2017

How questions are posed to a search engine An empiricial analysis of question queries in a large scale Persian search engine log

Mohammad Zahedi
Iran Telecommunication Research Center, Tehran University

Behrouz Mansouri
Tehran University

Shiva Moradkhani
Iran Telecommunication Research Center

Mojgan Farhoodi
Iran Telecommunication Research Center

Farhad Oroumchian
University of Wollongong in Dubai, farhado@uow.edu.au

Publication Details

Zahedi, M. Sadegh., Mansouri, B., Moradkhani, S., Farhoodi, M. & Oroumchian, F. 2017, 'How questions are posed to a search engine An empiricial analysis of question queries in a large scale Persian search engine log', 2017 3rd International Conference on Web Research, ICWR 2017, IEEE, United States, pp. 84-89.

How Questions are Posed to a Search Engine?

An empiricial analysis of Question Queries in a Large Scale Persian Search Engine Log

Mohammad Sadegh Zahedi[1,2], Shiva Moradkhani[1], Behrouz Mansouri[2], Mojgan Farhoodi[1],Farhad Oroumchian[3]
IT Faculty,Iran Telecommunication Research Center, Tehran, Iran [1]
Database Research Group,School of Electrical and Computer Engineering,University of Tehran, Tehran, Iran[2]
University of Wollongong in Dubai, Knowledge Village, Dubai, UAE[3]
{s.zahedi, sh.moradkhani}@itrc.ac.ir, b.mansouri@ut.ac.ir, farhoodi@itrc.ac.ir, farhadoroumchian@uowdubai.ac.ae

***Abstract*— In this paper we investigate the Persian search engine log and present a comprehensive analysis of question queries in three levels: structure, click and topic. By analyzing question queries characteristics, we explore behavior of Persian language users. Our experimental results show that question queries length are larger than normal queries. Most of these queries had question words "How" and "What" and their topic were mostly about health, policy, religion and society.**

Keywords—question query; query log analysis; large scale query log

I. INTRODUCTION

Large scale query log analysis has become one of the important research trend [1-4]. User activity recorded in these logs can be valuable resources for the research areas like information retrieval, data mining, natural language processing and machine learning. Recently researchers have studied the queries in form of question recorded in search engine log. Normally a question is the simplest way to express an information need. Search engines prefer keyword-based queries while users intend to express their queries by natural language form such as question. Previous studies show that the number of question queries being issued to search engine are increasing. Pang and Kumar analyzed Yahoo query log in 2010 and indicated that about 2% of entire queries were in question format [5]. In [6] one billion Russian question queries were analyzed and it was shown that 3-4% of total query traffic are questions.

Studies in [5,7,8] reported that search engines have lower performance when natural language question is asked compared to normal keyword-based query. Therefore according to increase in question queries, development of special purpose search algorithms for answering these queries are required. The latest Google algorithm in 2013 (Hummingbird) aimed to answer long natural queries[1]. As the number of these queries are increasing and they usually have special characteristics, analyzing them can give a better insight to search engine about answering them.

This paper brings into focus the analysis of Persian question queries by using queries submitted to Iranian search engine named Parsijoo[2] during one year period consisting of 27 million of queries. Our analysis indicates that about 1.6% of total queries submitted are in the form of natural language questions. Even though this is a small proportion of queries, studying these queries is valuable as users prefer to ask their information need using natural language format to better express and specialize their intent rather than short keyword query [9,10]. Our analysis on question query log has four main steps: parsing, extraction, cleaning and analyzing.

To the best of our knowledge, we make the first attempt to analyze Persian question queries in a large scale search log. The main purposes of our study are to seek the answers for these questions:

- How question queries are being formulated by Persian language users?
- What are the main characteristics of Persian question query?
- What are the common topics for Persian question queries?

The rest of the paper is organized as follows: In next section, discusses the previous studies about query log analysis. In Section 3 we explain how we have collected a question query log from the primary query log. Section 4 is where our analysis is performed on the extracted question queries and our analysis results are presented. We conclude with future work in Section 5.

II. RELATED WORK

Query log analysis was target of many studies to explore user intent in several dimensions, including semantic classes [11], goals [12] and topics [13]. Richardson [14] explored a one-year log of millions of users to understand the long-term dependencies of users' intents and preferences. His analysis

[1] https://en.wikipedia.org/wiki/Google_Hummingbird

[2] www.Parsijoo.ir

showed that by studying user behavior on long periods we can gain information that are not presented in shorter logs.

Question queries have been also studied in the context of long queries [8]. In 2011, [5] studied the question queries in search engine logs and analyzed their characteristics and showed that these type of queries are growing. Studies in [5,7,8] shown that the results for question queries are worse compared to keyword-based searches.

Völske et al. [6] reviewed one year query log of Yandex[3] and aimed at the topical categorization of questions retrieval systems. This type of categorization have been target of many studies. Cai et al. [15] enriched question queries with Wikipedia entries to solve the sparseness problem. Chan et al. [16] proposed a hierarchical question classification by using a set of kernels corresponding to different aspects of questions. Bailey et al. [17] considered sparse user interaction data to classify long queries by matching them with more popular and shorter queries categorized based on past users' behavior. In this paper we focus on analyzing question queries submitted to a search engine in different aspects including their content, format and topic.

III. THE QUESTION QUERY DATASET

For our analysis we use the Parsijoo query logs with ~27M queries and all users' interaction with search engine including associated clicks. These queries were submitted to search engine during the year 1394 (Shamsi Calendar; equivalent to March 2015 - March 2016). For creating Persian question query dataset from the original query log, 3 preprocessing steps are used:

- Parsing query logs
- Question query extraction
- Cleaning question query

A. *Parsing query logs*

In area of web search engines, a transaction log is considered as a record of interactions between a web search engine and user during a searching session. Web searching transaction logs are a specific type of transaction log file. This searching log format is similar to the extended file format, which contains data such as client computer's Internet Protocol (IP) address, search engine access time, user query, and referrer site, among other fields. For preparing data, in this phase we parsed these log files and saved the result in a relational database.

B. *Question query extraction*

To extract question query we propose similar method that was presented in [8]. First we prepared a list of 40 Persian question words. Then a query is tagged as a question if it contains one of 40 combinations of this question word. Some of these question words with their English translation are indicated in Table 1.

[3] http://yandex.ru

Table 1 : Common Persian question words

Persian	چگونـه	چرا	کـیست	کـجا	کـی	کـدام
English	How	Why	Who	Where	When	Which

By using the mentioned approach we extracted 431820 question queries which is about 1.6% of total queries submitted to Parsijoo.

C. *Cleaning question query*

Searches from both human and agents are recorded in query logs. As our goal is to analyze characteristics of question query submitted by users therefore in the first step we removed bot queries from the actual question queries that were extracted from the previous phase. A query with the following properties is considered as a bot query:

- In "UserAgent" field of transaction log record the word "Bot" exists.
- The query length is more than 50

This cleaning step removed 67483 queries. Our next cleaning step was to eliminate repeated question queries. Repeated queries are the ones that have been recorded more than once by the same user in the same session, in the query logs. This happens when a user clicks on SERP (search engine result page). This step eliminated 158566 queries.

Some question words have ambiguity and may result in wrong detection of question queries. For example the word "کیست" in Persian has both meaning of "who" and "Cyst". So our third cleaning step was to remove these ambiguous questions using manually created roles.

Finally we filtered out the single word questions as they are meaningless and are not a real question query. At the end of our cleaning step a total number of 205071 queries are considered as real question queries.

IV. PERSIAN QUESTION QUERY ANALYSIS

In this section we focus on three levels of analysis which are query, click and topic. Each of these levels are explored and the data analysis stage is stepped through.

A. *Query level analysis*

We first extract the 10 of the most common question words used in question queries shown in Table 2. Question words such as "چرا", "آیا", "چگونه" usually appear at the beginning of the sentence while question words like "چیست" and "کیست" emerge at the end. Also some of these terms have the same meaning but are used in different positions in a sentence. We then investigate the top 10 most asked questions from Parsijoo. Looking at details, Table 3 shows that these queries contain two common question words: "چیست" and "چگونه". The first four popular questions were asking for information while the other questions asked for an instruction to do a task like cooking, healthy diet. To better understand question query forms, we represent the top five most commonly used initials and ending 2-grams of these questions in Table 4. The word "چگونه" is mostly used with prepositions like "از", "در" and "با" at the start of questions. 2-grams like "یعنی چه" and "چگونه است" are common at the end of sentences.

Table 2 : 10 most occurring question words in queries

Rank	Persian Question Word	English Translation	Frequency
1	چی	What	179,527
2	چیست	What is it	151,214
3	چه	Which	130,758
4	چگونه	How	86,640
5	چرا	Why	54,987
6	آیا	Whether/If	54,580
7	چند	How many/much	47,127
8	کیست	Who	30,876
9	کدام	Which	28,016
10	کجا	Where	26,191

Table 3 : The 10 most popular question queries

Rank	Persian Query Title	English Translation	Frequency
1	پدافند غیر عامل چیست	What is passive defense	568
2	برجام چیست	What is Barjam	472
3	تلگرام چیست	What is Telegram	161
4	طلسم چیست	What is spell	130
5	ناهار چی بپزم که آسون باشه	What to cook lunch that is easy	109
6	چگونه چاق شویم	How to become obese	107
7	چگونه اینترانت به اینترنت تبدیل می شود	How Intranet becomes Internet	96
8	چگونه دعا کنیم	How to pray	89
9	چگونه لاغر شویم	How to become skinny	82
10	چگونه پسر دار شویم	How to have a boy	79

Table 4 : Top five starting and finishing 2-grams

Rank	Persian Query Title	English Translation	Frequency
1	چگونه از	How … from …	1065
2	چگونه در	How … in …	1005
3	چگونه با	How … with …	907
4	منظور از	Meaning of …	736
5	چرا در	Why in …	527
1	یعنی چه	…means what	4998
2	چگونه است	How is …	1248
3	نشانه چیست	Is the symbol of …	774
4	از کیست	Whom is it from	657
5	شده است	Has become	503

The average length of queries is 3.2 terms while for question queries it is about 6.1 because question queries are closer to natural language and are longer than simple keyword-based queries. Figure 1 gives information on question query length distribution which demonstrate approximately an exponential distribution, with the long queries in the tail.

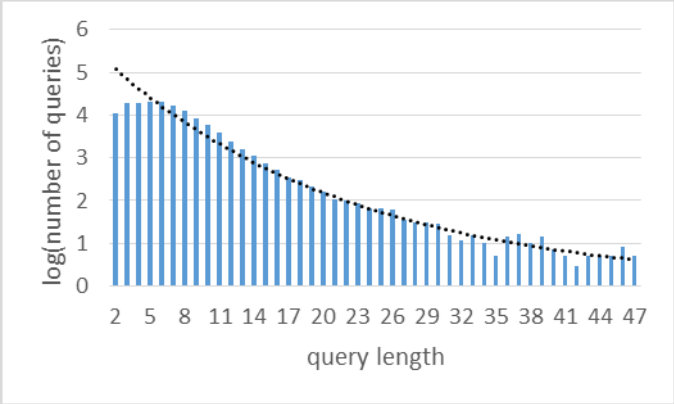

Figure 1 : Distribution of question query length

We also examined the temporal distribution of question queries by considering their distribution over hours of a day and monthly (In all of our analysis we consider Georgian calender). We can observe from Figure 2 that there is an upward trend in number of question queries being asked by users from search engine with the start of official hours (8'clock) and this amount starts to decrease by the midnight.

The result of our exploration on monthly ditribution is indicated in Figure 3. The monthly distribution is approximatley uniform. The increase in number of questions being asked by users from Parsijoo in the last months was due to new features and services of Parsijoo.

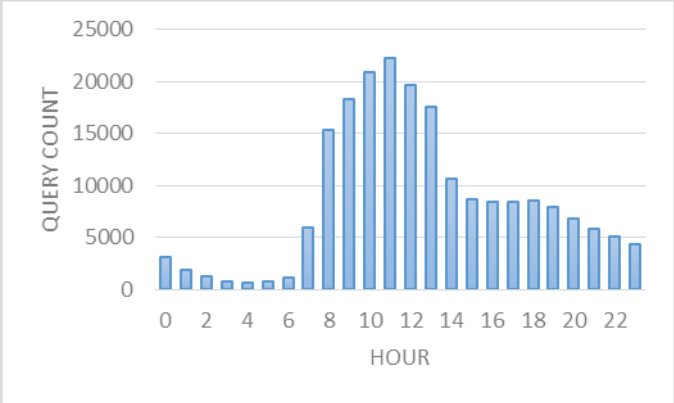

Figure 2 : Question Queries volume by hours of a day

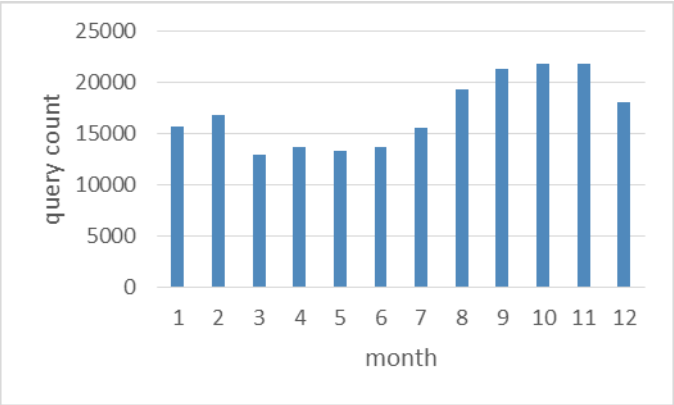

Figure 3 : Question queries volume by month

B. Click level analysis

In this section we carry on our analysis by looking at the click data in the search logs to better understand the user behavior toward the question queries. We first review the top 10 questions with the most clicks on their retrieved results at Table 5.

Table 5 : Top 10 most clicked question queries

Rank	Persian Query Title	English Translation	Frequency
1	خود باوری چیست	What is self-esteem	1,338
2	پدافند غیر عامل چیست	What is passive defense	1,332
3	چیست h.s.e	What is h.s.e	598
4	برجام چیست	What is Barjam	452
5	چرا حسین شهید شد؟	Why Hussein became a martyr?	263
6	تلگرام چیست	What is Telegram	238
7	چگونه چاق شویم	How to become obese	202
8	کم فروشی چیست	What is low retail	189
9	چگونه پولدار شویم	How to become rich	170
10	چگونه لاغر شویم	How to become skinny	168

We then introduce the top 15 most visited websites regarding the results retrieved by search engine for questions asked by users. As it can be seen from Table 6 most of these websites topic are about religion, society, culture and entertainment. Most of these sites provide content to users and also have forums to ask questions.

Table 6 : Top 15 visited websites

Rank	Website	Visits	Topic
1	www.tebyan.net	7,432	Cultural and Information
2	www.ninisite.com	7,233	Health
3	www.cloob.com	5,779	Social Network
4	www.persianv.com	4,184	Life skills
5	www.rasekhoon.net	3,962	Religious
6	www.aparat.com	2,408	video sharing
7	www.yjc.ir	2,170	News
8	www.kanoon.ir	2,137	Scientific
9	www.askdin.com	1,955	Religious
10	www.njavan.com	1,795	Scientific
11	www.delgarm.com	1,537	Entertainment
12	www.salamatnews.com	1,389	Health
13	www.bartarinha.ir	1,262	Life skills
14	www.askquran.ir	1,254	Religious
15	www.niniban.com	1,113	Health

C. Topical analysis

In this part we analyze question queries based on their topic. To do so, we need a query classification method to categorize them by their topic. We classified question queries based on an approach similar to [18]. We used an updated version of Hamshahri news dataset [19] as our corpus. Every news in this dataset has been categorized manually and we use these tags to categorize the questions. For each question we retrieve top 10 relevant documents from Hamshahri using Okapi BM25 probabilistic model [20]. The question query category is determined by using majority voting approach on the category of top 10 retrieved documents. *Table 1*Table 7 shows the top 10 relevant documents for the input query. As the majority of retrieved documents' category are related to health therefore the class of this question query is health.

Table 7 : Example of classification method

Rank	Document ID	Category
1	D1	Health
2	D2	Health
3	D3	Religion
4	D4	Economy
5	D5	Health
6	D6	Health
7	D7	Health
8	D8	Health
9	D9	Science
10	D10	Health

Having all question queries classified, we first check how these questions are distributed across different topics to find out the most interesting topic for Persian language users. As it can be seen from Figure 4 most of these questions concern health (20% of total question queries), policy (18% of total question queries), society and religion while few questions are posed concerning topics like environment and sports.

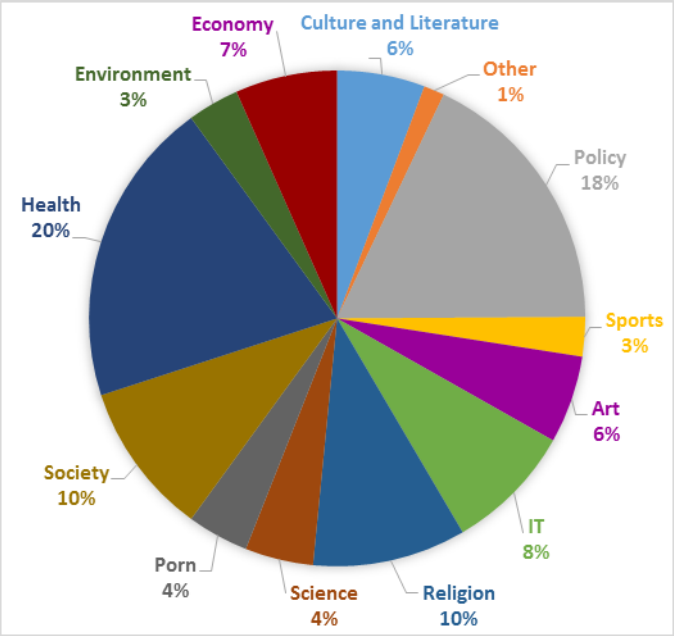

Figure 4 : Breakdown of categorized question queries.

For each topic, we calculated the average length of queries and the results are presented in Figure 5. According to this figure, the religious queries with average term 7.22, are the longest question queries meaning that people want to express their religious questions with more detail.

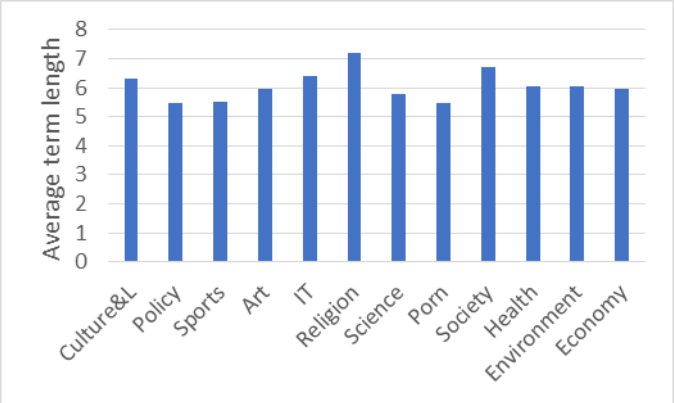

Figure 5 : Average query length for each topic

We now take a look at temporal aspect of these categories. Figure 6 shows how the number of related questions for each category changed during a year. Our investigation on question queries topic at month granularity revealed that health and policy were always on top during this period. Policy was one top during the first two and the last two months of the year while in the rest of the year health related news were on top. The reason that policy was on top during these months was Iranian Parliament election that took place in February. Another peak that was detected in policy question trend was in April when Barjam (Joint Comprehensive Plan of Action) was achieved.

Exploring religion question queries trend, at Ramadan(June-July), Muharram and Safar (November-December) which are the month with special virtues, we can detect increase in number of question queries being submitted to the search engine. The number of question queries for all four categories had decreased in March. This happened because of Iranian traditional Norooz holiday.

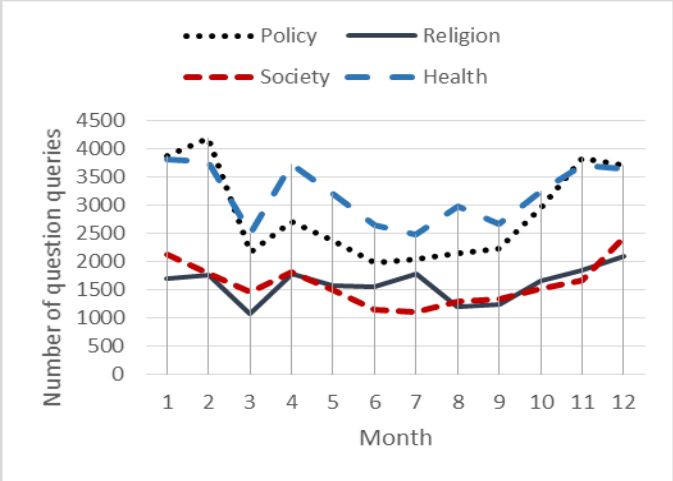

Figure 6 : Number of question posed regarding to each of four most popular questions' topic.

Figure 7 indicates the hourly distribution of the four most popular questions' topic. All these four categories have the same hourly distribution and from the 7 to 12 we can observe an increase in number of submitted questions while after that hour they start to decrease.

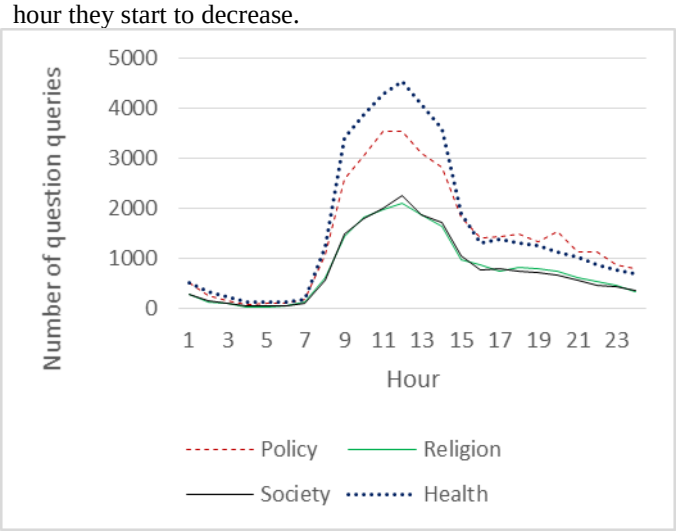

Figure 7 : Hourly analysis of four most popular questions' topic

Figure 8 gives information on the distribution of question query categories by daily hours over the entire dataset using heatmap diagram. As it can be seen from this figure, after health and policy which in all hours were the first two popular topics, Religion, Society and IT were competing for the third place.

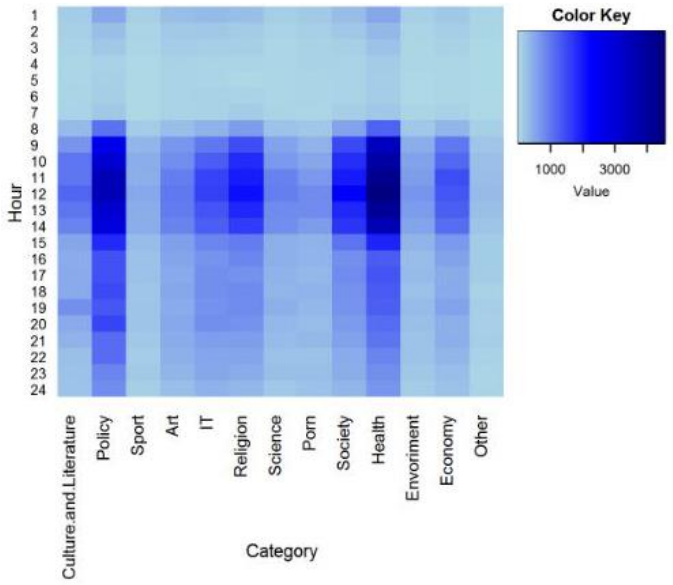

Figure 8 : Distribution of hourly question query volume over categories. The shading of grid cells represent the volume of question queries being posed to search engine during a specific hour considering special topic.

V. CONCLUSION

In this paper, we analyzed the characteristics of question queries submitted to a Persian search engine. Based on a large scale search log we investigate question queries at different levels. We started our analysis at query level and explored question queries characteristics and mainly focused on questions' form. We continued our analysis at click level and reviewed top 15 visited site regarding the results shown to users who asked a question from search engine. In the last part of our analysis, we focused on questions' topic. To classify our questions topically, we used an information retrieval method, using Hamshahri Dataset which contains categorized news. For each question we retrieved top 10 documents, and then used majority voting approach to decide the question's topic. Based on our analysis questions about Health and Policy were the most popular ones.

ACKNOWLEDGMENT *(Heading 5)*

We appreciate Parsijoo search engine for providing query logs with all necessary details and granting access. This study is conducted by support of Iran Telecommunication Research Center.

REFERENCES

[1] Jiang, Di, and Lingxiao Yang. "Query intent inference via search engine log." Knowledge and Information Systems (2016): 1-25.

[2] Palotti, João, et al. "How users search and what they search for in the medical domain." Information Retrieval Journal (2016): 1-36.

[3] Dumais, Susan, et al. "Understanding user behavior through log data and analysis." Ways of Knowing in HCI. Springer New York, 2014. 349-372.

[4] Lucchese, Claudio, et al. "Discovering tasks from search engine query logs." ACM Transactions on Information Systems (TOIS) 31.3 (2013): 14.

[5] Pang, Bo, and Ravi Kumar. "Search in the lost sense of query: Question formulation in web search queries and its temporal changes." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011.

[6] Völske, Michael, et al. "What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries." Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015.

[7] Aula, Anne, Rehan M. Khan, and Zhiwei Guan. "How does search behavior change as search becomes more difficult?." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2010.

[8] Bendersky, Michael, and W. Bruce Croft. "Analysis of long queries in a large scale search log." Proceedings of the 2009 workshop on Web Search Click Data. ACM, 2009

[9] Lau, Tessa, and Eric Horvitz. Patterns of search: analyzing and modeling web query refinement. Springer Vienna, 1999.

[10] Phan, Nina, Peter Bailey, and Ross Wilkinson. "Understanding the relationship of information need specificity to search query length." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

[11] Beitzel, Steven M., et al. "Varying approaches to topical web query classification." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

[12] Broder, Andrei. "A taxonomy of web search." ACM Sigir forum. Vol. 36. No. 2. ACM, 2002.

[13] Beeferman, Doug, and Adam Berger. "Agglomerative clustering of a search engine query log." Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

[14] Richardson, Matthew. "Learning about the world through long-term query logs." ACM Transactions on the Web (TWEB) 2.4 (2008): 21.

[15] Cai, Li, et al. "Large-scale question classification in cQA by leveraging Wikipedia semantic knowledge." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[16] Chan, Wen, et al. "Community question topic categorization via hierarchical kernelized classification." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[17] Bailey, Peter, et al. "Mining historic query trails to label long and rare search engine queries." ACM Transactions on the Web (TWEB) 4.4 (2010): 15.

[18] Ullegaddi, Prashant, and Vasudeva Varma. "A simple unsupervised query categorizer for web search engines." Proceedings of ICON-2010: 8th International Conference on Natural language Processing. 2011.

[19] AleAhmad, Abolfazl, et al. "Hamshahri: A standard Persian text collection." Knowledge-Based Systems 22.5 (2009): 382-387.

[20] Robertson, Stephen E., et al. "Okapi at TREC-3." NIST SPECIAL PUBLICATION SP 109 (1995): 109.