Chapter 10
SogouQ: The First Large-Scale Test Collection with Click Streams Used in a Shared-Task Evaluation

Ruihua Song, Min Zhang, Cheng Luo, Tetsuya Sakai, Yiqun Liu, and Zhicheng Dou

Abstract Search logs are very precious for information retrieval studies. In this chapter, we will introduce a real Chinese query log dataset, SogouQ, which was released by SogouQ corporation in 2010 for the NTCIR-9 Intent task. SogouQ contains more than 30 million clicks collected in 2008. It is the first large-scale query logs used in a shared-task evaluation (i.e., the NTCIR tasks). SogouQ has been adopted in a number of follow-up evaluation tasks, NTCIR-10 Intent-2, NTCIR-11 IMine, NTCIR-12 IMine-2, as well as in several Chinese domestic tasks. Moreover, SogouQ has a broader impact on other research areas, such as natural language processing and social science. It has been acquired by more than 200 institutions.

R. Song (✉)
Microsoft XiaoIce, Beijing 100080, China
e-mail: rsong@microsoft.com

M. Zhang · Y. Liu
Tsinghua University, Beijing 100084, China
e-mail: z-m@tsinghua.edu.cn

Y. Liu
e-mail: yiqunliu@tsinghua.edu.cn

C. Luo
MegaTech.AI, Beijing 100080, China
e-mail: luochengleo@gmail.com

T. Sakai
Waseda University, Shinjuku-ku Okubo 3-4-1 63-05-04, Tokyo 169-8555, Japan
e-mail: tetsuyasakai@acm.org

Z. Dou
Renmin University of China, Beijing 100872, China
e-mail: dou@ruc.edu.cn

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_10

10.1 Introduction

When we were preparing the NTCIR-9 Intent task that aims to investigate query intents and search result diversification (Song et al. 2011) in 2010, Sogou corporation was so generous to provide a real Chinese query log to NTCIR participants and further research communities. The data is called SogouQ and contains 30 million clicks collected in 2008. It is the first large-scale query logs used in a shared-task evaluation, such as NTCIR tasks.

The NTCIR-9 Intent task attracted 16 teams for Subtopic Mining subtask and 8 teams for Document Ranking subtask. It became the largest track in NTCIR-9 partially because participants are interested in SogouQ and how to use query logs for mining intents and diversifying document ranking. Since then SogouQ is used for NTCIR-10 Intent-2 task (Sakai et al. 2013), NTCIR-11 IMine task (Liu et al. 2014), and NTCIR-12 IMine-2 task (Yamamoto et al. 2016). The total number of participants groups is more than 80. They are from Australia, Canada, China, Germany, France, Japan, Korea, Spain, UK, and United States.

Later SogouQ had an even bigger impact on research. The usage of SogouQ data collection goes beyond the research on query intent. SogouQ is also used for improving fundamental natural language processing modules, such as name entity identification and new word discovery, user behavior studies, and Sociological topics. More than 200 institutes have acquired SogouQ related datasets from Tsinghua-Sohu Joint Laboratory on Search Technology. We believe that a more practical impact has happened but not been reported.

The remainder of this chapter is organized as follows: Sect. 10.2 describes the details of SogouQ and its related data collections. Section 10.3 briefly describes how organizers and participants use SogouQ in the NTCIR tasks. Section 10.4 reports more research impact beyond the works published in NTCIR proceedings. Section 10.5 concludes this chapter.

10.2 SogouQ and Related Data Collections

SogouQ was constructed by the Tsinghua-Sohu Joint Lab on Search Technology. It is a web query log of Sogou search engine for about one month (June 2008). There are about 30 million clicks included. The size of compressed SogouQ is about 1.9 gigabytes and is available for download.[1]

It should be noted that several similar click datasets were also released by several organizations for research purpose:

- **AOL Query logs** (2006/36M queries/English) includes user ids and click data. This dataset was intentional and intended for research purposes. However, the queries were not filtered and further lead to much controversy about privacy issues.

[1] http://www.sogou.com/labs/resource/q.php.

- **MSN Query logs** (2006/100M queries/English) includes session ids and click-through information, but not user ids (Craswell et al. 2009).
- **Yandex Query logs** (unknown time/210M queries/Russian) includes user sessions extracted from Yandex logs, with user ids, queries, query terms, URLs, their domains, URL rankings, and clicks. However, the user data is fully anonymized.[2]

The data format of SogouQ is as follows:

[Access time]\t[User ID]\t[Query]\t[Rank of the URL in the returned result]\t[The sequence number of user click]\t[URL that user clicked]\n

Here User ID is automatically assigned according to the cookie information when a user accesses the search engine by using the browser. Different queries that are input by the same browser correspond to the same user ID.

Compared to other search log data, SogouQ has several advantages. First, User ID and access time can provide information on sessions, which is important for session-based retrieval or mining-related searches by session. Second, in addition to the clicked URL, SogouQ provides the rank of clicked URL when it was shown to the user and which sequence the user clicked URLs for a query. Such information is valuable for research on user click modeling. Third, if we have only URLs, the content of URLs is difficult to obtain because the web keeps evolving. URLs may expire or the content of some URLs may change. Fortunately, Sogou released a document collection called SogouT[3] in 2010, which were crawled in June 2008. Therefore, researchers can get the corresponding page content at the same time.

We appreciate Sogou corporation and Tsinghua-Sohu Joint Lab of Search Technology. Due to their deep understanding of search and courage, research communities can have such valuable data collections.

10.3 SogouQ and NTCIR Tasks

The NTCIR-9 Intent task comprises the Subtopic Mining subtask (given a query, output a ranked list of possible subtopic strings) and the Document Ranking subtask (given a query, output a ranked list of URLs that are selectively diversified). In the Subtopic Mining subtask, a subtopic could be a specific interpretation of an ambiguous query (e.g., "microsoft windows" or "house windows" in response to "windows") or an aspect of a faceted query (e.g., "windows 7 update" in response to "windows 7"). The subtopics collected from participants were pooled, manually clustered, and thereby used as a basis for identifying the search intents of the query. The probability of each intent given the query was estimated through assessor voting. In the Document Ranking subtask, in contrast to traditional relevance assessments where the assessors determine the relevance of each pooled document with respect to a topic, we required the assessor to provide graded relevance assessments with

[2] https://www.kaggle.com/c/yandex-personalized-web-search-challenge/data.

[3] http://www.sogou.com/labs/resource/t.php.

respect to each intent of a given query. Finally, the relevance and diversity of the ranked subtopics or documents were evaluated using diversified information retrieval metrics (Sakai and Song 2014).

SogouQ was used by every participant for mining subtopics for given queries or estimating the importance of subtopics according to the number of clicks (Han et al. 2011; Wang et al. 2013; Xue et al. 2011; Yu and Ren 2014). The subtopics and their importance will influence document ranking then. Thus when user queries and clicks are introduced to the subtopic pool via SogouQ, our manually labeled intents or documents model the information needs of real users more accurately. Such an evaluation benchmark helps research on information retrieval in universities or labs without commercial search engines as experimental platforms.

In NTCIR-10 Intent-2 task, organizers provide the following instruction on subtopic:

> A subtopic string of a given query is a query that specialises and/or disambiguates the search intent of the original query. If a string returned in response to the query does neither, it is considered incorrect.
>
> e.g. original query: "harry potter" (underspecified) subtopic string: "harry potter philosophers stone movie" incorrect: "harry potter hp" (doe not specialise)
>
> It is encouraged that participants submit subtopics of the form "<originalquery> <additionalstring>"

Assessors were asked to provide a label for each intent cluster in the form "<originalquery><additionalstring>". Such a change provides valuable data to better understand a query in the perspective of two intent roles, i.e., kernel-object and modifier (Ren and Yu 2016; Yu and Ren 2012; Zheng et al. 2018). In contrast to the NTCIR-9 Intent task where we had up to 24 intents for a single topic, organizers of Intent-2 decided to select up to 9 intents per topic based on votes because search result diversification is mainly about diversifying the first search result page, which can only accommodate around ten URLs.

NTCIR-11 IMine task continued Subtopic Mining subtask and Document Ranking subtask and started a new subtask called TaskMine, which aims to explore the methods of automatically finding subtasks of a given task (e.g., for a given task "lose weight", the possible outputs can be "do physical exercise", "take calories intake", "take diet pills", etc.). In the Subtopic Mining subtask, participants are expected to generate a two-level hierarchy of underlying subtopics by analysis into the provided document collection, user behavior data including SogouQ, or other kinds of external data sources. For example, given the ambiguous query "windows", the first-level subtopic may be "microsoft windows", "software on windows platform", or "house windows". In the category of "microsoft windows", users may be interested in different aspects (second-level subtopics), such as "windows 8" and "windows update". The hierarchical structure of subtopics is closely related with the knowledge graph. However, the hierarchical subtopics here are used to describe users' possible information needs instead of the manually created knowledge structure of entity names. Organizers encouraged participants not to use the graph directly even when a knowledge graph exists for a given query. Therefore, user behavior data, such as SogouQ,

play important roles in creating the hierarchy of subtopics as real user queries reflect users' possible information needs.

NTCIR-12 IMine-2 task focuses on vertical intents behind a query as well as its topical intents because many commercial Web search engines merge several types of search results and generate a SERP (search engine results page) in response to a user's query. For example, the results of query "flower" now may contain image results and encyclopedia results as well as usual Web search results. We refer to such "types" of search results as verticals. Accordingly, the IMine-2 task comprises two subtasks: the Query Understanding subtask and the Vertical Incorporating subtask. The Query Understanding subtask is a successive task of the Subtopic Mining subtask but the difference is that participants are asked to identify the relevant verticals for each subtopic. For example, for the query "iPhone 6", a possible result list of the Query Understanding subtask is:

[tid] [subtopic] [vertical] [score]
IMINE2-E-000 iPhone 6 apple.com Web 0.9
IMINE2-E-000 iPhone 6 sales News 0.90
IMINE2-E-000 iPHone 6 photo Image 0.88
IMINE2-E-000 iPhone 6 review Web 0.78

The Vertical Incorporating subtask is also a successive task of the Document Ranking subtask. The difference is that the participants should decide whether the result list should contain vertical result or not. SogouQ is still a useful resource of user behaviors for Chinese subtasks. Similarly, Yahoo! Japan provides the participants of Japanese subtasks a Web search related query data, which is generated from the query log of Yahoo! Japan Search from July 2009 to June 2013.[4]

10.4 Impact of SogouQ

As by April 30, 2019, we can find 82 papers when we search the keyword "SogouQ" in Google Scholar.[5] Most of them are not published in NTCIR proceedings.

Some works such as Gu et al. (2016), Han et al. (2011), Ren et al. (2015), Xue et al. (2011), Kim and Lee (2015), and Zheng et al. (2015) use SogouQ to mine subtopics (Song et al. 2018; Wang et al. 2013; Yu and Ren 2014), or suggestions (Li and Wang 2014; Liu et al. 2017; Shu et al. 2013). Some works like Zheng et al. (2018) use SogouQ for better understanding a query in the perspective of two intent roles, i.e., kernel-object and modifier (Ren and Yu 2016; Yu and Ren 2012). Some other works investigate intent shifting (Wang and Chen 2011), query specification (Xiangbin et al. 2015), and search task identification (Du et al. 2018). Some works use SogouQ for improving some fundamental modules of natural language processing, such as unsupervised dependency parsing (Qiao et al. 2016), new word identification

[4] http://research.nii.ac.jp/ntcir/news-20150717-ja.html.

[5] http://scholar.google.com.

(Xuewei 2014), and person name recognition (Lv et al. 2013; Wen et al. 2013). Moreover, the rich information of SogouQ provides evidence to get statistics, e.g., query per second (Fang et al. 2017), sample queries (Liu and Li 2014); or mine a particular type of queries, e.g., time-sensitive search queries (Pei et al. 2016) and health search queries; or predict authoritative of website (Yu and Ren 2018).

Some usage of SogouQ is on broader research topics. Rao et al. (2014) constructs query co-occurrence network from SogouQ and compares the network with Named Entity Person co-occurrence network and the network based on the co-occurrence of words in sentences of news articles; Wang and Pleimling (2017) use it to investigate foraging patterns in online searches. Authors analyze three different click-through logs and discover an increased efficiency of the search engines. In the language of foraging, the newer logs indicate that online searches overwhelmingly yield local searches (i.e., on one page of links provided by the search engines), whereas for the older logs, the foraging processes are a combination of local searches and relocation phases that are power law distributed. It follows that good search engines enable the users to find the information they are looking for through a local exploration of a single page with search results, whereas for poor search engine, users are often forced to do a broader exploration of different pages.

According to the statistics from Tsinghua-Sohu Joint Lab on Search Technology, more than 200 institutions have acquired SogouQ related datasets. We believe that a more practical impact has happened but not been reported.

10.5 Conclusion

The problems that are explored in NTCIR Intent and IMine tasks require a data collection of query logs. With the great support of Sogou corporation, SogouQ becomes the first query logs that are used in a shared evaluation. Compared to other query logs, SogouQ has richer information on session, ranking, and orders of clicks, and corresponding documents if being combined with SogouT. Therefore, SogouQ does not only support research on query understanding of intent and vertical, but also enable many works on broader research topics on web search user behaviors. More than 200 institutes have acquired SogouQ data and they are using the query logs for various research and applications.

As query logs are too sensitive, it is difficult to obtain more shared query logs. Some efforts were done to simulate click-through data, such as Sogou-QCL (Zheng et al. 2018), to enable the neural-based works that need a larger amount of data.

References

Craswell N, Jones R, Dupret G, Viegas E (eds) (2009) Proceedings of the 2009 workshop on web search click data (WSCD'09). ACM, New York, NY, USA

Du C, Shu P, Li Y (2018) CA-LSTM: search task identification with context attention based LSTM. In: The 41st international ACM SIGIR conference on research & development in information retrieval. ACM, pp 1101–1104

Fang Z, Yu T, Mengshoel OJ, Gupta RK (2017) Qos-aware scheduling of heterogeneous servers for inference in deep neural networks. In: Proceedings of the 2017 ACM on conference on information and knowledge management. ACM, pp 2067–2070

Gu J, Feng C, Gao X, Wang Y, Huang H (2016) Query intent detection based on clustering of phrase embedding. In: Chinese national conference on social media processing. Springer, pp 110–122

Han J, Wang Q, Orii N, Dou Z, Sakai T, Song R (2011) Microsoft research Asia at the NTCIR-9 INTENT task. In: Proceedings of NTCIR-9

Kim SJ, Lee JH (2015) Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents. Inf Process Manag 51(6):773–785

Li L, Wang H (2014) Multi-strategy query expansion method based on semantics. J Digit Inf Manag 12(3)

Liu C, Li Y, (2014) Non-iteration parallel algorithm for frequent pattern discovery. In: 2014 13th international symposium on distributed computing and applications to business, engineering and science. IEEE, pp 127–132

Liu J, Li Q, Lin Y, Li Y (2017) A query suggestion method based on random walk and topic concepts. In: 2017 IEEE/ACIS 16th international conference on computer and information science (ICIS). IEEE, pp 251–256

Liu Y, Song R, Zhang M, Dou Z, Yamamoto T, Kato MP, Ohshima H, Zhou K (2014) Overview of the NTCIR-11 IMine task. In: Proceedings of NTCIR-11

Lv X, Wu R, Wen B (2013) Chinese personal name recognition in web queries via bootstrapping. In: 2013 9th international conference on computational intelligence and security. IEEE, pp 415–419

Pei J, Huang D, Ma J, Song D, Sang L (2016) Dut-nlp-ch@ NTCIR-12 temporalia tid subtask. In: Proceedings of NTCIR-12

Qiao X, Cao H, Zhao T (2016) Improving unsupervised dependency parsing with knowledge from query logs. ACM Trans Asian Low-Resour Lang Inf Process (TALLIP) 16(1):3

Rao L, Luo Z, Tang J, Wang T (2014) Research on the query co-occurrence networks. Management innovation and information technology, vol 61, pp. 275

Ren F, Yu H (2016) Role-explicit query extraction and utilization for quantifying user intents. Inf Sci 329:568–580

Ren P, Chen Z, Ma J, Wang S, Zhang Z, Ren Z (2015) Mining and ranking users' intents behind queries. Inf Retr J 18(6):504–529

Sakai T, Song R (2014) Evaluating diversified search results using per-intent graded relevance. In: ACM SIGIR 2011

Sakai T, Dou Z, Yamamoto T, Liu Y, Zhang M, Song R, Kato M, Iwata M (2013) Overview of the NTCIR-10 INTENT-2 task. In: Proceedings of NTCIR-10

Shu B, Niu Z, Jiang X, Mustafa G (2013) A novel query suggestion method based on sequence similarity and transition probability. In: Proceedings of the international conference on data mining (DMIN), The Steering Committee of The World Congress in Computer Science, p 1

Song R, Zhang M, Sakai T, Kato MP, Liu Y, Sugimoto M, Wang Q, Orii N (2011) Overview of the NTCIR-9 INTENT task. In: Proceedings of NTCIR-9

Song W, Liu Y, Liu Lz, Wang Hs (2018) Semantic composition of distributed representations for query subtopic mining. Front Inf Technol Electron Eng 19(11):1409–1419

Wang CJ, Chen HH (2011) Intent shift detection using search query logs. Int J Comput Linguist Chin Lang Process 16(3–4)

Wang Q, Qian Y, Song R, Dou Z, Zhang F, Sakai T, Zheng Q (2013) Mining subtopics from text fragments for a web query. Inf Retr 16(4):484–503

Wang X, Pleimling M (2017) Foraging patterns in online searches. Phys Rev E 95(3):032145
Wen B, Xiao S, Luo Y, LV X, (2013) Unsupervised Chinese personal name recognition using search session. J Comput Inf Syst 9(6):2201–2208
Xiangbin T, Wei L, Xiaojuan Z, Shihao H (2015) Feature analysis and automatic identification of query specificity. Data Anal Knowl Discov 31(2):15–23
Xue Y, Chen F, Zhu T, Wang C, Li Z, Liu Y, Zhang M, Jin Y, Ma S (2011) THUIR at NTCIR-9 INTENT task. In: Proceedings of NTCIR-9
Xuewei L, Xueqiang L, Kehui L (2014) Chinese new words identification from query log by extending the context. Data Anal Knowl Discov 30(11):59–65
Yamamoto T, Liu Y, Zhang M, Dou Z, Zhou K, Markov I, Kato MP, Ohshima H, Fujita S (2016) Overview of the NTCIR-12 IMine-2 task. In: Proceedings of NTCIR-12
Yang H, Feng Y (2018) Authoritative prediction of website based on deep learning. In: 2018 IEEE 4th international conference on big data computing service and applications (BigDataService). IEEE, pp 208–212
Yu H, Ren F (2012) Role-explicit query identification and intent role annotation. In: Proceedings of the 21st ACM international conference on information and knowledge management. ACM, pp 1163–1172
Yu HT, Ren F (2014) Subtopic mining via modifier graph clustering. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 337–347
Zhang X, Han S, Lu W (2018) Automatic prediction of news intent for search queries: an exploration of contextual and temporal features. Electron Libr 36(5):938–958
Zhang Z, Sun L, Han X (2015) Learning to mine query subtopics from query log. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: short papers), vol 2, pp 341–345
Zheng Y, Fan Z, Liu Y, Luo C, Zhang M, Ma S (2018) Sogou-QCL: a new dataset with click relevance label. In: The 41st international ACM SIGIR conference on research & development in information retrieval. ACM, pp 1117–1120