Journal of Applied and Advanced Research 2017, 2(3): 170–178
doi.: 10.21839/jaar.2017.v2i3.92
http://www.phoenixpub.org/journals/index.php/jaar
ISSN 2519-9412 / © 2017 Phoenix Research Publishers

Research Article – Computer Engineering

The Extraction of Social Networks from Web Using Search Engines

Faranak Salman Mohajer*

Department of Electrical Sciences, Computer & Information Technology, Islamic Azad University of Zanjan, Zanjan Branch, Iran

Abstract

In this paper, our purpose is to create a large collection of related vocabularies and concepts to the user's favorite field (articles, people, conferences, books, etc.) from the available information on the infinite and vast source of web which is expressed in the form of social network. In the other words, we introduced a way to help the researchers to be able to specify their favorite topic in a particular field and by this way, observe and extract the social network of the related concepts to that topic. In order to extract the nodes of this network, we used the sampling of web pages through the Google search engine, text processing techniques, and information retrieval. The topic of the extracted social network in this research is the scientific conferences in the field of computer sciences. In order to evaluate the effectiveness of this method, the extracted network from the results of the search engine is compared with the scientific conferences available in the DBLP[1] database. The obtained results from the social network analysis showed that the extracted network is of very high accuracy.

Key words: social network, extraction, web, search engine, information retrieval, overlap, link of the results, query

Introduction

Information retrieval in web, with several billion Internet pages is a tough work. In recent years, the field of information retrieval has been developed and changed due to the progress made in information technology and computational techniques. It has caused the research on information retrieval to be one of the most exciting and most important areas of research. The process of searching on the Internet, despite the efforts and widespread quest, still suffers from many problems. It could be argued that these problems and limitations are not rooted in the methods and search strategies, but the main cause of it is the current nature of the web; because the current web is full of meaningless data and a search engine does not understand the meaning of data having inserted a subscript, and often conducts the search based on the keywords received from the user.

One of the new research fields in this case is the extraction of social networks from web. The social networks are created and formed from the most intelligent components, i.e., human beings. Due to the various aspects of human beings, there are different types of social networks in all areas and their diversity can vary from physicists' network to doctors' network, and even football fans network. The considered social networks are generally pre-designed networks in which people become a member of them and communicate with each other; Facebook[2] , Twitter[3] , and orkut[4] are examples of such networks. The issue is that these people have selected their friends based on their common interests and these friends are so limited. Now suppose that we want to find relevant people

Received: 17-05-2017; Accepted 08-06-2017; Published Online 10-06-2017

*Corresponding Author

Faranak Salman Mohajer, *Department of Electrical Sciences, Computer & Information Technology, Islamic Azad University of Zanjan, Zanjan Branch, Iran*

[1]Digital Bibliography and Library Project
[2]www.facebook.com
[3]www.twitter.com
[4]www.orkut.com

to a subject through the web or that we want to extract the relevant concepts from the web. For example, we want to find relevant conferences to a special conference such as SIGMOD[5]. It is clear that the extraction of this information from the social network is not possible easily. There is a data source which is not only vast, nut also diverse, dynamic and above all, free and accessible to all: Internet, typically web.

The automatic generation of social network through concepts extraction from web and also the development and growth of semantic web is the main incentive to create such networks in different domains. One of the main problems in this area is the access to a valid and perfect vocabulary collection for the production of such networks. In this research, an automatic method is proposed for the production of social network for research domain in computer sciences, using a sampling method of first pages, text processing algorithms, and also information retrieval techniques. This method is based on the structure of the social network under web and also based on data obtained from the results of the users' search in their favorite fields such as articles, people, conferences, and books related to the subject of the research. One of the objectives of this research is to automatically provide a great collection of vocabulary and main concepts with a high and acceptable accuracy in order to facilitate and accelerate the production of network. For this purpose, relevant pages to the social network have been extracted, using a crawler.

Performed tasks

Kautz (1997) presented an extraction system of social network from web and called it Referral Web. This system was applicable to the concurrence of names on the web [1].

Peter Mika (2005) presented a system for extraction of social networks of a group of researchers in the field of semantic web and called it Flink. In this system, he extracted the social network from various data sources such as web pages, publications archive, and profiles created by the people. In addition, in order to enhance the accuracy of the results returned by the search engines, the keyword 'semantic web or ontology' was added to the query [2].

Matsuo et al (2007) presented a system for extraction information from the web and called it POLYPHONET. This system applies different advanced technologies to extract the relationships between people in order to identify the groups and obtain keywords for a person. The method is based on the concurrence of two persons' names that can be in a line or in a few lines. In order to calculate the cooperation between two people, a number of results found by the search engine are used for each name and composition of two names [3].

Bekkerman & McCallum (2005) designed a system to extract a social network. This system selects certain people by e-mail and finds their personal web pages and then prepares a list of their specifications such as phone book. Then a relationship is established between the owners of personal pages and people who have been identified within the pages [4].

Harada et al (2004) designed a system for extracting the relationship between people from the web [5].

Culotta et al (2004) presented a system for extracting social network in which the existing information in the e-mail inbox and people's contact addresses are used. In this method, a set of keywords is extracted for each person and his relationships with other people are specified using the keywords [6].

Faloutsos (2004) created a social network containing 15 million people from 500 million web pages, using their concurrence through a slippery 10-words-frame in the texts [7].

Adamic (2003) collected the relationship between students through the structure of web connections and textual information, using social networks of M.I.T and Stanford Universities [8].

Aleman & Meza (2006) extracted and presented a social network by combining two techniques including 'knowing 'in FOAF and 'co-author' of the DBLP database [9].

Tang et al (2007) attempted to extract the researchers' social network. They created profiles for each person and through this, extracted the

[5]Special Interest Group on Management of Data

people associated with him [10]. Mahyuddin & Shahrul (2011) presented a method to extract the social network of academic researchers. They extracted the names of people and their relationships using online DBLP database and the mathematical relationships of the collections and graphs [11].

Yun Hong et al (2012) proposed a method for analysis and scientific extraction to the researchers by combining social networks and semantic analysis of concepts. In this method, data is collected and filtered from the collection of concepts sand massive data on the web such as scientific websites, researchers' personal pages, social networks' websites, and then is created in the form of a twofold network model which contains a layer of concepts and a layer of researchers; accordingly, a number of recommendations of the study is presented to the user [12]. Ferrara et al (2014) provided an overview of data extraction from the web. They divided the data extraction applications into two categories: the organizational category, and the social web. In the organizational category, the data extraction techniques from the web are used for data analysis in business and intelligent competitive systems. These techniques are useful in the social web category to gather a large amount of structured data which is constantly being produced and published [13].

In the field of the identification of concurrence of communication in social networks, Yu Xin *et al* (2015) proposed a method called SLW, based on the determination of weight for links between nodes which evaluates the weight of each current edge between the nodes of each model and accordingly, the concurrence of relations based on the current weight is obtained [14].

Siersdorfer *et al* (2015) presented a new method for efficient extraction of social networks from the web. For this purpose, they use patterns in the queries and develop these patterns by the help of Bootstrapping method. Their results showed that the proposed algorithm offers high quality results in large scales. This method is used for social network extraction of influential people [15].

Armentano et al (2014) presented a search engine which collects the published information on the web based on the academic websites and researchers' personal pages; then, it processes the considered textual results based upon the methods of natural language processing, extracts the useful information in brief and returns as a result for non-technical users [16].

The purpose of the research

This research aims to create a social network using the existing information in web pages. This network plays an important role in reducing the volume of information and increasing the speed of access to meaningful information. Moreover, it can be a user-friendly environment for people; because through the providing of interest for the user, the possibility of providing relevant and meaningful results to the user's search would be opened up at a right time and would improve the process of search and accelerate the speed of finding accurate and relevant answers to the needs of the users on the Internet.

The method in this research is that at first the considered query of the researcher is given to the search engine as the primary seed and the textual contexts of search results are extracted through exploratory software, designed for this purpose; then, the nodes and social network relationships are created using information retrieval and text processing mechanisms.

Research methodology

The subject of social network

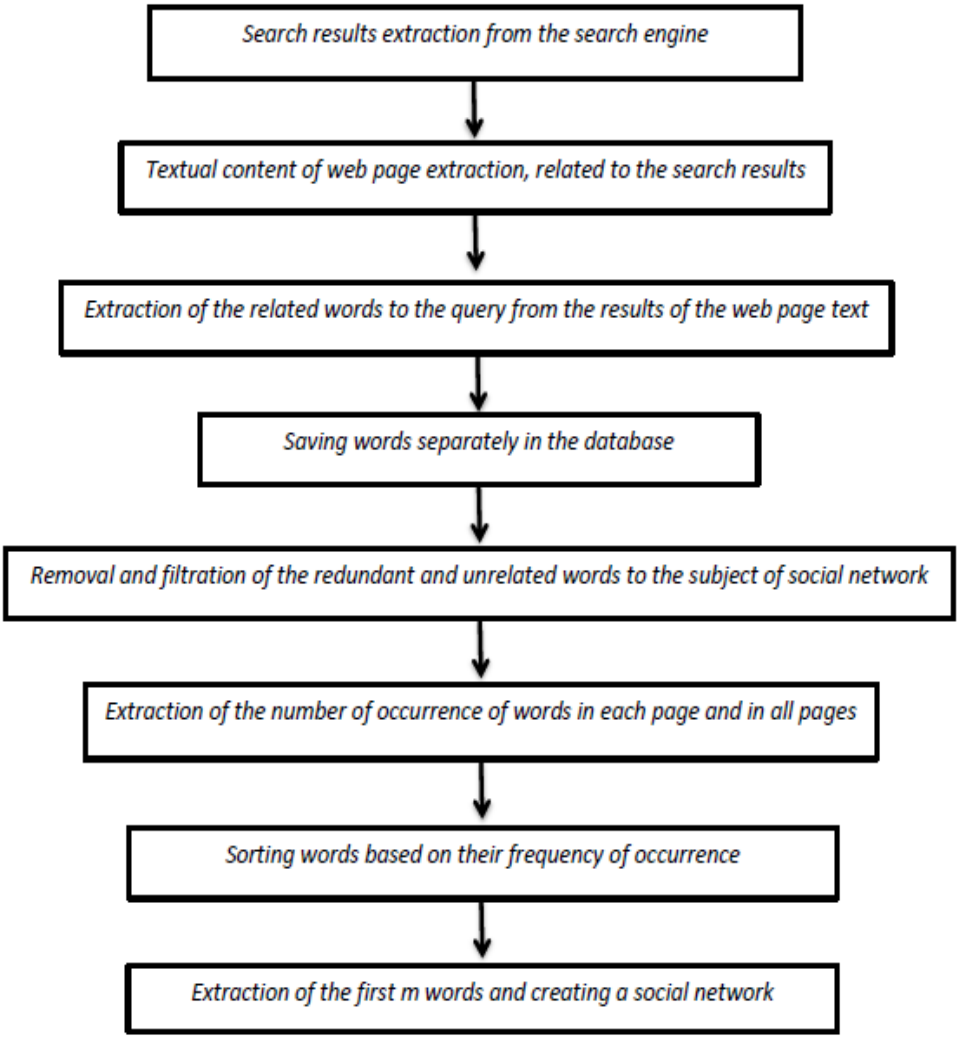
The subject used in this study is the name of scientific conferences on the field of computer sciences and to present the name of conference into the search engine in order to retrieve the considered information from the DBLP database, which is a large database of researchers and scientists in the field of computer sciences and a collection of research records and names of the scientific conferences in the field of computer sciences and has about 3 million research records (DBLP, July 2015) [17,20]. SEKE[6], ICSE[7], ICCIT[8] are among the queries used in this research.

[6]International Conference on Software Engineering and Knowledge Engineering

[7]International Conference on Software Engineering

[8]International Conference on Computer Science and Information Technology

Figure 1. The process of social network nodes extraction from the search engine



The formulation of query

Regarding the searched words in the context, some words may be assed to the query for more accurate search. In this research, since the aim is to extract the relevant links to the name of the conference, the keyword 'conference' was added along with the name of the conference. For example, if we consider the selected query a conference entitled 'SEKE', then the user's query is modified and converted into 'SEKE Conference'.

The selection of search engine

In this research, the Google search engine is used for extracting information from the web' this search engine accounts for 63.9% of people's use of search engines and is the most widely used search engine at the present time (comScore qSearch Data, November 2015) [21].

The number of extracted links

The number of extracted links from the search engine results is considered the first 20 links of results; because approximately 89.8% of clicks made on the search engine results were in the first pages of results and the users often check the first pages of search engine results [18, 19].

The extracted social network verification

For this purpose, the DBLP database is used. DBLP data in the form of file sand with XML structure can be downloaded from the Internet.

This large collection of research records of all authentic world conferences is built up and numerous researchers have used this dataset in their scientific articles in order to test and evaluate their algorithms.

The process of social network extraction from the search engine results

Figure 1 shows the process of social network nodes extraction from the search engine results. After submitting the query to the Google search engine, the first 20 links of the results are extracted buy the software that has been designed for this purpose.

The extraction of the textual content of web pages related to the search results

At first, by conducting the survey of results page, the link of the results along with the textual parts and their ratings are saved in the database and the crawler begins to extract the contents of pages with certain links, considering the URL related to the first n returned results. In this step, in order to reduce the volume of web pages, uploading audio, image, video, flash or scripts should be prevented and only textual contents and HTML of the pages should be extracted. 'No Flash' software is used for this purpose. After the considered pages were fully loaded, the textual contents and their HTML would be saved in the database designed for this purpose in order for the further processing to be conducted on them. The content of these pages are saved in the form of HTML tags. The problem that can take place in this step is that there may be error in the extraction or loading of some pages; these errors should be identified and rectified. In the cases where it is not possible to rectify the error, we should proceed to remove these pages from the search list. For example, the textual content of the considered link might be removed or the considered website addresses do not exist anymore; therefore, it is possible to extract the first 20 results from the search engine results, but during the loading of related web pages, we realized that it is not possible to save the textual content of some pages.

The extraction of related words to the query from the text of results web pages

In order to extract the concepts from a retrieved document for a query, we need to identify the keywords in the document. To extract these concepts, we require the text processing steps include the following:

- Preparation and pre-processing data
- Lexical separation
- Removal of common words
- Removal of words with lowercase letters

Preparation and pre-processing data

In this step, the content of HTML is converted into textual content. To do this, the HTML or XML tags are removed from the saved text in the database. By doing this step and removal of all existing tags, the remaining text is saved in another field of the same table in the database.

Lexical separation

In this step, a parser scrolls the text and identifies the words' separating marks such as empty space and marks at the end of a sentence such as '!', ',', '؟', '?', '.'; then, identifies, separates, and saves words in the database.

Removal of common words

The common words are words that have been repeated in the text and the removal of them does not have a role in the determination of the keywords of the text. These words have no value in the context and include words such as pronouns, adverbs, prepositions, conjunctions, and some repeated words that are very common in any text.

Removal of words with lowercase letters

Since our considered case is the social network of conferences, so in the technique to extract this social network, the names of the conferences are written in abbreviation form and with uppercase letters. As a result, in this survey all words that are written with lowercase letters are removed from the database.

Removal and filtration of the redundant and unrelated words to the subject of social network

There are a lot of words in a text that are not involved in the considered social network of the user. These words should be removed from the extracted text. The method used in this step can be different; it means that in the social network of people, we should follow the names of people in the extracted words. Since we aim to create a

social network of conferences, the following should be done in this step:

- Removal of existing words in the dictionary
- Removal of unrelated words to the name of the conference

The removal of existing words in the dictionary

As we know, all the existing words in the English Language are available in the dictionaries and there is no word in the present massive dictionaries that cannot be found. Consequently, the database of current dictionaries is a good source for us to identify the English words. Since we aim to find the name of the conference from the extracted text, and with this supposition that the name of the conference is an English abbreviated name therefore this name cannot be found among the existing words in the dictionary. As a result, all existing words in the processed text are compared with the existing words in the dictionary. If a considered word is not found in the dictionary, it means that this word is not of English words and can be a part of our considered words.

The removal of unrelated words to the name of the conference

A lot of words, especially English abbreviated words, are not parts of the names of the conferences. These words are not of English words and do not exist in the dictionaries; therefore, they are not removed in the step of comparison with the dictionary. That solution that can be represented for the removal of these words is to add these words to the dictionary database; so if they exist in the text, they can be removed in this step. Words ISBN or WWW are such words.

Extraction of the number of occurrence of words in each page and in all pages

The frequency of each word in our current list of retrieved pages by the search engines can be an indicator of the importance of that word. If a word from the list had a high repetition in the presented results in the search, it can be added as the subsequent node to the primary node in the social network structure. There are two modes for the number of repetition of a word:

- The number of repetition of the considered word in the first n results of the search engine
- The total number of repetition of the considered word in all pages

The number of repetition of the considered word in the first n results of the search engine

It means that each word in our list of wards have been found in the few results of the first 20 results. For example, a word can be found in the 8 results of the first 20 results. Therefore, the number 8 is the number of repetition of this word in the presented results by the search engine.

The total number of repetition of the considered word in all pages

It means that how many each word in our list had repeated in each page and in the following, how many the sum of them in all pages is. For example, if a word is repeated in the 8 results of the first 20 results, we should consider the number of that word's repetition in each of 8 mentioned pages and then calculate the sum of it.

Sorting words based on their frequency of occurrence

Due to the fact that the purpose is to extract the keywords as nodes of the social network which are related to the primary node (primary query), so the words should be prioritized and sorted based on their frequency in the results pages. In this case, the words which have high frequency are guided to the top of the list and create the social network as the part of related nodes to the primary node. In order to sort the words, at first we sort them based on the frequency of the considered word in the first few results of the search engine and then prioritize them based on the sum of frequency of the considered word sorted in all pages; because if a word could be found in a lot of pages retrieved by the search engine, it would have higher priority.

Extraction of the first 9 words and creating the social network

In this step, due to the selected level of association, which is considered 9 in this research, we should select the first 9 words of the sorted list as the nodes of first depth which are directly

connected to the primary node in the social network.

Creating social network in a greater depth

In order for the created social network to be expanded and the number of nodes and user options to be added, the social network can be developed by all the above methods. Of course it should be considered that by adding too much depth to social network, more time is certainly needed for the system calculations and sometimes too much development of the network not only has no practical use, but also it will make the network complicated and confuse the user.

Therefore, we have extended the social network in this research to a depth of 2 and the obtained results were also acceptable. After creating the first depth in the social network in order to create the second depth, each of words or the very same nodes of the first depth should be given to the system as the initial seed; so, all steps would be performed for each of them and the social network with a depth of 2 is created.

Data analysis

Thirty names of the conferences were extracted from the DBLP database and were given to the crawler software as the initial seed. A social network with the level of association of 9 in the depth of 2 was extracted from the search engine per each initial seed. Therefore, the number of extracted social networks is 30. The figures 2 and 3 show the created networks, with SEKE query in the depth of 1 and 2.

Table 1. The number of extracted nodes

Number of nodes	1 query	30 queries
Number of extracted nodes in the depth of 1	9	270
Number of extracted nodes in the depth of 2	81	2430

Table 2. The number and percentage of concurrence of the extracted nodes and DBLP

Extracted Nodes	Number	Percentage
Were found in DBLP	2041	84%
Were not found in DBLP, but are the names of conferences	263	11%
Are not the name of the conference	126	5%

Figure 2. The created network for SEKE query in the level of association of 9 to the depth of 1

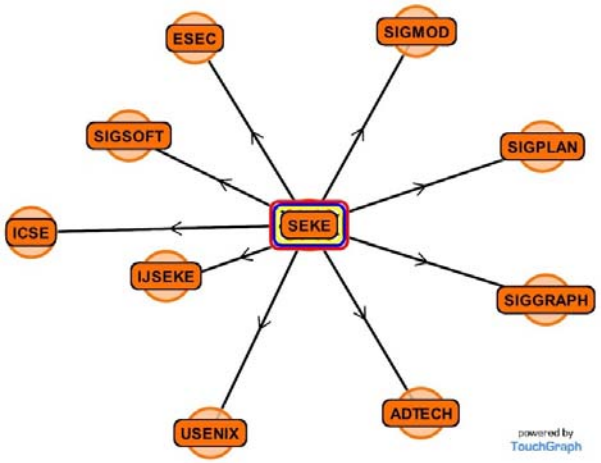
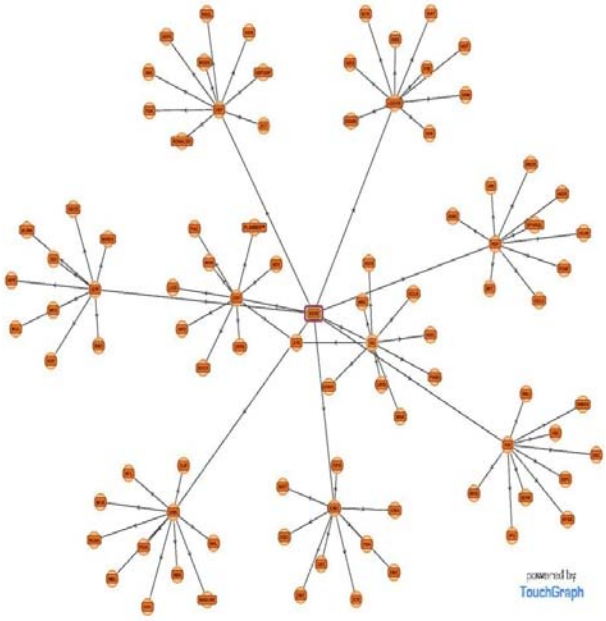


Figure 3. The created network for SEKE query in the level of association of 9 to the depth of 2

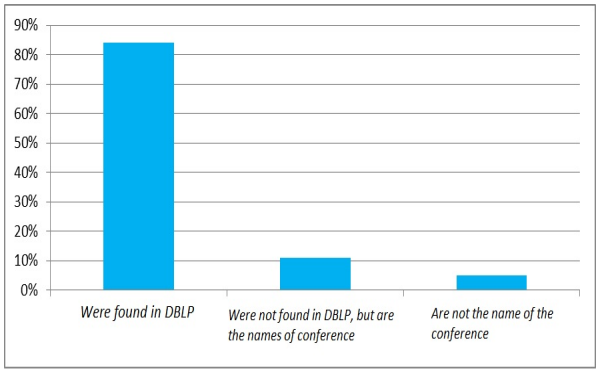


As shown in table 1, the number of extracted nodes for a query to the depth of 1, is 9 and to the depth of 2 is 81; therefore, for 30 queries, the number of nodes in the depth of 1 is 270 and in the depth of 2 is 2430.

By extracting the social network for 30 queries and comparison of each of nodes with DBLP database, according to table 2 and diagram 1, it was found that 2041 items of 2430 extracted nodes were available in DBLP database; 263 items were

not found in DBLP but they were the names of the conferences and ultimately only 126 items were not the names of the conferences and were considered error.

Diagram 1. The percentage of concurrence of the extracted nodes and DBLP



Conclusion

According to the obtained results, it was specified that social networks can be created from the available pages on the web and using search engines; because the information in the web pages are highly rich and each of social networks in different subjects can be extracted from the free and vast source of web, using the proposed techniques; this network will be a dynamic network due to the information update in the web and new links and nodes will be added and removes. Meanwhile, these networks are directed because it is possible that the node A be a part of top 5 related nodes to the B, but the node B is not a part of top 5 related nodes to the a. Moreover, due to the fact that the percentage of error in the extraction of social networks' nodes was only 4 percent which is very low; therefore, it is a proper and reliable method for researchers in finding their considered names of conferences.

References

1. Kautz,B.Selman,M.Shah, The hidden Web, AI Mag v18 , n2 (1997), pp 27–35
2. Mika Peter, "Flink: Semantic web technology for the extraction and analysis of social networks", Web Semantics: Science, Services and Agents on the World Wide Web, v3 (2005), pp 211–223.
3. Matsuo Yutaka, Mori Junichiro, Hamasaki Masahiro, Nishimura Takuichi, Takeda Hideaki, Hasida Koiti, Ishizuka Mitsuru, "POLYPHONET: An advanced social network extraction system from the Web", Web Semantics: Science, Services and Agents, on the World Wide Web, v5 (2007), pp 262–278.
4. Bekkerman Ron, McCallum Andrew, "Disambiguating Web appearances of people In a social network", WWW '05 Proceedings of the 14th ACM International World Wide Web (2005), pp 463-470.
5. Harada Masanori, Sato Shin-ya, Kazama Kazuhiro, "Finding authoritative people from the Web", JCDL'04, (2004), pp306 – 313.
6. Culotta Aron, Bekkerman Ron, McCallum Andrew, "Extracting social networks and contact information from email and the Web", American Association for Artificial Intelligence.
7. Faloutsos Christos, S. McCurley Kevin, Tomkins Andrew, "Fast discovery of connection subgraphs", KDD'04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (2004), pp 118-127.
8. A.Adamic Lada, Buyukkokten Orkut, Adar Eytan, "Social network caught in the Web", First Monday Peer-Reviewed Journal On The Internet (2003), v8, n6.
9. Aleman-Meza Boanerges, Nagarajan Meenakshi, Ramakrishnan Cartic, Ding Li, Kolari Pranam, P. Sheth Amit, Arpinar I. Budak, Joshi Anupam, Finin Tim, "Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection", WWW '06 Proceedings of the 15th ACM international conference on World Wide Web (2006), pp 407-416.
10. Tang Jie, Zhang Duo, Yao Limin, "Social Network Extraction of Academic Researchers", ICDM 2007. 7th IEEE International Conference on Data Mining (2007), pp 292-301.

11. Mahyuddin K. M. Nasution, Shahrul Azman Noah, "Extraction of Academic Social Network from Online Database", IEEE International Conference on Semantic Technology and Information Retrieval (STAIR) (2011), pp 64-69.
12. Yunhong Xu, Xitong Guo, Jinxing Hao, Jian Ma, Raymond Y.K. Lau, Wei Xu, "Combining social network and semantic concept analysis for personalized academic researcher recommendation", Decision Support Systems, v54 (2012), pp 564–573.
13. Ferrara Emilio, De Meob Pasquale, Fiumara Giacomo, Baumgartner Robert, "Web data extraction, applications and techniques: A survey", Knowledge-Based Systems (2014), v70, pp 301-323.
14. Yu Xin, Jing Yang, Zhi-Qiang Xie, A Semantic Overlapping Community Detection Algorithm Based on Field Sampling, Expert Systems with Applications (2015), v 42, Issue 1, pp 366–375.
15. Siersdorfer Stefan, Kemkes Philipp, Ackermann Hanno, Zerr Sergej, "Who With Whom And How? - Extracting Large Social Networks Using Search Engines", CIKM ’15 Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (2015), pp 1491-1500.
16. G. Armentano Marcelo, Godoy Daniela, Campo Marcelo, Amandi Analia, "NLP-based faceted search: Experience in the development of a science and technology search engine", Expert Systems with Applications (2014), v41, pp 2886–2896.
17. Alwahaishi Saleh, Martinovic Jan, Snasel Vaclav, Kudelka Milos, "Analysis of the DBLP Publication Classication Using Concept Lattices", Dateso (2011), pp 132-139.
18. J. Jansen Bernard, Spink Amanda, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs", Information Processing and Management v 42 (2006), pp 248–263.
19. Spink Amanda, J.Jansen Bernard, Blakely Chris, Koshman Sherry, " A study of results overlap and uniqueness among major Web search engines ", Information Processing and Management , v42 (2006), pp 1379–1391.
20. http://dblp.uni-trier.de, (1394/9/14)
21. http://www.comscore.com, (1394/9/14)