International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 4, No. 2, 2017, pp. 135-144.
ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

The overlap among Social Networks Extracted from Major Web Search Engines

Faranak Salman Mohajer

Master in Computer Engineering, Faculty of Electrical Engineering, Computer and Information Technology, Islamic Azad University, Zanjan, Iran

Abstract

The performance and functionality of web search engines is one of the popular issues considered by researchers. One of the new research areas is extraction of social networks from website by search engines. This study aimed to evaluate the overlap among three Google, Yahoo, and Bing search engines in extracting social networks. For this purpose, the overlap among search engines results links was investigated. It was expected that the overlap would be very low. Then, the textual content of links was extracted and texts were processed to extract the related words to queries. These words are the nodes of our social networks. By conducting this process for each of the search engines and extracting social networks from search engines, the overlap of these social networks can be evaluated. The findings showed that the percentage of unique links in search engines is 59 percent and the percentage of unique nodes in social networks is 26 percent. The percentage of shared links in two and three search engines is 20 percent and 3 percent and the shared nodes are 37% and 21%, respectively. This indicated higher overlap among search engines in extracting social networks than overlap among search engines in results links.

Keywords: Social Network, Extraction, Web, Search Engine, Information Retrieval, Overlap, Results Links, Inquiry, Node.

1. Introduction:

Today, the Web is considered as the first source for information among many people and more than 80% of web searches are conducted using search engines [1]. Daily, millions of users use search engines. Therefore, the performance and capabilities of search engines to retrieve information is one of the issues are considered by researchers in scientific research [2]. For this reason, many attempts have been conducted in recent years to improve retrieval of information from web. Also, the ability of search engines has been increased every day to deliver better results. The web search processes suffer from one problem: current nature of web. The current web is dominated by meaningless data.

One new research area is social networks extraction from web. The social networks are consisted of most intelligent elements; i.e. human beings. Due to differences among people, there are a variety of social

networks in all areas which range from physicists networks to doctors networks and even football fans networks.

When user enters the keyword in a search engine and different results are provided in different search engines, the user will get confused about which of the search engine results are more accurate and more appropriate response to his/her query (2). This study aims to evaluate the overlap among search engines in finding social networks. In first stage, the social network is extracted by search engine. The social network data may be extracted from web search engine results.

Then, the social network are made by three Google, Yahoo, and Bing search engines separately. These social networks are compared with each other to determine the overlap of three mentioned search engines in extracting social networks. It is expected that the overlap to be high. Therefore, the social network will be more reliable source for users.

Ding and Marchionini (1996) investigated the search results of three Infoseek, Lycos, and OpenText search engines in five inquiries. They found that there is very little overlap among these search engines [3].

Baharat and Broder (1998) used random search to investigate the overlap among search engines. They showed that the overlap among Hotbot, AltaVista, Excite, and Infoseek search engines is 1.4% [4].

Sterling (2007) investigated four Google, Yahoo, Ask Jeeves, and MSN search engines. In this study, 19332 queries were entered to search engines and the links provided on the first page of search engines were compared with each other. The results showed that the number of unique links was 88.3% and overlaps between two search engines, three search engines, and four search engines was 8.9%, 2.2%, and 0.6%, respectively [7].

Esfandiari Moghadam and Bahari Movafagh (2012) investigated four Yahoo, Live Search, Google, and Ask search engines. In this study, 10 queries in medical field were entered to search engines. The results showed that Yahoo and Google had the highest unique results and Google had the most overlap with other search engines. Also, the overlap among four search engines was 11% [12].

Peter Mika (2005) provided a system of extracting social networks for semantic web researchers and called it (Flink). In this system, the social network was extracted from various information sources such as Web pages, emails, publications archive, and profiles created by individuals. Also, the keywords (semantic web or ontology) were added to queries to enhance the accuracy of search engines results [14].

Harda and colleagues (2004) designed a system for extracting relationships among individuals from the web [17].

Kuluta and colleagues (2004) provided a system for extracting the social network of individuals. For this purpose, the information of email box and contact addresses were used. In this method, a set of keywords are extracted for each person and the relationship between individuals are determined by these keywords [18].

2. Methodology:

The DBLP database (Digital Bibliography and Library Project) was used to extract the names of 31 scientific conferences in the field of computer science. The DBLP is a database of computer authors and scientists and has a set of research records in computer science. It has been collected at University of Tier and has around 3 million records (DBLP, July 2015) [29].

The IGI, ICSR, SEKE, ISC, ICSE, SIGMOD, ESEC, ICSC, SIGPLAN, ICDE, APA, CFP, SIGSOFT, INSPEC, CCC, MLA, ADTECH, SPIE, KSI, SIGGRAPH, HIPAA, ETAPS, ICCS, IDC, SIGIR, SCI, ICCI, SIAM, ACS, USENIX, and SIGACT conferences were used as inquiry.
In this study, the three major web search engines were selected: Google, Yahoo, and Bing. According to Table 1, they are the most used search engines at present (97.3%) (ComScore qSearch Data, November 2015).

3. Analysis of data

Providing 31 queries to each of search engines and extracting 20 first results links, the number of extracted links in each search engine was 620 links and in three search engines was 1860 links. After extracting all links, they should be compared to determine the overlap of links in search engines.
According to Table 1, from 620 extracted links in Google, Yahoo, and Bing search engines, 524, 302, and 294 links were unique.

Table 1. Number and percentage of unique links in search engines

Search engines	Number of results	Percentage of uniqueness
Google	524	%84
Yahoo	302	%49
Bing	294	%47

According to Table 2, comparing the extracted links in two Google and Bing search engines, it can be seen that only 52 of 620 links were shared in these search engines. The number of shared links in Google and Yahoo search engines was 44 links; the shared links in Yahoo and Bing were 274 links. Finally, comparing the links of three search engines, it can be seen that only 24 links of 620 links in any search engine was shared with other search engines.

Table 2. Number and percentage of shared links in search engines

Search engines	Number of results	Percentage of shared links
Google – Bing	52	%8
Yahoo – Bing	274	%44
Google – Yahoo	44	%7
Google – Bing _Yahoo	24	%3

According to table 3, the number of unique links in three search engines was 1120 links from 1860 links, 370 links in two search engines, and 24 links from 1860 links in three search engines.

Table 3. Number and percentage of shared links in one, two, and three search engines

Search engines	Number of results	Percentage of unique links
Unique links	1120	%59
Two search engines	370	%20
Three search engines	24	%3

Therefore, it can be concluded that the provided links in each search engine for user queries are very different from provided links in other two search engines. Therefore, the overlap is very low among results links provided by search engines.

To investigate the overlap among search engines in extracting social networks, the queries were entered to search engines as primary nodes and first 20 links were extracted. Using social networks extraction method, 30 first words were connected directly as first depth nodes of social network to primary node. Then, the obtained nodes (maximum 30 nodes) were again entered to search engines and first 20 links and first 30 words were extracted from each search engine; therefore, second depth nodes were created. It should be noted that it is possible the number of extracted words from 20 results links to be less than 30. In other words, it is possible that some pages do not load or the remained words will be less than 30 words in loading pages by applying keyword extraction method. To coordinate with investigation of overlap among search engines links, however, 31 queries were entered to search engines and first 20 links were extracted.

In social network extraction process, 31 queries were entered to search engines and 30 words were extracted for each of the queries; therefore, the maximum number of extracted nodes will be 930 nodes. Due to mentioned reasons, however, the number of nodes retrieved from each search engine is as Table 4.

Table 4. Number of nodes extracted from each search engine

Search engines	Number of nodes
Google	846
Yahoo	751
Bing	773

Figure 2 shows the created network by Google search engine which has expanded to depth two.

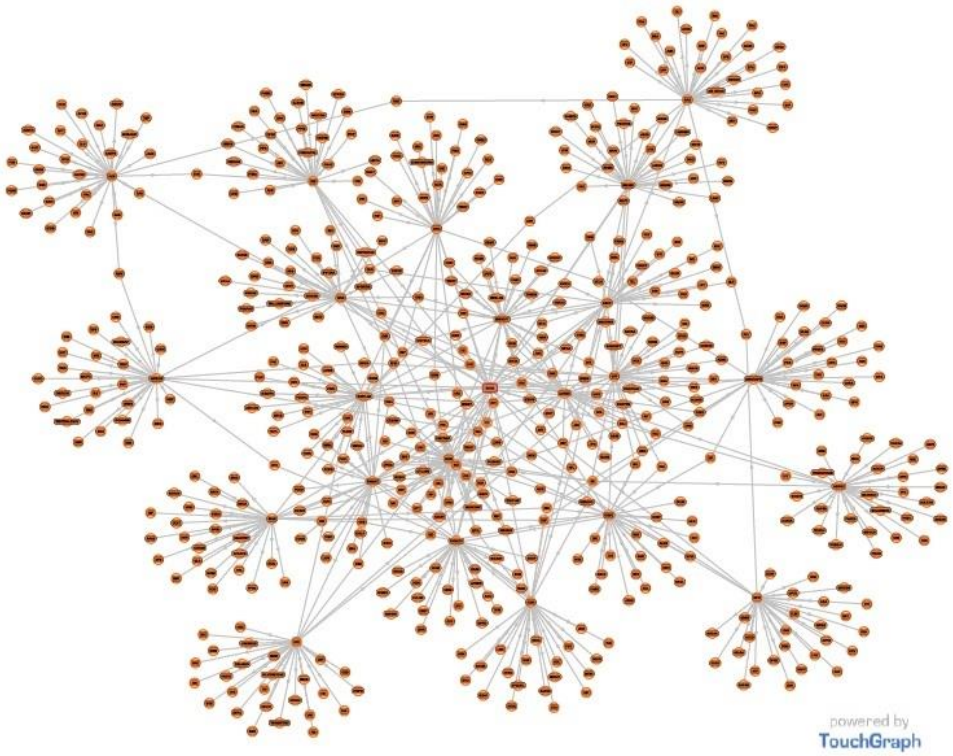


Figure 2. The created network by Google search engine

According to table 5, from 864 extracted nodes in Google search engine, 375 nodes were unique; from 751 extracted nodes in Yahoo, 121 nodes were unique; and from 773 extracted nodes in Bing, 140 nodes were unique.

Table 5. Number and percentage of unique nodes in search engines

Search engines	Number of results	Percentage of unique nodes
Google	375	%44
Yahoo	121	%16
Bing	140	%18

According to Table 6, comparing the nodes extracted from two Google and Bing search engines, it can be seen that about 237 nodes extracted from each search engine were shared. In Google and Yahoo search engines, the number of shared nodes was 234 nodes; in Yahoo and Bing, the number of shared nodes was 396 nodes. Finally, comparing the nodes extracted from three search engines, it can be seen that 166 nodes in each search engine were shared with other search engines.

Table 6: Number and percentage of shared nodes in search engines

Search engines	Number of results	Percentage of shared nodes
Google – Bing	237	%29

Google – Yahoo	234	%30
Yahoo – Bing	396	%52
Google – Bing _Yahoo	166	%21

According to table 7, the number of unique nodes in three search engines was about 636 nodes; the number of shared nodes in two search engines was 867 nodes; and the number of shared nodes in three search engines was 166 nodes.

Table 7. Number and percentage of unique and shared nodes in one, two, and three search engines

Search engines	Number of results	Percentage of unique nodes
Unique node	636	%26
Two search engine	867	%37
Three search engine	166	%21

Now, the shared links of search engines and social networks' nodes may be compared with each other. The percentage of unique nodes compared with unique links (Table 8 and Figure 1) has decreased in three search engines. For example in Google search engine, the unique links percentage has decreased from 84 percent to 44 percent in unique nodes; this is significant and shows an increase in level of overlap.

Table 8. Percentage of unique results links and network nodes in search engines

Search engine	Link	Node
Google	%84	%44
Yahoo	%49	%16
Bing	%47	%18

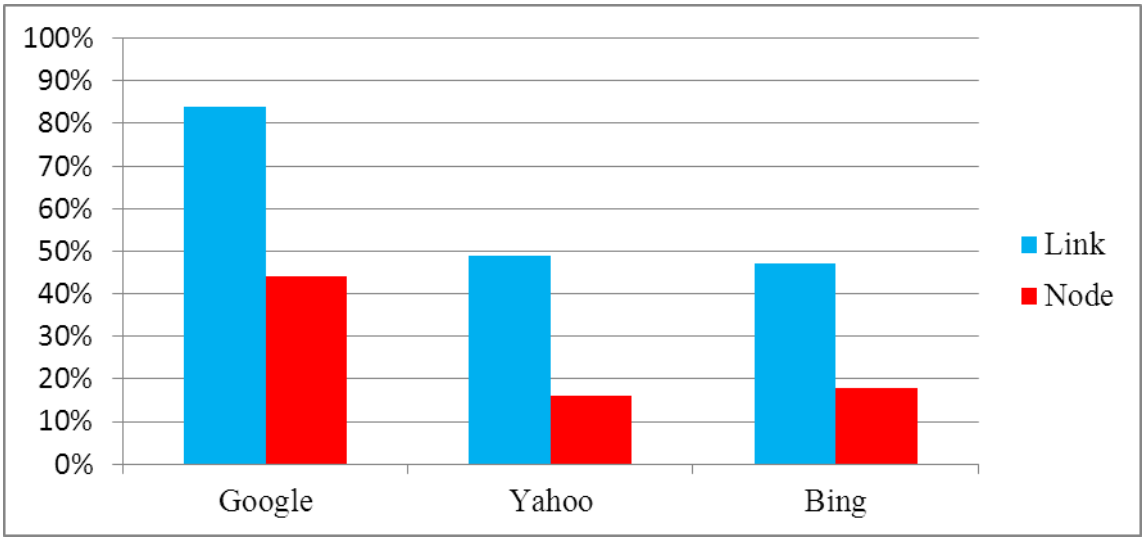


Figure 1. Percentage of unique extracted results links and network nodes

The percentage of shared nodes compared with shared links (Table 9 and Figure 2) has increased in three search engines. For example the percentage of shared links in Google and Bing search engines was 8% which has increased to 29 percent in shared nodes. Also, the percentage of shared links in three search engines was 3% which has increased to 21 percent in shared nodes. This is significant and shows an increase in level of overlap.

Table 9. Percentage of shared results links and network nodes in search engines

Search engine	Link	Node
Google – Bing	%8	%29
Bing – Yahoo	%44	%52
Yahoo – Google	%7	%30
Google – Bing _Yahoo	%3	%21

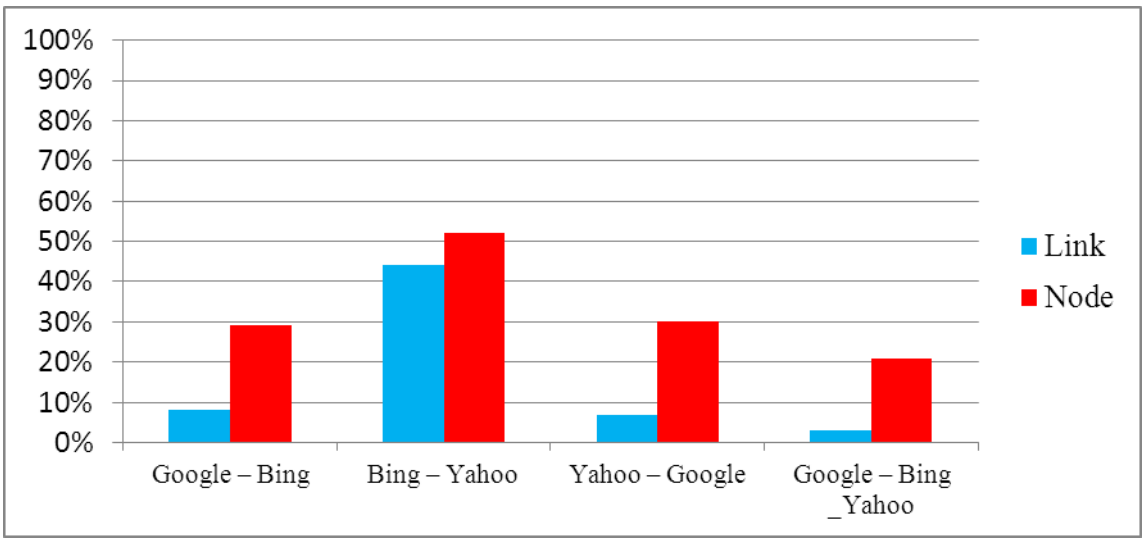


Figure 2. Percentage of shared extracted links and network nodes

According to table 10 and figure 3, the percentage of unique network nodes is very low compared with unique links provided by each of the search engines. This indicates a decrease in unique results of search engines. Also, the percentage of shared nodes between two search engines or three search engines is higher than shared links between two or three search engines. This suggests an increase in shared results provided by search engines.

Table 10. Percentage of unique results links and network nodes in one, two, and three search engines

Search engines	Link	Node
Unique link	%59	%26
Two search engine	%20	%37
Three search engine	%3	%21

4. Conclusion:

In this study, proper inquiries were entered in three Google, Yahoo, and Bing search engines. Then, the first results links presented by these search engines were compared with each other. According to conducted tests, it was determined that the number of unique links in each of the search engines is high and the number of shared links among these three search engines is very low. Therefore, it can be concluded that the overlap in search engines results links is very low.

Then, the results links in these three search engines were used to extract the textual content of results pages. Processing the texts, the query related words were obtained and stored in the database. The list of words was sorted based on their frequency. Then, the social networks' nodes were extracted from list. Comparing the created networks by each of the search engines, it was determined that the unique extracted nodes in three search engines are much lower than primary links and the number of shared nodes among search engines is high. Therefore, it can be concluded that the overlap among social network extraction of search engines is higher than overlap in results links of these three search engines.

References:

1- 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

2- 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

3- Ding Wei, Marchionini Gary, "A Comparative Study of Web Search Service Performance", Proceedings of the ASIS Annual Meeting, v33 (1996), pp136-142

4- Bharat Krishna, Broder Andrei, "Technique for measuring the relative size and overlap of public Web search engines", Computer Networks and ISDN Systems, v30 (1998), pp 379-388

5- Bar-Ilan Judit, Mat-Hassan Mazlita, Levene Mark, "Methods for comparing rankings of search engine results", Computer Networks, v50 (2006), pp 1448–1463

6- Rather Rafiq Ahmad, Lone Fayaz Ahmad, and Shah Gulam Jeelani, "Overlap in Web Search Results: A Study of Five Search Engines", Library Philosophy and Practice (e-journal), (2008), Paper 226

7- Sterling Greg, "Different Engines, Different Results", A Research Study by Dogpile.com, http://searchengineland.com/dogpile-search-results-on-major-engines-diverging-11364, (1394/9/17)

8- Lewandowski Dirk, "The retrieval effectiveness of web search engines: considering results descriptions", Journal of Documentation, v 64 No 6(2008), pp 915-937

9- Usmani Tauqeer Ahmad, Pant Durgesh, Bhatt Ashutosh Kumar, "A Comparative Study of Google and Bing Search Engines in Context of Precision and Relative Recall Parameter", International Journal on Computer Science and Engineering (IJCSE), v4 (2012), pp 21-34

10-Kumar Kailash, Bhadu Vinod, " A COMPARATIVE STUDY OF BYG SEARCH ENGINES", American Journal of Engineering Research (AJER), v2, Issue 4 (2013), pp-39-43

11-AJAYI Olusola Olajide, ELEGBELEYE Damilola Matthew, " Performance Evaluation of Selected Search Engines ", Computer Engineering and Intelligent Systems, v5, no1(2014)

12-Isfandyari Moghadam Alireza, Bahari Movaffagh Zohreh, " The Overlap Rate Of Searching Medical Keyword In General Search Engines", Health Information Management, v9, no 2 (2012), pp 2003 – 2014
13-Lewandowski Dirk , "Evaluating the retrieval effectiveness of Web search engines using a representative query sample ", Journal of the Association for Information Science and Technology, v 66, Issue 9 (2015), pp 1763–1775
14-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
15-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
16- 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
17- Harada Masanori, Sato Shin-ya, Kazama Kazuhiro, "Finding authoritative people from the Web", JCDL'04, (2004), pp306 - 313
18-Culotta Aron, Bekkerman Ron, McCallum Andrew, "Extracting social networks and contact information from email and the Web", American Association for Artificial Intelligence
19-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
20-A.Adamic Lada, Buyukkokten Orkut, Adar Eytan, "Social network caught in the Web", First Monday Peer-Reviewed Journal On The Internet (2003), v8, n6
21-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
22-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
23-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
24-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
25-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
26-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
27-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
28-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
29- Alwahaishi Saleh, Martinovic Jan, Snasel Vaclav, Kudelka Milos, "Analysis of the DBLP Publication Classication Using Concept Lattices", Dateso (2011), pp 132-139