Search Engine Pictures: Empirical Analysis of a Web Search Engine Query Log

Farzaneh Shoeleh[1,2] , Mohammad Sadegh Zahedi[1,2], Mojgan Farhoodi[2]
School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
Iran Telecommunication Research Center, Tehran, Iran
{f.shoeleh, s.zahedi, farhoodi}@itrc.ac.ir

***Abstract*— Since the use of internet has incredibly increased, it becomes an important source of knowledge about anything for everyone. Therefore, the role of search engine as an effective approach to find information is critical for internet's users. The study of search engine users' behavior has attracted considerable research attention. These studies are helpful in developing more effective search engine and are useful in three points of view: for users at the personal level, for search engine vendors at the business level, and for government and marketing at social society level. These kinds of studies can be done through analyzing the log file of search engine wherein the interactions between search engine and the users are captured. In this paper, we aim to present analyses on the query log of a well-known and most used Persian search engine. Our analyses are presented in three main categories: 1) Stats-based analyses, 2) Temporal-based analyses, and 3) Topic-based analyses. The obtained results are promising. Mobile users often posted queries in weekends, whereas Web users utilize the search engine in workweeks. The majority of queries posted form most-populated cities. Additionally, Iranians are mostly interested in political, social, and economical topics.**

Keywords—Web Search engine; Query Log analyses; Information Search and Retrieval

I. INTRODUCTION

The internet is an important source of knowledge about anything for everyone. Nowadays, as the use of internet has incredibly increased, Web search engines become the common approach to find and retrieve needed information. A Web search engine is a software system that is developed to help users to find their needed information through searching on the World Wide Web. Users enter queries into a Web search engine when they seek information, help, or advice.

With the exponential growth of information, powerful and successful search engines are needed to find the exact information that users are seeking for. One of the open directions in information retrieval domain is researching on methodologies to analyze and investigate the effectiveness of a search engine. One potential way to find whether a search engine is successful in retrieving information from Web or not, is analyzing its users' behaviors and satisfaction. To analyze the behavior of search engine users, there are two main approaches: 1) Through human evaluation and analyses on their observations and viewpoints [1,2], where the users' opinions are qualitative. 2) Through automatic analyses on the search engine log file which is quantitative. It should be noted that the human based measurement is more accurate. However, the second category has become an important role in search engine evaluation because of the expensive and time-consuming of the first category.

In general, log analysis is an art and science seeking to make sense out of the records of a software system. In special, search log analysis is a kind of Web analytics software that parses a search engine log file from a search engine, and based on the values contained in the log file, derives indicators about when, how, and by whom a search engine is visited. Reports are usually generated immediately, but data extracted from the log files can alternatively be stored in a database, allowing various reports to be generated on demand.

The search logs capture a large and varied amount of interactions between users and search engines, which is less susceptible to bias and enables identifying stronger relationships between the data. Previous works show that there are differences between users living in different world regions because they have their own language, vocabulary and cultural bindings. For example, authors in [3, 4] indicate that Europeans' queries are more about people and places, while the U.S. users' ones are more focused on ecommerce. Therefore, in this paper, we plan to study the specificities of the behavior of Persian Web search engine users. This study is done in *Webazma*[1] laboratory, located in Iran Telecom Research Center, wherein the researches officially focus on evaluating and analyzing Persian search engines' services such as text, image, video, news, and etc. [5, 6, 7, 8, 9].

Query log analysis has become one of the important research trend [10-12]. In this paper, we aim to analyze the Iranians' behavior through Web search engine based on search log gathered form *Parsijoo*[2] search engine, the most well-known Persian search engine, through 3 January and 11 February 2017. Our analyses are categorized in three main categories: 1) Stats-based analyses, 2) Temporal-based analyses, and 3) Topic-based analyses. The first category consists of analyses about the overall statistics of the queries posed by users. The second one contains the analyses which time has important role. The last but not least category presents analyses on the content of queries.

[1] http://webazma.itrc.ac.ir/
[2] http://parsijoo.ir/

The paper is structured as follows: Section 2 covers the most important related work. Section 3 describes the methodology and logs dataset from which we based our study and Section 4 presents the analyses on search engine users and Section 5 finalizes with the discussion of results and conclusions.

II. Related Work

Most of researches in this area generally are similar, and their differences are in the underlying search engine and also their proposed analyses on the data gathered from search engine.

In [13], authors analyzed the log of *Alta Vista* search engine through three weeks. The analyses are provided in two main categories: 1) First order analyses of queries and 2) Second order analyses of queries. In former category, only one item is considered to be analyzed. The analyses are done on the properties of queries such as average length of queries, categorization queries based on their number of words, the percentages of usage of operators like *and* and *or* and also the query frequency. In later category, the analyses are done on the pair of items, for example considering where two words like *computer* and *programming* are occurred simultaneously or not. The results of such analyses on *Alta Vista* indicate that 77 percentage of users' sessions contain only one request, i.e. query. In addition, the average length of queries is about 2.3 words. A few of queries are repeated through one day and the 25 most frequent queries are entered through 43 days.

In [4], authors tried to investigate the user behavior of *NAVER*, the most famous Korean search engine. The log of this search engine consists of about 40 million queries requested through one week. The authors firstly preprocess the log file to eliminate the informal and invalid requests and ignoring the other additional data. Then, classified the queries based on the services of search engine which is used by the user to enter the query. Authors categorize the queries in two main classes: 1) Initial input queries, the queries that user enter them for the first time. 2) Subsequent queries, the queries which are generated by a user after her/his first entering query. For example, user adding or eliminating a/some words from her/his first queries. The analyses done on *NAVER* log file in this research consists of the number of queries according to its type, the distribution of quires per session, the number of results clicked by users per each query. The results indicate that users often enter short-length queries and seldom use advance search service. Also, most of the time, they only click on the first link in the result page. On of the most interesting findings in this study is that users generally tend to change their queries completely instead of adding or omitting a/some words.

In [3], there is a research study about the users' behavior of *Tumba* as a Portuguese search engine between 2003 and 2004. Since unmoral and invalid queries and session may cause negative effects on their studies, the author suggested that before any analyzing the log file must be cleaned from these kinds of issue, such as incomplete queries and sessions without any query or having more than 1000 queries. The results are presented via three main analytic categories:

1) Queries: the number of queries according to their type, the number of visited pages per query, the users' clicks statistics, the distribution of users' clicks, the length of queries, the frequency of queries.

2) Queries' words: the frequency of words included in queries and topic query classification.

3) Session: the average of sessions' time and the average of number of queries per session.

According the obtained results, users enter 749914 and 333781 queries through 2003 and 2004, respectively. In addition, the average of the number of queries in each session is 2.94 and 2.49 through 2003 and 2004, respectively. In sum, the average length of queries is about 2.2 and 1.4 pages are clicked by users for each query.

In [14], authors analyzed the log file of two well-known search engines: *Excite,* as an American search engine and *FAST*, as a European search engine. The average length of Americans' queries are longer than Europeans' one. Also, American users open less results pages than the European users. The most interesting topics among Americans' are Business, Travel and Tourism, Employment and Economy, whereas the Europeans' most interesting topics are People and Place. Also, authors in [15] claimed that the behavior of *Excite's* users does not change dramatically over the time.

In [16], is was found that among 15000000 queries entered into *MSN* [17] search engine, 124422 queries have religious topic. In [16], a method is proposed to classify this type of queries according to their words. They show that people with different religious have different searching patterns. Additionally, they demonstrated that both the length of session and also the average length of religious queries are mostly longer than other queries.

III. The User Behavioral Analyses

There are many approaches to analyze the behavior of search engine users. One of the most reliable approach is parsing, analyzing and data mining of a search engine logs. In this paper, we aim to provide some statistics analyses about the users of a well-known Persian search engine. To do so, three following main steps must be done sequentially: 1) Data Gathering, 2) Data Preparing and 3) Data Analyzing.

We use the web server apache log file of search engine that contains a series of requests with many items that only some of which concern us here. Each request consists of following useful items:

1. **Time Stamp:** to indicate when the query was submitted.
2. **Session ID:** to identify whether a query is submitted through same session.
3. **Query Title:** the exact title of a query submitted to search engine.
4. **IP:** the IP of user device.
5. **User Agent:** to identify the application type, operating system, software vendor or software version of the requesting software user agent.

A. Data Gathering

Data gathering can be executed through an independent program to collect useful data according to each query posted into a search engine. On the other hand, one possible solution for data gathering is utilizing the search engine log file which are accessible via the specific web service provided by the search engine. Here, as mentioned before, we use the apache log of search engine.

B. Data Preparing

The content of a search engine log file is parsed and inserted data into a relational Data Base (DB) after creating corresponding tables with proper columns to storage this data structurally. While inserting the data, we eliminate the incorrect, invalid, broken requests and also the additional information which are not useful in our analyzing process. It is worth mentioning that we scan the request properly and detect the robot requests manually. Then, we omit such automated/tested requests from DB.

C. Data Analyzing

To analyze the users' behavior, we classify our analyses into three main categories:

1) Stats-based Analyses, contains analyses about the overall statistics of the queries posed by users, such as:

- Query length distribution.
- Query frequency distribution.
- Number of terms per query.
- How often distinct queries are asked.
- Geographical distribution of the number of queries posted by both mobile and Web users.

2) Temporal-based Analyses, contains the analyses which time has important role, such as:

- Distribution of query count over 24-hours for both mobile and Web users.
- Weekdays query count distributions for both mobile and Web users.

3) Topic-based Analyses, presents analyses on the content of queries.

- Ten most frequent queries.
- Word cloud of users' queries for three salient social events.
- Topics of users' queries.

Our analytic diagrams, the methodology of each analysis and also the obtained results are presented in results and discussion section.

IV. Results And Disscusion

In this section, for each analyses category, we firstly describe the category and its results, obtained during the time between 2017-01-03 and 2017-02-10, would be presented.

A. Stats-based Analyses

The aim of this type of analysis is to extract the overall search engine stats in view of different aspects. In this section we present an analysis of Persian search engine's query log containing almost 4 million entries to **find its users' behavior.**

Table 1 shows the statistics summarizing the query log contents used in our experiments. Figure 1 illustrate the query length distribution. As results show, the majority of the queries have 2 or 3 terms, i.e. words, and the average length of queries in Persian domain is about 3.46. On the other hand, as the previous work on the log analyses of search engine which are developed in different languages show, the average length of queries in *AltaVista* is 2.35, in *Excite* is 2.6, in *NAVER* is 1.13 and in *Tumba* is 2.2. In addition, the query length has a semi-power law distribution, as shown in Figure 1. Note that less than 14% of the queries had 5 or more terms.

Similarly, Figure 2 and Table 2 show the statistics concerning "How often distinct queries are asked", called query frequency. As Figure 2 indicates, the distribution of queries frequency is semi-power law, means that the number of queries which occur only one time is much bigger than the sum of queries occurring more than 7 times. In other words, 94% of queries occur less than 7 times.

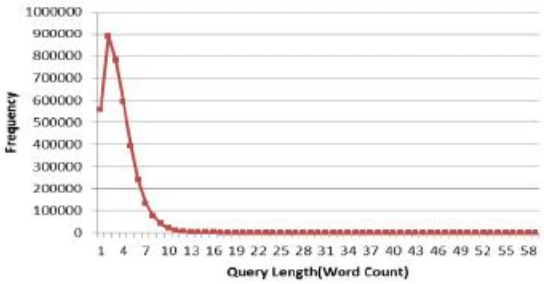

Fig. 1. Query length distribution.

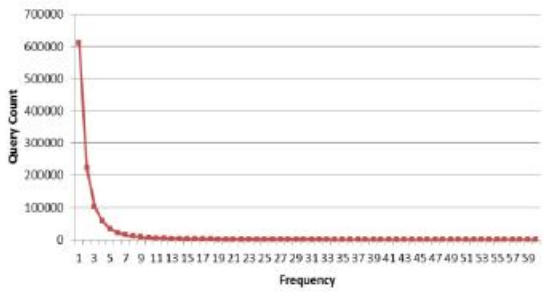

Fig. 2. Query frequency distribution.

TABLE 1 Overall Statistics concerning the number of terms per query.

1 term in query	15%
2 terms in query	24%
3-5 terms in query	47%
>5 terms in query	14%
Max #terms in query	79
Average #terms in query	3.46
STD of #terms in query	2.12

TABLE 2 OVERALL STATISTICS CONCERNING HOW OFTEN DISTINCT QUERIES ARE ASKED.

Query occurs 1 time	54%
Query occurs 2 time	20%
Query occurs between 3 and 7	20%
Query occurs >7 time	6%
Max query frequency	16022
Average query frequency	3.33
STD of query frequency	37.62

Fig. 3. Geographical distribution of query count for Mobile users.

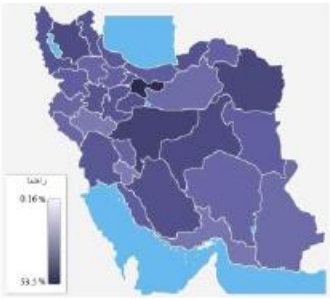

Fig. 4. Geographical distribution of query count for Web users.

Recent investigations show that nowadays mobile phone penetration rate has been rapidly increased in Iran (currently it is about 99%). Mobile phone penetration rate is the number of active mobile phone users within a specific population[3]. Therefore, in this paper, we also analyze the mobile users' behavior independently. Through Figure 3. and Figure 4, we show the geographical query count distribution for both mobile and Web users, respectively. As results indicate, most of queries are posted from Tehran by both mobile and Web users. 75.5% of mobile users' queries and 53.3% of Web users' queries are posted by the users who connect the internet in Tehran, means their IPs are provided by the ISP which are reserved for Tehran. Note that in both Figure 3 and 4, the Iran map is colored based on the log of number of queries posted from each province. Results illustrate that the number of posted queries form each province is proportionate to its population.

B. *Temporal-based Analyses*

Here, we analyze the behavior of search engine users during 24-hours and weekdays. Figure 5 and Figure 6 show the number of queries which are posted during 24-hours by Web users and mobile users, respectively. The number of queries distribution over 24-hours for We users is approximately normal distribution. Most of the Web users' queries are posed to the search engine between 9am and 13pm, i.e. peak working hours. On the other hand, as Figure 6 shows, the number of queries distribution over 24-hours for mobile users is approximately left skewed distribution. In contrast to Web users, most of the mobile users' queries are posed to the search engine after 5pm, i.e. after work hours.

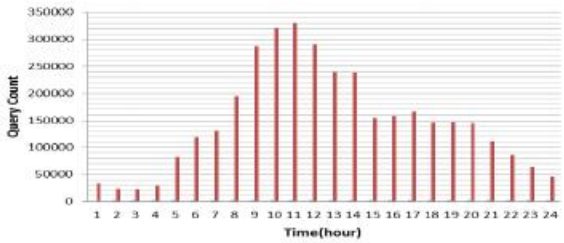

Fig. 5. distribution of query count over 24-hours for Web users.

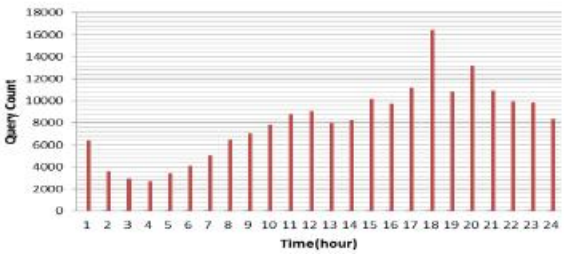

Fig. 6. Distribution of query count over 24-hours for mobile users.

Figure 5 and Figure 6 show the weekdays query count distributions for both Web and mobile users, respectively. These distributions indicate how many queries are posed during weekdays by Web or mobile users. Mobile users often use search engine in weekends, whereas the Web users utilize the search engine in workweek. It is worth mentioning that Thursday and Friday are weekend in Iran.

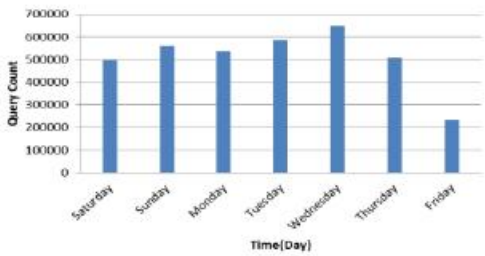

Fig. 7. Weekdays query count distributions for Web users.

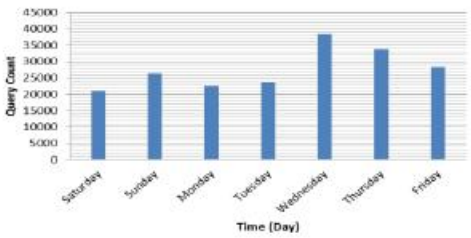

Fig. 8. weekdays query count distributions for Mobile users.

[3] http://www.khabaronline.ir/detail/601069/Economy/macroeconomics

C. Topic-based Analyses

Here, the analyses in which the context of query is important would be presented. Table 3 lists the ten most frequent queries for both mobile and Web users, independently.

TABLE 3 TEN MOST FREQUENT QUEIRIES.

	Mobile Users		Web Users	
rank	Query (Persian)	Query (English)	Query (Persian)	Query (English)
1	تقویم	Calendar	دیوار	Divar
2	اذان تهران	Tehran Azan	تقویم	Calendar
3	حقیر اخلاقی>	Sextual Query	قیمت سکه ارز	Price of Gold Coin
4	ورزش۳	Varzesh3	مسابقات ورزشی	Sport Competition
5	خبرگزاری فارس	Fars news	گوگل	Google
6	دیوار	Divar	ورزش۳	Varzesh3
7	اپارات	Aparat	ضمن خدمت فرهنگیان	In-service teachers
8	جی ال ایکس	GLX	دیجی کالا	Digikala
9	ایران خودرو	IranKhodro	ایران خودرو	IranKhordo
10	فارس	Fars	شهر خبر	ShahreKhabar

Figure 9, 10 and 11 illustrate the word cloud [18] of users' queries for three most important events happening through our evaluation duration time (between 2017-01-03 and 2017-02-10). These events are:

1. Death and state funeral of Hashemi Rafsanjani, the fourth President of Iran and the country's Chairman of Expediency Discernment Council.
2. Plasco fires and collapsing buildings, was an iconic 17-story high-rise landmark building in Tehran, the capital city of Iran.
3. Dahe-ye Fajr, i.e. dawning of new age, is a ten-day celebration of Ruhollah Khomeini's return to Iran in 1979.

Note that the first event happened in 8 January 2017, the second one happened in 19 January 2017. The last event is an annual celebrations last from 1 to 11 February. The start of the celebration coincides with the date of Ruhollah Khomeini's arrival and the ending with Revolution's victory, a day which is called Islamic Revolution's Victory Day or 22 of Bahman. To find the word clouds of each event, we firstly extract the users' queries which were posted during the event period from log file. Then, we calculate the frequency of each distinct word and weight each word based on its frequency in the word cloud visualization.

Figure 9, 10 and 11 indicate that these events, as social events, was reflected in users' queries. The words related to these events have bigger weight in the corresponding word cloud. For example, for the first event the words "هاشمی" and "رفسنجانی" (referred to "Hashemi" and "Rafsanjani", respectively) occur among ten first words, for the second event the word of "پلاسکو" (referred to "Plasko") has rank of 4 and for the last event the words of "دهه" and "فجر" (referred to "Dahe" and "Fajr" respectively) have high rank. As results, one of the advantageous of search engine log analyses is to detect the social events and consequently it opens directions to analyze the user's behavior.

We analyze users' queries based on their topics. We classified the queries into 12 specific classes. To classify the queries, an IR-based method similar what proposed in [19] was implemented. We use the last version of Hamshahri news dataset [20] and Okapi BM25 probabilistic model [21] to find the topic of a given query. The topic is determined by using majority voting approach on the category of top 10 retrieved documents. Having all queries classified, we first check how these queries are distributed across different topics to find out the most interesting topic for Persian language users. Figure 12 shows the ratio of each 12 categories in a pie chart. As it can be conducted from Figure 12, most of these queries concern Society (15% of total queries), Economic (10% of total queries), Art and Culture while few queries are posed concerning topics like Environment and Religious.

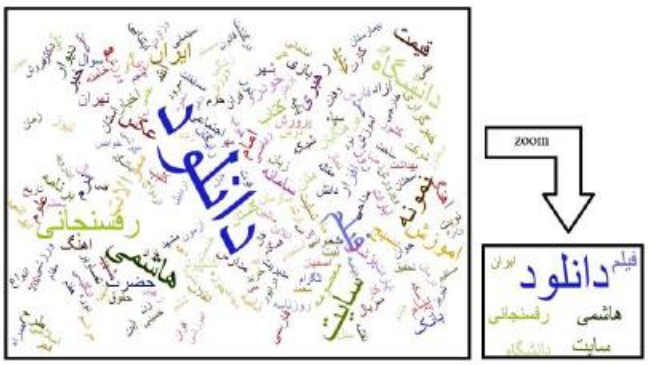

Fig. 9. The word cloud of users' queries for *Death and state funeral of Hashemi Rafsanjani* event.

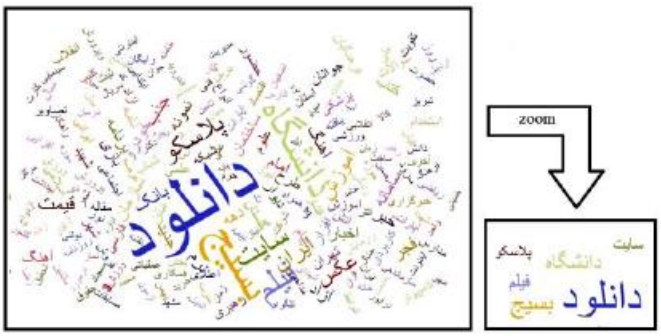

Fig. 10. The word cloud of users' queries for *Plasco fires and collapsing buildings* event.

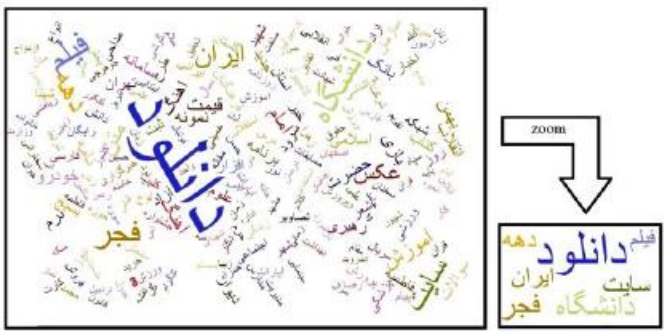

Fig. 11. The word cloud of users' queries for *Dahe-ye Fajr* event.

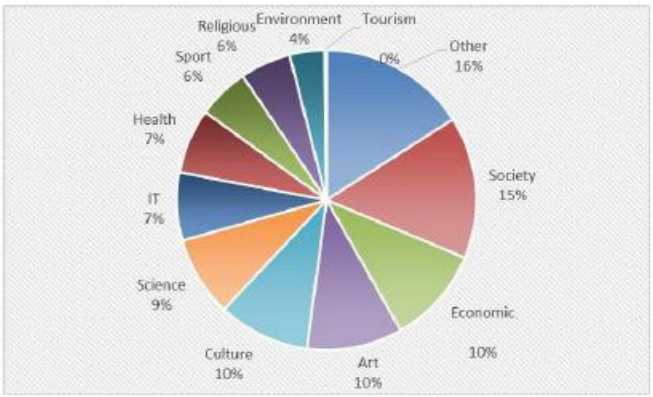

Fig. 12. Breakdown of categorized users' queries

V. Conclusion and Future Work

In this paper, we presented an analysis of Persian search engine's query log containing almost 4 million entries to **examine the search behavior of a large number of users.** We categorize the analyses into three main classes: 1) Stats-based Analyses, Temporal-based Analyses, and 3) Topic-based Analyses. The first category consists of analyses about the overall statistics of the queries posed by users.

The second one contains the analyses which time has important role. The last but not least category presents analyses done on the content of queries. Stats-based analyses shows that the average length of queries in Persian is longer than queries in English. Also, the number of queries which occur only one time is much bigger than the sum of queries occurring more than 7 times. In other words, 94% of queries occur less than 7 times. Through geographical analyses, it is found that the number of posted queries form each province is proportionate to its population. Temporal-based Analyses indicates that Mobile users often use search engine in weekends or posted after 5pm, i.e. after work hours, during workweek. On the other hand, Web users utilize the search engine in workweek, especially between 9am and 13pm, i.e. peak working hours. Topic-based Analyses illustrate that social events may be reflected in users' queries. Hence, analyzing the users' queries might be useful to detect social events.

Research can be conducted to further investigate the findings of this study. This study quantitatively analyzed Web users' search behavior based on search engine logs. Thus, further research might be needed to address issues using scripts which gathers the needed information about user requests and his/her behaviors such as identifying unique users or analyses about their clicks.

Acknowledgment

This study is conducted by support of Iran Telecommunication Research Center. We appreciate *Webazma* team members who have helped us most throughout our research.

References

[1] C. H¨olscher and G. Strube, "Web search behavior of internet experts and newbies," Computer networks, vol. 33, no. 1, pp. 337–346, 2000.
[2] B. J. Jansen and A. Spink, "How are we searching the world wide web? a comparison of nine search engine transaction logs," Information processing & management, vol. 42, no. 1, pp. 248–263, 2006.
[3] M. Costa and M. J. Silva, "A search log analysis of a Portuguese web search engine," Proc. of the 2nd INForum-Simp´osio de Inform´atica, vol. 525, no. 536, p. 5, 2010.
[4] S. Park, J. H. Lee, and H. J. Bae, "End user searching: A web log analysis of naver, a korean web search engine," Library & Information Science Research, vol. 27, no. 2, pp. 203–221, 2005.
[5] F. Shoeleh, A. Azimzadeh, A. Mirzaei, and M. Farhoodi, "Similarity based automatic web search engine evaluation," in Proceedings of the 8th international symposium on Telecommunication, IEEE, 2016.
[6] R. Badie, M. Azimzadeh, M. S. Zahedi, and S. Samuri, "Automatic evaluation of search engines: Using webpages' content, web graph link structure and websites' popularity," in Telecommunications (IST), 2014 7th International Symposium on, pp. 556–562, IEEE, 2014.
[7] M. Mahmoudi, R. Badie, M. S. Zahedi, and M. Azimzadeh,"Evaluating the retrieval effectiveness of search engines using persian navigational queries," in Telecommunications (IST), 2014 7th International Symposium on, pp. 563–568, IEEE, 2014.
[8] M. Azimzadeh, R. Badie, and M. M. Esnaashari, "A review on web search engines' automatic evaluation methods and how to select the evaluation method," in Web Research (ICWR), 2016 Second International Conference on, pp. 78–83, IEEE, 2016.
[9] M. S. Zahedi, B. Mansouri, S. Moradkhani, M. Farhoodi, F. Oroumchian, "How Questions are Posed to a Search Engine? An empiricial analysis of Question Queries in a Large Scale Persian Search Engine Log." Web Research (ICWR), 2017 Third International Conference on. IEEE, 2017.
[10] D. Jiang and L. Yang, "Query intent inference via search engine log," Knowledge and Information Systems, vol. 49, no. 2, pp. 661–685, 2016.
[11] J. Palotti, A. Hanbury, H. M¨uller, and C. E. Kahn Jr, "How users search and what they search for in the medical domain," Information Retrieval Journal, vol. 19, no. 1-2, pp. 189–224, 2016.
[12] S. Dumais, R. Jeffries, D. M. Russell, D. Tang, and J. Teevan, "Understanding user behavior through log data and analysis," in Ways of Knowing in HCI, pp. 349–372, Springer, 2014.
[13] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," in ACm SIGIR Forum, vol. 33, pp. 6–12, ACM, 1999.
[14] A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen, "Us versus european web searching trends," in ACM Sigir Forum, vol. 36, pp. 32–38, ACM, 2002.
[15] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "From e-sex to e-commerce: Web search changes," Computer, vol. 35, no. 3, pp. 107–109, 2002.
[16] R. Wan-Chik, P. Clough, and M. Sanderson, "Investigating religious information searching through analysis of a search engine log," Journal of the American Society for Information Science and Technology, vol. 64, no. 12, pp. 2492–2506, 2013.
[17] P. Serdyukov, G. Dupret, and N. Craswell, "Wscd2013: workshop on web search click data 2013," in Proceedings of the sixth ACM international conference on Web search and data mining, pp. 787–788, ACM, 2013.
[18] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization," in Visualization Symposium (PacificVis), 2010 IEEE Pacific, pp. 121–128, IEEE, 2010.
[19] P. Ullegaddi and V. Varma, "A simple unsupervised query categorizer for web search engines," in Proceedings of ICON-2010: 8th International Conference on Natural language Processing. Macmillan Publishers, 2011.
[20] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard persian text collection," Knowledge-Based Systems, vol. 22, no. 5, pp. 382–387, 2009.
[21] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., "Okapi at trec-3," Nist Special Publication Sp, vol. 109, p. 109, 1995.