CORE
Metadata, citation and similar papers at core.ac.uk
Provided by UCT Computer Science Research Document Archive

Investigating Per-user Time Sensitivity Of Search Topics

Jivashi Nagar and Hussein Suleman

Department of Computer Science, University of Cape Town,
Private Bag X3, Rondebosch, 7701, Cape Town, South Africa.
{jnagar,hussein}@cs.uct.ac.za

Abstract. Search engines give the same results for the same query. They do not consider that a user's topics of interest may diverge at different times even if the query terms are the same. This paper presents the findings of a study into how different topics of interest of a user are influenced by time. The results show that most of the users have time sensitive search patterns, indicating that they have different topics of interest that are dominant at different times.

Keywords: Information Retrieval, Query log analysis, Topic Modelling, Users search behaviour, Time sensitive search patterns

1 Introduction

Search engines are used to search and retrieve information from the Web. The Web has information on almost every topic but the search engines do not consider diverging interests of a user and retrieve the same results for a query even if it is issued at different times.

An example of this could be a user who is a computer science student who also likes sports. So, it can't be said that (s)he will search only for the topics related to computer science. It is possible that at some point of time, (s)he will search for sports also. So, if (s)he issues a query ”tag”, during study hours, (s)he may be looking for HTML tags but, in some leisure time, the same query may mean Tag Sports Gear, a sporting goods brand.

Although the user queries are mostly small and ambiguous in nature [12], better results can be provided if a user's topics of interest and search patterns are known. Search patterns have been studied before based on the query logs. Query logs serve as an excellent store of knowledge as they have complete information about what the users have searched in a given time frame. It has been observed by previous studies that a user's search behaviour varies from workplace to home [19]. Change in topical categories and search query volume also varies with time [4][24][3]. According to these studies, after analysing the query log, they found that there is a pattern in search queries. But a general search pattern cannot be applicable to all users.

2 Research Question

Do the topics of interest of a user vary with time of the day?

Motivated by the observations of past studies, this study explored the time sensitive search pattern of a user to find his/ her different topics of interest that are dominant at different times. This study observed queries issued by 100 different users in an AOL query log [1] and analysed the query set of each one of the 100 users individually. The details and outcomes of this study are presented in this paper. Section 2 shows the related work on user's search behaviour, pattern identification and topic modelling. Section 3 covers the methodology of our work. Section 4 presents the analysis of the data. Section 5 includes limitations of the study. Section 6 is the conclusion of the work and future directions.

3 Related Work

3.1 User's Search Behaviour And Pattern Identification

It is crucial to analyse a user's search behaviour to provide effective and efficient search services. The query log data can be used to know how users use the search engines and also about their diverse interests and preferences [9]. Rose and Levinson [20] made an attempt to understand users' search goals. They analysed an Alta Vista query log and found that the goal of users' searches is less navigational and more resource seeking. In a work by Tyler and Teevan [22], the authors analysed repeated queries and user behaviour. According to this study, search engines can capitalize this re-finding behaviour of the user to improve the user's search experience. In the same line of re-finding behaviour, Tyler et al. [23] in their study found that repeated re-finding behaviour also contains diversification. According to Srivastva et al. [21], identifying the patterns in Web usage can be helpful for marketers in placing advertisements focusing on a certain target group. Temporal analysis of the sequence patterns can prove useful in finding the trending topics.

Temporal analysis of query logs has also been done by many researchers to explore users' search behaviour and search patterns. A significant outcome of the study by Rieh [19] is that the author was able to find the difference in search behaviour of the users. According to this study, users' search behaviour differs in their workplace from that at home. The websites visited during working hours were mostly related to their work while, at home, the search was of diverse nature. In a similar work [25], Yuye and Alistair analysed an MSN query log. They found a general pattern in the volume of queries. There was a peak early in the week that dropped steadily until Friday and decreased sharply over the weekend. This pattern speaks about the weekly routine of a common working person. They also observed an hourly pattern and found a rise in volume of queries from early morning, peaking at noon and decreasing steadily through midnight. Judit et al. [3] analysed the same MSN query log for topic specific analysis. They found that, during weekdays, queries related to work were dominant. In a comparatively recent work by Michael et al. [24], the authors analysed a Russian query log spanning one year. According to this study, queries related to categories like "Health" and "Beauty and Style" were distributed more or less constantly throughout the year. Some categories like "Education" observed a drop during vacation periods. John Cosley [7] has shown that the query terms and search patterns of users vary with time of the day and also with device (Mobile and PC). He also compared the search patterns of weekdays and weekends. According to this study, queries regarding task completion were dominant during the morning on weekdays, while entertainment and shopping related queries have shown their dominance in the evening on Mobiles and Tablets. All these studies have analysed the query logs as a whole and suggested that there is a pattern in users' search behaviour. Some of them [24][3][7] also analysed the time dependent popularity of some topics. They did not consider and analyse each individual user's search pattern. A common trend of topic change and popularity cannot be applied to improve the search experience of an individual user. Each user may have a specific search pattern, which is different from the others. As reported by Michael et al. [24], the queries related to the category "Education" observed a decline during the vacation period; this trend or pattern cannot be generalized for each user. A user, who is looking for extra classes or lessons, may search for "education" even in the vacation period.

3.2 Topic Inference

Every user has different topics of interests. To find the different topics from query logs, most of the previous studies about topic based personalized information retrieval systems have relied on Open Directory Project(ODP) categories [16][14] [5][16]. Jansen et al. [11] used the Google Directory topical hierarchy to classify the queries into subject categories. Arguably, Web search is not limited to these categories because of the rich nature of the Web. It is not ideal to put the wide range of a user's interests into predefined categories. Unsupervised Machine Learning may be a better tool to learn latent topics from users' search queries [15].

4 Methodology

The goal of this study is to investigate the time sensitivity of a user's topics of interest. We analysed queries submitted by each user separately to explore the time sensitive search pattern of that user. This section presents the details of our study in the following steps:

4.1 Data Collection

In this study, the search history of 100 users from an AOL query log [1] has been analysed. The AOL query log is publicly available log data for research and analysis [1][18]. The AOL query log has been analysed before by many researchers.

Duarte et al. [8] identified queries with children intent from this AOL query log. This log collection contains about 20M Web queries from 650K users issued in three months from March 2006 to May 2006. The data is anonymized and consists of: UserID, Query, QueryTime, ClickedRank, DestinationDomainUrl. Each UserID represents a unique user. For this study, UserID, Query and QueryTime fields of the query log were considered. It was assumed that each UserID is representing a unique user. The logs of each unique user were cleaned and pre-processed for further analysis. The details of data cleaning and pre-processing are described below.

4.2 Data cleaning

The queries of each of the 100 users were processed individually. The entries with empty queries were removed. The same queries that were submitted on the same date within a time difference of less than 10 seconds were also not considered for analysis. According to Odjik et al. [17], queries issued within a few seconds of time are more likely to be spelling correction or substitution type of formulations of the previous query issued. These queries were not considered as different queries but some modification of the previous ones. This process of removing incomplete, irrelevant or duplicate data is called data cleaning.

4.3 Pre-processing

Pre-processing, also known as text normalization, gives a syntactical view of the original text. Pre-processing was accomplished by using the Natural Language Toolkit (NLTK). NLTK is a free and open source community-driven project [10]. It is the most used platform to work with natural human language data. This involves tokenization, stopword and punctuation marks removal, and lemmatization.

The process of dividing a phrase or sentence into tokens is called tokenization. The tokens may represent words, digits or punctuation marks. After tokenization, stopwords were detected from the data. Stopwords are common and high frequency words that are independent of any topic like a, an, the, for and and. Following detection of stopwords and punctuation marks, they were removed from the query terms and the query terms were lemmatized.

Lemmatization aims to remove inflectional endings and to return the base or dictionary form of a word, which is called the lemma. This step utilizes vocabulary along with a morphological analysis of words.

Table 1 shows the aggregate number of queries of 100 users and the queries that remained for analysis after cleaning and pre-processing.

4.4 Topic modelling

Topic modelling is a technique that is used to identify the latent topics present in a corpus. Topic models are the algorithms that are used to find the main themes

Table 1. Queries analysed

Original number of queries	Queries remained after cleaning
22096	6850

or ideas in a data collection. In a number of previous works, authors have utilized predefined topical categories like the Open Directory Project [13][5][16] and the Google Directory topical hierarchy [11] to find the topics of interest of a user. According to Mehrotra [15], to learn the latent topics of interest from the user search logs, unsupervised machine learning would be a better tool.

Latent Dirichlet Allocation (LDA) is an unsupervised approach for topic modelling. It is a generative probabilistic mode for a text corpus [6] and the most commonly used approach for topic modelling. LDA is a three-level hierarchical Bayesian model. In LDA, each document of a corpus is modeled as a finite mixture over an underlying set of topics and each topic is modeled as an infinite mixture over an underlying set of topic probabilities [6].

In this study, we assumed that each user has at least 4 different topics of interest. After applying LDA on the queries of each user, topics were assigned to the individual queries. According to LDA, a document may have more than one topic. So, in this case, if a query fell under more than one topic, all those topics were assigned to that query. The reason behind doing this is that most of the Web queries are ambiguous in nature i.e. they may belong to many topics. The aim of this study is to find the temporal dominance of topics of interest of a user. We analysed which topics were dominant at a particular time. This notion of assigning multiple topics to a query along with finding temporal dominance of topics can be utilized to disambiguate the ambiguous queries. After assigning the topic(s), the queries were grouped according to the Time-bins.

4.5 Time-bins

The aim of this study is to find time sensitive search patterns. We analyse the relation between a topic of interest and the time when it is searched dominantly. According to Rieh [19], users search behaviour differs in their workplace from that at home. For the purpose of our study, we divided the time of a day into four customized bins according to the common daily routine of a working person. The four Time-bins are: Early morning (6h-8hr), Working hours (8h-18h), Evening time (18h-24h) and Midnight (24h-6h).

5 Results and Analysis

In this section, search patterns of all of the 100 users were analysed. Keeping in mind that this AOL query log is from 2006 i.e. more than 10 years old, people searched less frequently because of Internet availability and usage cost. People used to search for fewer topics. In present times, due to fast Internet connections, easy availability and more advanced communication devices, users search more frequently. Moreover, the number of topics of interest has also increased. In spite of this limitation, some promising facts are revealed from its analysis. Table 2 shows the number of Time-bins used by the users to issue queries. As can be seen from Table 2, out of 100 users, the majority of the users (50) have utilized 2 Time- bins to issue their queries. 35 users have made their queries in 3 Time-bins while 12 users have searched in 4 Time-bins. Very few users (3) have searched only in one Time-bin. Table 3 presents the time-sensitive search patterns of users along with their respective numbers of dominant topics. Every unique number in Example Pattern indicates the different dominant topic in a user's search pattern. For instance, pattern 123 represents the search pattern of a user who has searched in 3 Time-bins and in every Time-bin the dominant topic is different. One of the possible pattern, 1122, when a user searches for 2 dominant topics in 4 Time-bins, was not found in any of the 100 patterns analysed. Out of 100 users, 1 user has searched for 4 different dominant topics in the 4 Time-bins, 13 users have searched 3 Time-bins with 3 distinct dominant topics and 42 users have 2 different dominant topics for the 2 Time-bins they searched in. So, 56 users have searched for different dominant topics in every Time-bin. Among the users who searched in 4 Time-bins, 6 have 3 dominant topics and 3 have 2 dominant topics. 19 users have 2 dominant topics in 3 Time-bins they searched. Thus, there are 28 users who have at least either 3 or 2 dominant topics of interest. Only 13 users have shown the same dominant topic in every searched Time-bin and, as shown in Table 2, 3 users have searched in only 1 Time-bin.

Table 2. Number of Time-bins used

Number of Users	Number of Time-bins Used
12	4
35	3
50	2
3	1

The following figures represent search patterns of some of the categories shown in Table 3.

Figure 1 shows the search pattern of Category 2 users. Each time the user has searched, the dominant topic is different. Topic3 which is not searched much at Time1 and Time3, becomes dominant at Time4.

Figure 2 represents the search pattern of a Category 4 user. The user has searched only in 2 Time-bins but, in both the Time-bins, the dominant topics are different. Topic2, which has not been searched at Time3, is dominating at Time4. Figure 3 shows the search pattern of a Category 5 user who has made searches in 3 Time-bins. At Time1, only Topic0 is searched and it also dominates at Time3. Topic2, which has not been searched either at Time1 or at Time3, clearly dominates at Time4.

Table 3. Dominant Topic patterns and calculated entropy

Category	Number of users	Number of Dominant Topics	Example Pattern	Entropy
1	1	4	1234	1.386
2	13	3	123	1.098
3	6	3	1233	1.039
4	42	2	12	0.693
5	18	2	122	0.636
6	4	2	1333	0.215
7	16	1	1111,111,11,1	0

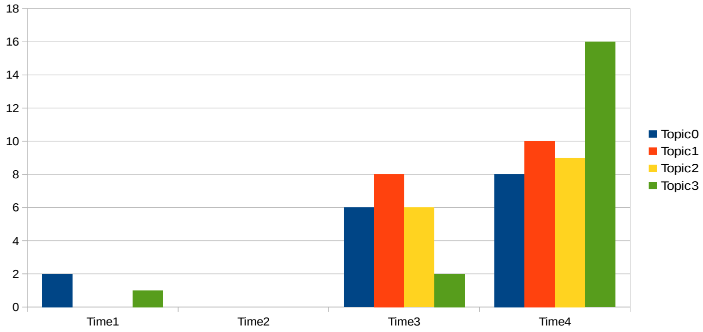

Fig. 1. Search pattern of Category 2

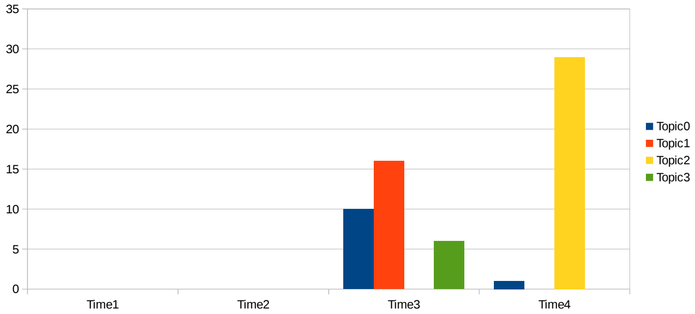

Fig. 2. Search pattern of a Category 4 user

Figure 4 shows the search pattern of a Category 6 user. The user has searched Topic0 and Topic1 in all the 4 Time-bins. While Topic1 is dominating in 3 Time-

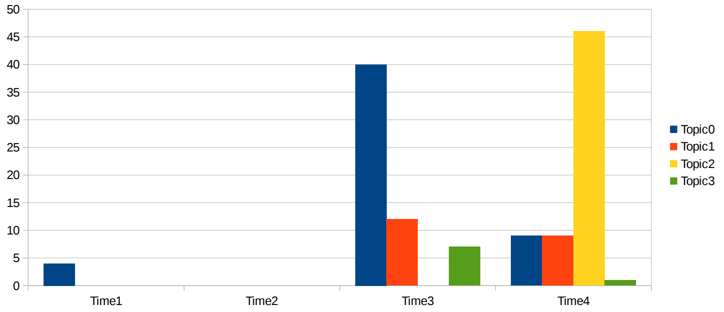

Fig. 3. Search pattern of a Category 5 user

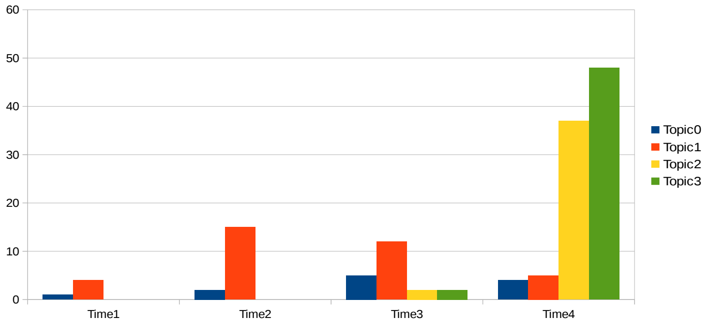

Fig. 4. Search pattern of a Category 6 user

bins (Time1, Time2 and Time3), at Time4, Topic3 is dominating followed by Topic2.

Figure 5 shows the search pattern of a Category 7 user. The Category represents the users who have the same dominant topic in every Time-bin whether they searched in 4, 3 or 2 Time-bins.

These search patterns indicate that users have different topics of interest and they prefer to search about them at different times.

For analysing the variability in patterns of dominant topics, entropy was calculated. Entropy is a measure of disorder or randomness and refers to the number of possible states a variable can have. It is calculated as:

$$H(X) = -\sum_{i=1}^{n} l_n p(x_i).p(x_i)$$

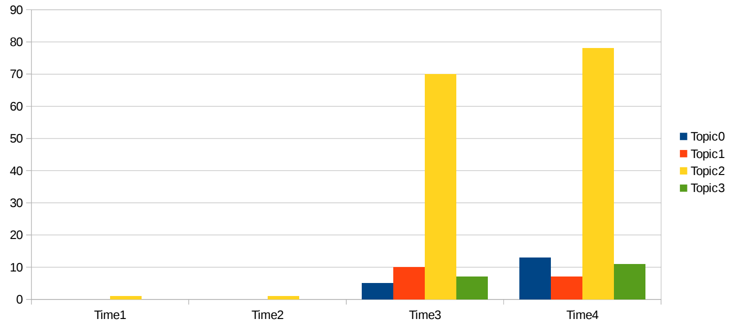

Fig. 5. Search pattern of a Category 7 user

A greater value of entropy points to more possible states or randomness of a variable. Table 3 shows the calculated entropy for different patterns of dominant topics, suggesting an ordering and grouping of different topic patterns based on variability.

80 users have entropy greater than or equal to 0.636. Even if a user searched in at least 3 Time-bins, the dominant topics were unique in 2 Time-bins. In other words, some topics are dominantly searched in a particular time interval. It clearly means that there is variability and uncertainty in a user's topics of interest. A user's topics of interest differ in different time intervals and so we can say that the topics of interest of a user are time-sensitive.

This inference can be utilized to disambiguate the queries and provide more useful and relevant search results.

6 Limitations

The aim of this study is to explore the temporal dominance of topics of interest of a user. For this purpose, we selected the queries submitted by 100 users from an AOL query log. The time of the day was divided into four bins and it is assumed that each user has at least four topics of interest. After processing and analysing, it was found that most of the users search for different topics at different times. It can be said that topics are time sensitive.

This study was able to find the time-sensitive search patterns of a user but it has some shortcomings also.

1. The query log data was old but readily available for analysis.
2. The queries of only 100 users were analysed because the query set of every user was cleaned manually.
3. It did not figure out the exact number of topics of interest of each user. As we divided the time of the day into 4 Time-bins according to the routine

of a common working person, it was assumed that each user has at least 4 topics of interest for 4 time bins.

4. Although we were able to find search patterns based on these Time-bins, each user may have different search Time-bins.

7 Conclusions and Future work

We have studied an AOL query log to explore the time sensitive search patterns of users. We have analysed the queries of 100 different users and found that, out of 100 users, 84 users have at least 2 different dominant topics searched at different times. Only 13 users have searched for the same topic in every Time-bin and 3 users have searched only in 1 Time-bin. This study concludes that most of the users have time sensitive topics. They search for different topics at different times. Different topics are dominant at different time intervals. The goal of future work is to explore and exploit the time sensitive search patterns of a user to model a user's time sensitive search behaviour, which could prove helpful to search engines in disambiguating the short and ambiguous queries and also providing users with more relevant search results.

References

1. http://www.cim.mcgill.ca/ dudek/206/Logs/AOL-user-ct-collection/user-ct-test-collection-01.txt/.
2. Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi. Preprocessing in natural language processing. *Editorial board*, 152, 2013.
3. Judit Bar-Ilan, Zheng Zhu, and Mark Levene. Topic-specific analysis of search queries. In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 35–42, New York, NY, USA, 2009. ACM.
4. Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 321–328, New York, NY, USA, 2004. ACM.
5. Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 185–194. ACM, 2012.
6. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
7. John Cosley. Hearing the rhythms of human search behavior: What weve learned. http://searchengineland.com/human-behavior-influences-search-marketing-197486/, July 2014.
8. Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. Query log analysis in the context of information retrieval for children. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 847–848. ACM, 2010.

9. Yi Fang, Naveen Somasundaram, Luo Si, Jeongwoo Ko, and Aditya P Mathur. Analysis of an expert search query log. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1189–1190. ACM, 2011.
10. Mansi Gera and Shivani Goel. Data mining - techniques, methods and algorithms: A review on tools and their validity. *International Journal of Computer Applications*, 113(18), 2015. Copyright - Copyright Foundation of Computer Science 2015; Last updated - 2015-04-14.
11. Bernard J Jansen, Zhe Liu, Courtney Weaver, Gerry Campbell, and Matthew Gregg. Real time search on the web: Queries, topics, and economic value. *Information Processing & Management*, 47(4):491–506, 2011.
12. Bernard J Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*, 36(2):207–227, 2000.
13. Hyoung R Kim and Philip K Chan. Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 101–108. ACM, 2003.
14. Jin Young Kim, Kevyn Collins-Thompson, Paul N Bennett, and Susan T Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222. ACM, 2012.
15. Rishabh Mehrotra. Topics, tasks & beyond: Learning representations for personalization. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 459–464. ACM, 2015.
16. Ashish Nanda, Rohit Omanwar, and Bharat Deshpande. Implicitly learning a user interest profile for personalization of web search using collaborative filtering. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 2, pages 54–62. IEEE, 2014.
17. Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. Struggling and success in web search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1551–1560. ACM, 2015.
18. Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, InfoScale '06, New York, NY, USA, 2006. ACM.
19. Soo Young Rieh. Investigating web searching behavior in home environments. *Proceedings of the American Society for Information Science and Technology*, 40(1):255–264, 2003.
20. Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM.
21. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, January 2000.
22. Sarah K Tyler and Jaime Teevan. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 191–200. ACM, 2010.
23. Sarah K Tyler and Yi Zhang. Multi-session re-search: in pursuit of repetition and diversification. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2055–2059. ACM, 2012.

24. Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. What users ask a search engine: Analysing one billion russian question queries. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1571–1580, New York, NY, USA, 2015. ACM.
25. Yuye Zhang and Alistair Moffat. Some observations on user search behaviour. *Austr. J. Intelligent Information Processing Systems*, 9(2):1–8, 2006.