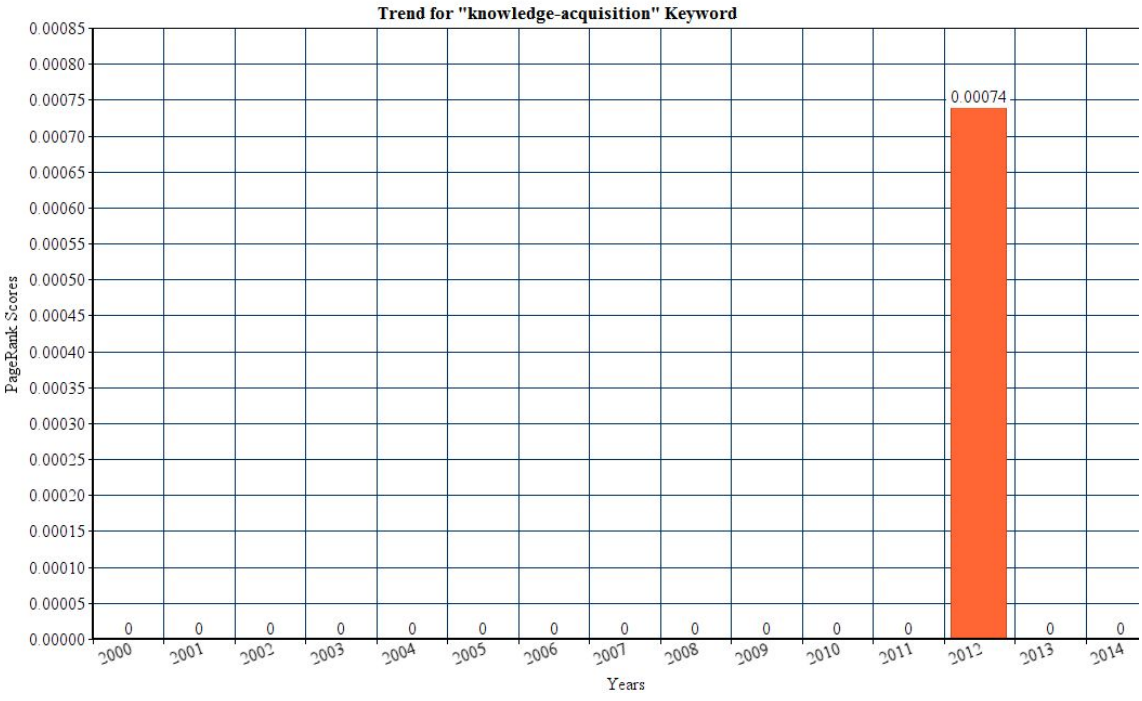

Trend Plot for "algorithmic-advertising" Keyword
PageRank Scores
Years
0.001391
0.000833
0.0016
0.0015
0.0014
0.0013
0.0012
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

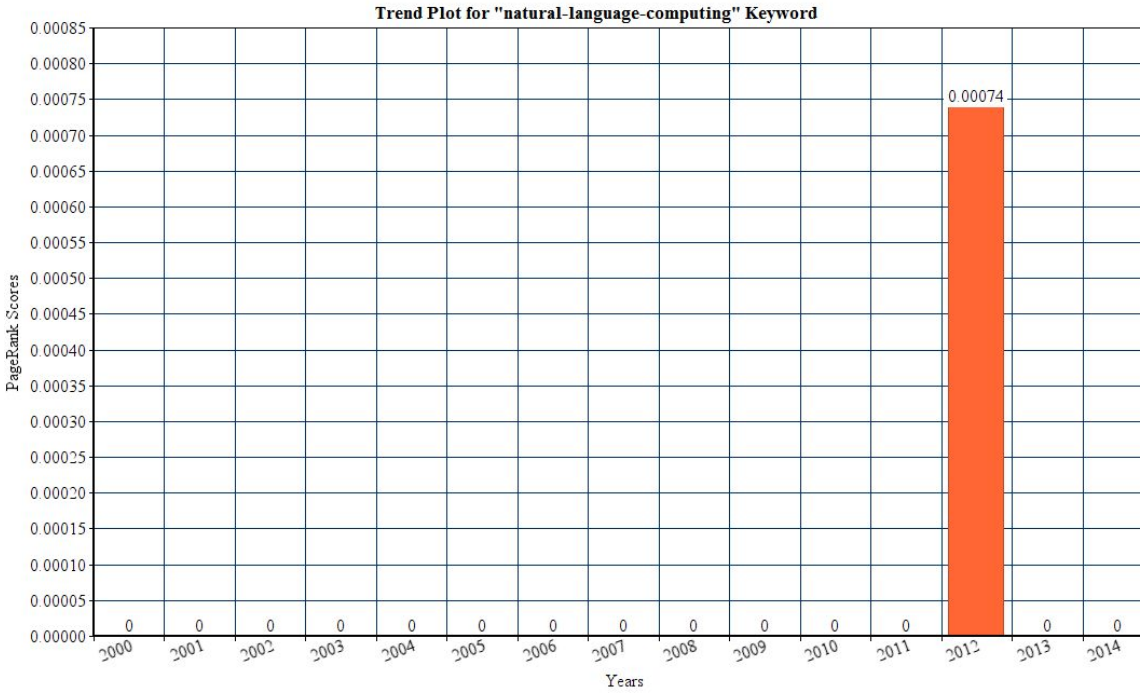
C) Trend Plots for Some Selected Keywords that have low PageRank

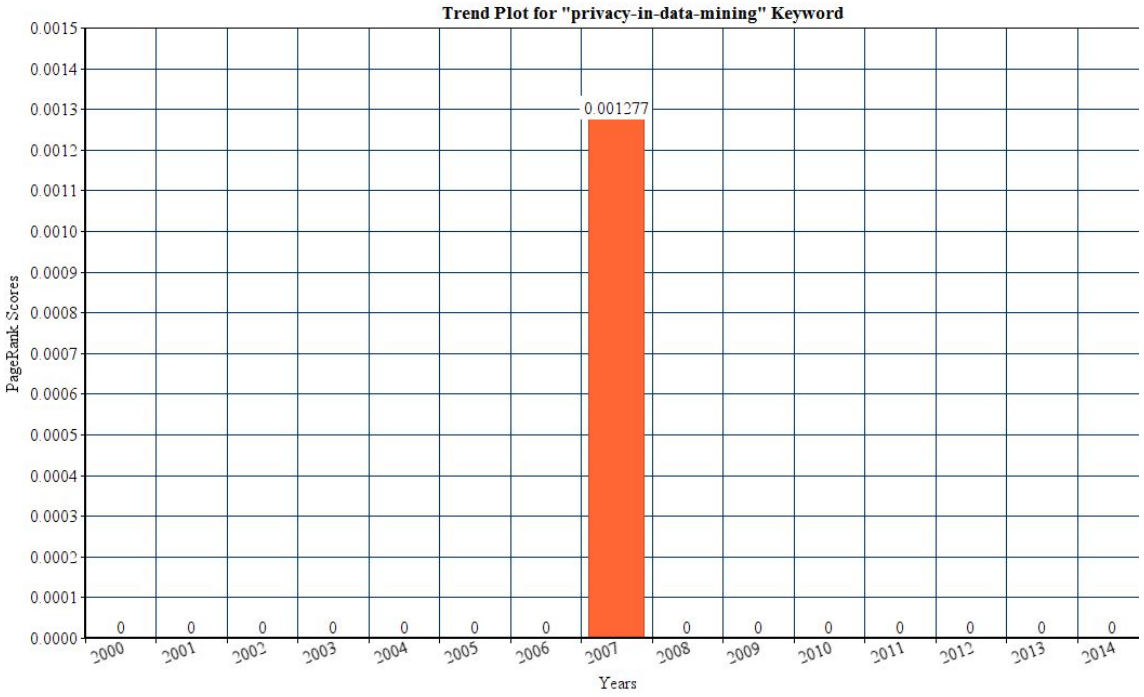


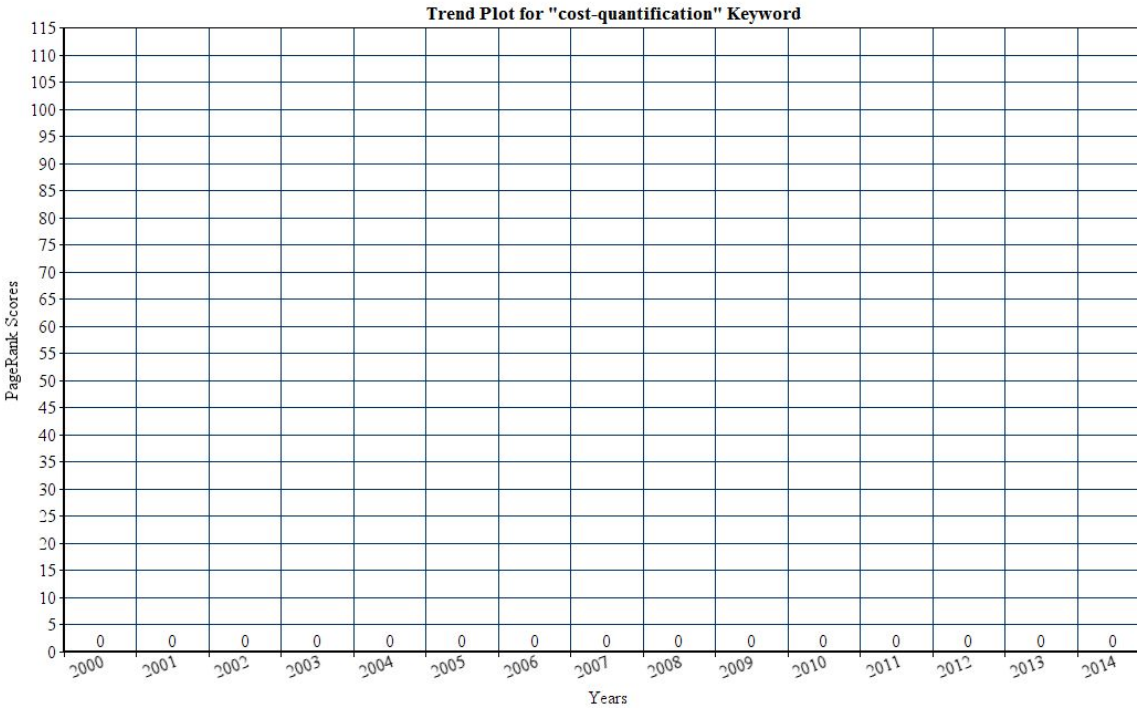


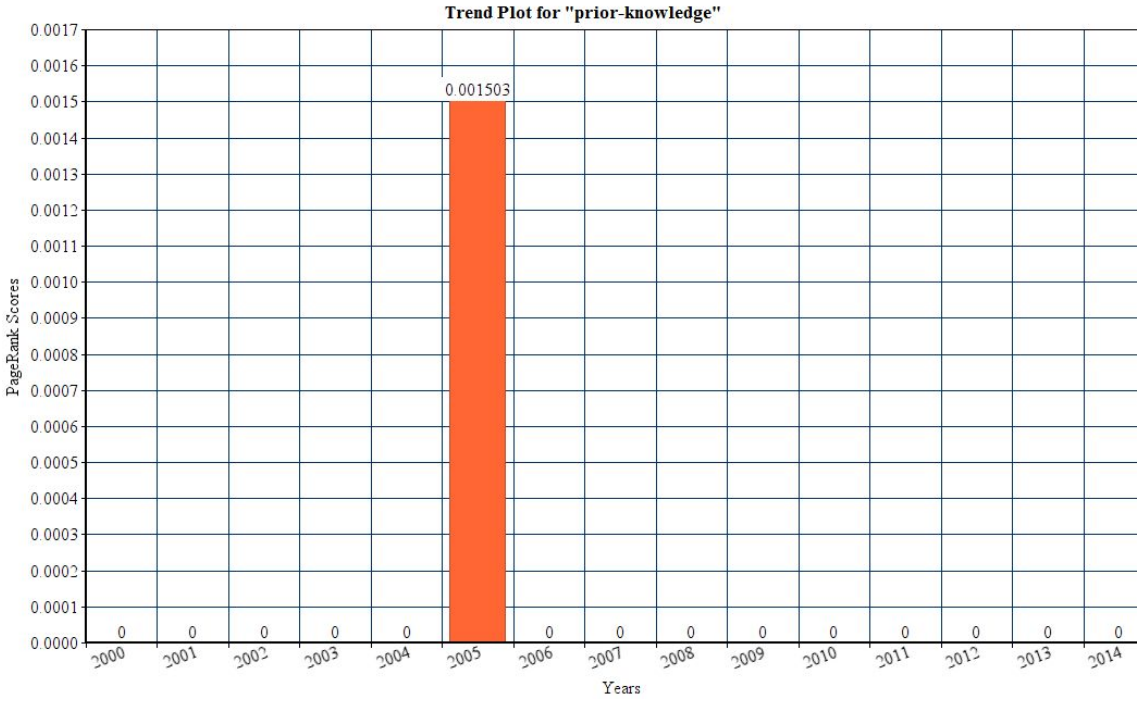

Trend Plot for "toolbar-trails" Keyword
PageRank Scores
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Years


Trend for "knowledge-acquisition" Keyword
PageRank Scores
0.00085
0.00080
0.00075
0.00070
0.00065
0.00060
0.00055
0.00050
0.00045
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0.00074
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Years


Trend Plot for "natural-language-computing" Keyword
PageRank Scores
0.00085
0.00080
0.00075
0.00070
0.00065
0.00060
0.00055
0.00050
0.00045
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0.00074
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Years


Trend Plot for "privacy-in-data-mining" Keyword
PageRank Scores
0.0015
0.0014
0.0013
0.0012
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0.001277
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Years


Trend Plot for "cost-quantification" Keyword
PageRank Scores
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Years


Trend Plot for "prior-knowledge"
PageRank Scores
0.0017
0.0016
0.0015
0.0014
0.0013
0.0012
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0.001503
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Years

APPENDIX E

PAGERANK RANK and SCORES of AKG

Table 8. Vertices Listed in Top-68

Rank	Vertex Label	PageRank Score
1	data-mining	0.0012732709430581897
2	clustering	0.0010301057575731493
3	classification	8.456338242459594E-4
4	social-networks	7.813305361538582E-4
5	text-mining	6.523049411130134E-4
6	machine-learning	5.938739471634679E-4
7	jieping-ye	5.71188772000735E-4
8	christos-faloutsos	5.696042577902038E-4
9	philip-s.-yu	5.54509969920494E-4
10	anomaly-detection	5.478728197270317E-4
11	jiawei-han	5.318816796897532E-4
12	graph-mining	5.124258275749302E-4
13	ravi-kumar	4.277199847856868E-4
14	collaborative-filtering	4.217804089909378E-4
15	time-series	4.205290476874286E-4
16	information-extraction	4.129103125987376E-4
17	feature-selection	4.093905808395928E-4
18	social-network	3.9058382905162895E-4
19	recommender-systems	3.8706525408146693E-4
20	hui-xiong	3.697588560082048E-4
21	active-learning	3.544215495190989E-4
22	insider-threat	3.542608504730144E-4
23	online-advertising	3.451696932749227E-4
24	bing-liu-0001	3.419019117835106E-4
25	huan-liu	3.4148158344898913E-4
26	tao-li	3.3592659007308356E-4
27	social-media	3.317602265718343E-4
28	deepak-agarwal	3.3043878256066363E-4
29	privacy	3.2266478115915045E-4
30	chengxiang-zhai	3.1953072498660807E-4

31	srinivasan-parthasarathy	3.1643833239856326E-4
32	thorsten-joachims	3.1556859009324083E-4
33	support-vector-machines	3.1271914401272106E-4
34	semi-supervised-learning	3.0377552611307384E-4
35	topic-modeling	3.018624381453292E-4
36	web-mining	3.0045967058305565E-4
37	text-classification	2.966773426644717E-4
38	ranking	2.946906091445229E-4
39	heikki-mannila	2.9034131451645417E-4
40	jian-pei	2.852111629700574E-4
41	information-retrieval	2.8465630632164014E-4
42	link-prediction	2.8211256392259693E-4
43	sampling	2.7737638248738787E-4
44	logistic-regression	2.756834905777371E-4
45	naren-ramakrishnan	2.735970100955793E-4
46	scalability	2.6968798059519153E-4
47	large-scale-learning	2.6939505646376337E-4
48	sentiment-analysis	2.6912873949734605E-4
49	vipin-kumar	2.669300694080652E-4
50	evimaria-terzi	2.664139649114918E-4
51	visualization	2.6585770271299184E-4
52	giles-hooker	2.6417512159942954E-4
53	outlier-detection	2.6283852025490734E-4
54	Zhongfei-(mark)-zhang	2.6187223524738783E-4
55	rayid-ghani	2.6173103723294163E-4
56	event-detection	2.6081502605709216E-4
57	martin-ester	2.6012077291096697E-4
58	nikolaj-tatti	2.6003204337585824E-4
59	udo-miletzki	2.5848876073550196E-4
60	ruoming-jin	2.5817940535504693E-4
61	xifeng-yan	2.581492908226886E-4
62	aristides-gionis	2.572374524472866E-4
63	jianyong-wang	2.5669536828856125E-4
64	transfer-learning	2.559010443357059E-4
65	twitter	2.555558375230038E-4
66	spatial-data-mining	2.5447305519634073E-4
67	wei-fan	2.5289817297465575E-4
68	wynne-hsu	2.506938796311164E-4