Investigating Session Search Behavior with Knowledge Graphs

Xiangsheng Li[1] Maarten de Rijke[2,3] Yiqun Liu[1*] Jiaxin Mao[4]
Weizhi Ma[1] Min Zhang[1] Shaoping Ma[1]
[1]Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing, China
[2]University of Amsterdam, Amsterdam, The Netherlands [3]Ahold Delhaize, Zaandam, The Netherlands
[4]Renmin University of China, Beijing 100872, China
yiqunliu@tsinghua.edu.cn,m.derijke@uva.nl

ABSTRACT

Knowledge graphs are widely used in information retrieval as they can enhance our semantic understanding of queries and documents. The main idea is to consider entities and entity relationships as side information. Although existing work has achieved improvements in retrieval effectiveness by incorporating information from knowledge graphs into retrieval models, few studies have leveraged knowledge graphs in understanding users' search behavior. We investigate user behavior during session search from the perspective of a knowledge graph. We conduct a query log-based analysis of users' query reformulation and document clicking behavior. Based on a large-scale commercial query log and a knowledge graph, we find new user behavior patterns in terms of query reformulation and document clicking. Our study deepens our understanding of user behavior in session search and provides implications to help improve retrieval models with knowledge graphs.

KEYWORDS

Session search; Knowledge graph; Query log analysis

ACM Reference Format:
Xiangsheng Li, Maarten de Rijke, Yiqun Liu, Jiaxin Mao, Weizhi Ma, Min Zhang, and Shaoping Ma. 2021. Investigating Session Search Behavior with Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463107

1 INTRODUCTION

As search systems and their functionality continue to be improved, the complexity and scope of users' information needs increase [1]. Users often have multiple types of interactions with search systems, including issuing queries, examining and clicking search results, to fulfill their information needs. To better understand users' information needs, query logs have been extensively studied to mine users' search intent [17]. Typically, query logs are partitioned into *search sessions*, i.e., sequences of queries and clicks issued by the same user within a short time interval. Previous studies [4, 20] have identified search behavior patterns by analyzing query reformulation and search result click preferences in search sessions. These findings deepen our understanding of users' search behavior and provide implications for improving the design of retrieval systems.

A *knowledge graph* (KG) is a repository of entities and their relationships and attributes that are represented as a graph; KGs are used extensively to better estimate relevance of a document for a query [16]. Existing KG based methods can be divided into implicit and explicit entity relationship exploitation methods. The implicit methods directly model the interactions between entities in the query and document by pretrained network embeddings. For example, Xiong et al. [21, 22] consider a bag-of-entities representation, and use an interaction matrix between bag of words representations and bag of entity representations for document ranking. Lu et al. [14] measure users' entity preferences by using the entity vector similarity from TransE [2]. Explicit methods explicitly exploit the relationship or the neighborhood information in the modeling. Liu et al. [13] utilize the entity description in a query to enrich the query representation. Ma et al. [15] use KGs to build explicit explainable patterns and rules for recommendation.

So far, few publications have utilized KGs to investigate users' search behavior [18]. Using a KG, we can investigate the relationship between the entities of two consecutive queries in the same session and identify patterns in sequences of query reformulations by looking at connected paths of mentioned entities in a KG. To better understand users' search behavior using KGs, we conduct a query log analysis of users' query reformulation and document clicking behavior by using a large-scale KG. We extract entities in the queries and documents from the same session and then analyze how they are connected in the KG. Specifically, we study the following research questions using a KG: (RQ1) How do users formulate queries in a session? (RQ2) How are entities in the previous query related to entities in the reformulated query? (RQ3) How are entities in the issued query and clicked documents related?

To address these research questions, we use a large-scale, publicly available query log, Tiangong-ST [5], and an English-Chinese bilingual KG, XLore [19], as our experimental data. We first study the commonness of query entities in a session and summarize the patterns of how users formulate queries. Then, the relationship between entities in query reformulations, and the differences in actions and other dimensions (such as reformulation type, session

*Corresponding author.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463107

Table 1: Statistics of our dataset.

#queries	#sessions	#avg. query per session	#avg. click per query	#avg. doc per query
231,142	89,737	2.58	1.26	9.58

Table 2: Number of entities in queries and documents.

	0	1	2	3	4	≥ 5
query	5.85%	29.4%	24.1%	20.7%	11.4%	8.50%
doc	0.96%	4.41%	11.9%	10.6%	17.5%	54.6%

length) are reported. Finally, we compare the differences between clicked documents and other documents according to their relationships with the query entities. Our findings provide a deeper understanding of user behavior in session search and reveal a rich opportunity to build better retrieval models with KGs.

2 DATASET

We conduct experiments on a large-scale, publicly available query log from a Chinese commercial search engine, Sogou.com, namely Tiangong-ST[1] [5]. Tiangong-ST provides web search session data extracted from an 18-day search log. We use the training set for our analysis and filter out sessions according to the following rules: (1) sessions without any clicks; (2) all queries in a session are the same; (3) the number of entities in all queries in a session is zero. Table 1 shows the statistics of the resulting dataset. To avoid working with very large numbers of entities in documents, we use document titles instead of the full document content.

For entity annotation, we utilize XLore [19]. XLore is an English-Chinese bilingual KG built from English and Chinese Wikipedia, Baidu Baike and Hudong Baike. It contains 16,284,901 entities, 2,466,956 concepts and 446,236 relations. The relations have four types: *subclass*, *instanceof* and *same*, *related*, where the proportions are 2.6%, 42.1%, 7.9%, 47.3%, respectively. The relations *subclass* and *instanceof* are further categorized as Derivation type while *same* and *related* are considered to be Equivalency type. Following [13], query and document entities are annotated by CMNS [9], the commonness (popularity) based entity linker. Table 2 shows the distribution of the number of entities in queries and documents; the average number of entities in queries and documents is 2.28 and 4.03, respectively.

3 RESULTS

3.1 Analysis of the issued queries

To answer **RQ1**, we report the word and entity statistics in the sessions and reveal the query patterns during session search. The results are shown in Table 3. Specifically, *Number* indicates the change in the number of words (or entities) between two consecutive queries in a session. *Overlap with first query* and *Overlap previous query* indicates the overlap rate between each query and its first (previous) query in the same session. *Common elements in session* denotes the ratio of common words (or entities) in a session.

We first observe that the trends of the number of words and entities are similar; users tend to use more words (entities) to express

[1] http://www.thuir.cn/tiangong-st/.

Table 3: Word and entity statistics of the queries. A one-sample Wilcoxon signed rank test is used to test significant differences in the number of words/elements between consecutive queries in a session, * means $p \leq 0.01$, two-tailed.

	Word	Entity
Number	1.266*	0.280*
Overlap with first query	0.319	0.283
Overlap previous query	0.355	0.316
Common elements in session	0.290	0.256

their information need as the session goes on. This indicates that users gradually arrive at a clearer way of expressing their information needs thus submit longer queries. Overlap with the first and the previous query and the common elements rates in the whole session are all around 30%, which suggests that there exists a key concept in a session and users will modify their query based on this key concept during the session search.

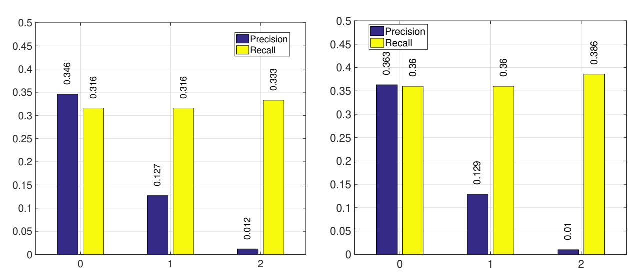
In conclusion, during session search users will use a key concept and attempt to modify this key with additional words.

3.2 Analysis on query reformulation

To answer **RQ2**, we study query reformulations in different scenarios by analyzing the corresponding entity relationships in KG. We first analyze the length of the shortest path between entities in the current query and the context. Then, we study the difference in entity relationships in different types of query reformulation and different session lengths.

3.2.1 Entity connections. To study how entities in the current query are connected to entities in the context, we take entities in the previous query (or all contextual queries) as the starting points and the entities in the current queries as the target points. We study how they are connected in the KGs. If a target point and a starting point are connected within a given hop, they are a *positive entity pair*. Since the size of the KG is large, we only consider entity pairs within 2 hops from each other. We define *precision* and *recall* as the ratio of all positive pairs in all 2-hop neighborhoods and in all start-target pairs, respectively. Assume, for example, that two consecutive queries contain entity A and entity B, respectively. A and B are within 2-hop in the graph. A has 5 2-hop neighbors in the graph (one of them is B). In this case, the precision is 1/5, recall is 1/1.

The results are reported in Figure 1.



(a) From the previous query (b) From contextual queries
Figure 1: Precision and recall of the positive entity pairs from different sources.

Table 4: Entity relationships according to different reformulation types. A one-sample Wilcoxon signed rank test is performed to test significant difference, * means $p \leq 0.01$, two-tailed. The relation ratio of Equivalency is one minus Derivation.

Reformulation type	Definition	Proportion	Δ entity	Δnew entity	Avg. hops	Relation ratio (Derivation)
Generalization	$Q_i \cap Q_{i+1} \neq \emptyset$; $\lvert Q_i\rvert > \lvert Q_{i+1}\rvert$	11.49%	-1.09*	0.157*	0.984	92.9%
Specialization	$Q_i \cap Q_{i+1} \neq \emptyset$; $\lvert Q_i\rvert < \lvert Q_{i+1}\rvert$	37.35%	1.07*	0.325*	0.910	92.9%
Word substitution	$0 < \lvert Q_i \cap Q_{i+1}\rvert < \lvert Q_i\rvert$; $\lvert Q_i\rvert = \lvert Q_{i+1}\rvert$	12.69%	-5.44e-2	0.270*	1.027	93.4%
Repeat[2]	$Q_i = Q_{i+1}$;	12.52%	-2.82e-4	2.54e-4	0.827	92.6%
New	$Q_i \cap Q_{i+1} = \emptyset$;	25.96%	2.42*	2.42*	1.921	92.5%

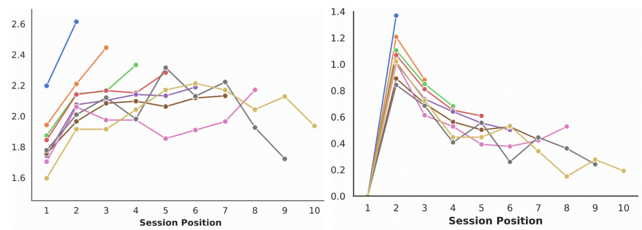
In Figure 1, 0-hop refers to the repeated entities between two queries. We observe that when considering the entity pairs between the current query and all contextual queries rather than the pairs between the current query and the previous query, the recall increases from 0.316 to 0.360, which means that using more context information is helpful to infer the next query. However, the overlap rate is only about 30%, which means that the majority of entities are not from the context.

When considering entity pairs with 1-hop, we find that the difference with the result in 0-hop is not significant. Statistically significant differences do occur when we consider 2-hop entities. Hence, by only using 1-hop neighborhoods we cannot effectively infer the next query; at least 2-hop neighborhoods are required.

In addition, although recall increases, precision drops drastically when we use broader neighborhoods. This means the noisy entities are also incorporated and we need a reasonable strategy to filter out these noisy entities.

3.2.2 Query reformulation types. Next, we consider query reformulations. We use the reformulation types described in [11], i.e., *Generalization*, *Specialization*, *Word Substitution*, *Repeat* and *New*. For details on the categorization we refer to [11]. We analyze the relationship between the entity pairs in two consecutive queries according to different reformulation types, as shown in Table 4. The distribution of reformulation types is similar to [11], where *Specialization* is the majority type. For the change in the number of entities (i.e., Δentity), only *Generalization*, *Specialization* and *New* use significantly different numbers of entities compared with the previous query while *Word Substitution* and *Repeat* do not. For new entities (i.e., Δnew entity), we find that all reformulation types except *Repeat* use new entities. *Avg. hops* measures the distance between entities in two consecutive queries in the KG. In addition, although *Specialization* introduces the largest number of new entities, the connections between new entities and the entities in the previous query are still close (with a low avg. hops). For *Word Substitution*, the hop distance between the entities in the current query and previous query is larger. In our KG, we have two main relation types: Derivation (*subclass* and *instanceof*) and Equivalency (*same* and *related*); we find that most relations are Derivation, which suggests that users are less likely to use entities with similar meanings.

3.2.3 Session length. We split the search sessions according to their length. The number of (new) entities at different positions in a session is shown in Figure 2. Shorter sessions tend to have



(a) (b)

Figure 2: Average number of entities (a) and new entities (b) across session iterations. The number of new entities is based on the previous query in the same session. Session position indicates the query order in a session.

more entities at each session position, which indicates that users in short sessions are clearer about their search intent. Generally, the number of entities increases as the session evolves because users gradually gain insight in their search intent, which is similar to the discussion in Table 3. As to the number of new entities, the biggest growth happens in the first reformulation, which is probably because users do not know how to formulate their initial query. Once they browse the result pages, they are clearer about how to reformulate the query.

3.3 Analysis of document clicking

To answer **RQ3**, we study the relationships between entities in the query and the returned documents (i.e., the common entities and how they are connected in the KGs). First, we contrast clicked and non-clicked documents. Then we analyze the distance between entities in the current query and the documents. Finally, we discuss the difference in different search intents.

3.3.1 Click behavior. We split query-document pairs according to whether the document is clicked. The results are shown in Table 5, where $E(Q)$ and $E(D)$ are the sets of entities in the query Q and the document D, respectively. Also, $E(Q) \cap E(D) \neq \emptyset$ and $\frac{E(Q)\cap E(D)}{E(Q)}$ refer to the proportion of query-document pairs with at least one common entity and the percentage of query entities appearing in the document.

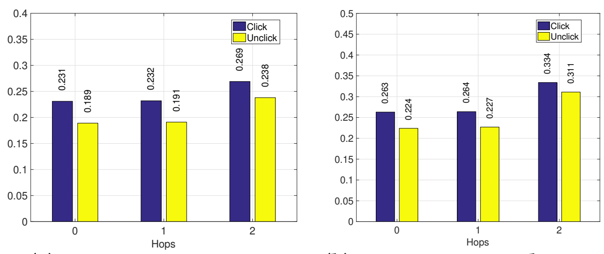
Table 5: Entity relationships in the clicked and unclicked documents. * denotes a significant difference under Wilcoxon signed rank test ($p \leq 0.01$).

	$E(Q) \cap E(D) \neq \emptyset$	$\frac{E(Q)\cap E(D)}{E(Q)}$	Avg. hops	Relation ratio (Derivation)
Click	70.36%	0.334	0.641	94.0%
Unclick	60.91%*	0.294*	0.694*	93.6%

[2] Repeat indicates that Q_i and Q_{i+1} have the same terms but the formats may be different, e.g., upper and lower case.

The differences in $E(Q) \cap E(D) \neq \emptyset$, $\frac{E(Q) \cap E(D)}{E(Q)}$, and avg. hops between clicked and non-clicked documents are all significant.

Clicked documents contain more entities from the query (exact matching) and their entities are closer to entities in the graphs (semantic matching). This indicates that we can use both exact and semantic matching signals to predict users' click behavior. In addition, the proportion of query-document pairs with at least one common entity is only 70.36%, which means that about 30% of the relevant documents have entities that are not in the query. These hard to retrieve documents require retrieval strategies beyond exact matching signals to infer user's search intent.



(a) From current query **(b) From contextual queries**

Figure 3: Recall of hit entities in clicked and non-clicked documents when spreading out entities from different sources.

3.3.2 Entity connection. Similar to the discussion in Section 3.2.1, we study how query entities are connected to document entities. We construct entity pairs by connecting query entities and document entities. Entity recall is defined in Section 3.2.1, i.e., the ratio of the entity pairs within a given hop in the KG among all entity pairs. See Figure 3. Our observations are similar to those for query reformulation: (1) not only the current query but also the contextual queries are helpful to infer the entities in the clicked documents; and (2) the number of hit entities does not increase significantly when using 1-hop neighborhoods but becomes significant when using 2-hop neighborhoods. Hence, 1-hop neighborhoods do not suffice to effectively infer the entities in clicked documents. Instead, we should use at least 2-hop neighborhoods.

3.3.3 Query intent. We randomly select 200 sessions and annotate the query intents according to the definitions in [3]. The labels are determined by the majority votes of three experts. The results are shown in Table 6. For each intent we consider $E(Q) \cap E(D) \neq \emptyset$, $\frac{E(Q) \cap E(D)}{E(Q)}$, and the avg. hops between query and clicked documents, as defined in Section 3.3.1. The proportion of query-document pairs with at least one common entity, $E(Q) \cap E(D) \neq \emptyset$, and the percentage of query entities appearing in the document, $\frac{E(Q) \cap E(D)}{E(Q)}$, are both largest for navigational query and smallest for transactional queries. Hence, exact matching signals based on entities are more important for navigational queries, and transactional queries require other matching signals for click prediction. For the average number of hops between entities in queries and clicked documents, informational queries have the biggest distance, which suggests that exact matching signals may not perform well for informational queries. The majority of derivation relations in informational queries indicates that users with an informational intent tend to search targets beyond the issued query.

Table 6: Entity relationships between query and clicked documents in different query intents.

	$E(Q) \cap E(D) \neq \emptyset$	$\frac{E(Q) \cap E(D)}{E(Q)}$	Avg. hops	Relation ratio (Derivation)
Navigation	81.3%	0.639	0.350	0.905
Information	74.0%	0.535	0.796	0.973
Transaction	70.8%	0.491	0.540	0.946

4 CONCLUSION

Knowledge graphs (KGs) have been used extensively in retrieval models to better estimate the relevance of a document to a query. Few studies have utilized KGs to investigate users' search behavior. To better understand user behavior, we conduct a query log analysis with a large-scale KG. We first analyze the distribution of words and entities in the query and find that queries in a session are based on a key concept and users tend to use longer queries as the session evolves. In query reformulation, using 1-hop neighborhoods is not effective to infer the next query and at least 2-hop neighborhoods are required. We also discuss differences between reformulation types and session lengths. For click behavior, the relationship of query and clicked documents is not always based on exact matching, which motivates the design of retrieval models that go beyond exact matching signals. Comparing three query intents, we find that exact matching signals on entities are more important for navigational queries while transactional queries require other matching signals for click prediction.

In future work, we plan to use our findings to help design and improve knowledge-aware retrieval models, for example, by incorporating different types of KG-based information for queries with different search intents. By deepening our understanding of users' search behavior using knowledge graphs, we expect to be able to design more effective retrieval models.

REPRODUCIBILITY

The code and data used to produce the results in this paper are available at [https://github.com/lixsh6/KnowledgeAnlysis-SIGIR2021-SP].

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61902209, 62002191, 61532011), Beijing Academy of Artificial Intelligence (BAAI) and Tsinghua University Guoqiang Research Institute. This project is also funded by China Postdoctoral Science Foundation (2020M670339) and Dr Weizhi Ma has been supported by the Shuimu Tsinghua Scholar Program. This research was (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Eugene Agichtein, Ryen W White, Susan T Dumais, and Paul N Bennet. 2012. Search, Interrupted: Understanding and Predicting Search Task Continuation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 315–324.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in neural information processing systems*. 2787–2795.

[3] Andrei Broder. 2002. A Taxonomy of Web Search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.

[4] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Query Reformulation Behavior of Search Users. In *China Conference on Information Retrieval*. Springer, 39–51.

[5] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. TianGong-ST: A New Dataset with Large-scale Refined Real-world Web Search Sessions. In *Proceedings of the 28th ACM International on Conference on Information and Knowledge Management*. ACM, 2485–2488.

[6] Michael J Cole, Chathra Hendahewa, Nicholas J Belkin, and Chirag Shah. 2015. User Activity Patterns during Information Search. *ACM Transactions on Information Systems (TOIS)* 33, 1 (2015), 1–39.

[7] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive Graph for Multi-hop Reading Comprehension at Scale. *arXiv preprint arXiv:1905.05460* (2019).

[8] Zhengxiao Du, Chang Zhou, Ming Ding, Hongxia Yang, and Jie Tang. 2019. Cognitive Knowledge Graph Reasoning for One-shot Relational Learning. *arXiv preprint arXiv:1906.05489* (2019).

[9] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2017. Entity Linking in Queries: Efficiency vs. Effectiveness. In *Proceedings of the 39th European Conference on Information Retrieval*. Springer, 40–53.

[10] Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach Machine How to Read: Reading Behavior Inspired Relevance Estimation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 795–804.

[11] Chang Liu, Jacek Gwizdka, Jingjing Liu, Tao Xu, and Nicholas J Belkin. 2010. Analysis and Evaluation of Query Reformulations in Different Task Types. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–9.

[12] Jiqun Liu, Shawon Sarkar, and Chirag Shah. 2020. Identifying and Predicting the States of Complex Search Tasks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 193–202.

[13] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Entity-duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval. *arXiv preprint arXiv:1805.07591* (2018).

[14] Shuqi Lu, Zhicheng Dou, Chenyan Xiong, Xiaojie Wang, and Ji-Rong Wen. 2020. Knowledge Enhanced Personalized Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 709–718.

[15] Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. 2019. Jointly Learning Explainable Rules for Recommendation with Knowledge Graph. In *The World Wide Web Conference*. 1210–1221.

[16] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2020. Knowledge Graphs: An Information Retrieval Perspective. *Foundations and Trends in Information Retrieval* 14, 4 (2020), 289–444.

[17] Fabrizio Silvestri. 2009. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval* 4, 1–2 (2009), 1–174.

[18] Svitlana Vakulenko, Maarten de Rijke, Michael Cochez, Vadim Savenkov, and Axel Polleres. 2018. Measuring Semantic Coherence of a Conversation. In *ISWC 2018: 17th International Semantic Web Conference*. Springer, 634–651.

[19] Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. 2013. XLore: A Large-Scale English-Chinese Bilingual Knowledge Graph. In *Proceedings of the 12th International Semantic Web Conference*. 121–124.

[20] Dietmar Wolfram, Peiling Wang, and Jin Zhang. 2009. Identifying Web Search Session Patterns Using Cluster Analysis: A Comparison of Three Search Environments. *Journal of the American Society for Information Science and Technology* 60, 5 (2009), 896–910.

[21] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. 2016. Bag-of-entities Representation for Ranking. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. 181–184.

[22] Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. In *Proceedings of the 26th international conference on world wide web*. 1271–1279.