Query refinement: type aware approach

Zahra Taherikhonakdar
University of Windsor
taherik@uwindsor.ca

Aaditya Pradipbhai Parekh
University of Windsor
parekh23@uwindsor.ca

ABSTRACT

This paper provides a concise overview of the query reformulation, the history of query reformulation and state of the art methods in this domain.

KEYWORDS

Query Refinement, Query Types, Information retrieval

1 QUERY REFORMULATION

A web search does not yield relevant results. Because, first, the user's initial query consists of keywords that may be related to multiple topics. For example, the word 'python' refers to the language programming and the snake. Second, the query could be too short to apprehend what the user is looking for. Third, the users are often not sure about what they are looking for until they see the results because they just begin the exploration. Modern search engines use query reformulation to assist users in getting their intended search results to address their information needs. In query reformulation, the input query would be changed by deleting/changing/adding related terms. Query reformulation can be used in query suggestions to help users reformulate their queries to achieve desirable results. It can also be used by search engines to refine the input query without noticing the user and return related results. In this paper, we used query reformulation/suggestion/refinement interchangeably.

1.1 History

Query refinement was first used in 1960 by Maron, and Kuhns [16] as a new technique for literature indexing and searching in a mechanized library system. In 1971, Rocchio [19] developed a relevance feedback algorithm as a method for query reformulation and is used for automatic document processing and text categorization by information retrieval (IR) systems. In this method, the user gives feedback on the retrieved documents based on the initial query by indicating the relevance of the results. Then, the user's feedback is incorporated into the information retrieval process to improve the final result. Later, Rocchio's work were extended in new techniques like collection-based term co-occurrence and cluster-based information retrieval [10, 11] and automatic text processing [20]. These works were developed before the advent of search engines in the 1990s and social media in the early 2000s. In this era, huge amounts of data started being published on the web and social media, which has continued to grow at an accelerating rate until today. Although there are a huge amount of data, users tend to use short queries for searching on the web or social networks. This increases the queries ambiguity and the necessity of query reformulation. Nowadays, query reformulation techniques are used in question answering systems, web, social media, desktop, and email searches [17].

2 PRIOR WORK

Several query suggestion models have been introduced to overcome the major challenge in query suggestion which is understanding the real user's search intent for predicting the next query. In this context, several studies used the "wisdom of crowds" by analyzing the search log of a user (frequency and popularity of last queries). As an example, Huang et al. [9] considered the query co-occurrences to suggest the next query. Although, query suggestion models based on co-occurrence have some issues like data sparsity and the inability to cover rare or unseen data. This approach can not be used in a specific area where there is a wide range of possible queries and systems are rarely used, like academic search [13]. To reduce initial query ambiguity using query reformulation, researchers try to take advantage of different information, like user's search log information, user's social networks information, and query types. In this paper, we consider different query types and evaluate the different query refinement techniques based on them to select the appropriate refining methods. This would result in information retrieval(IR) methods improvement and the user's search satisfaction. In this section, we investigate recent works in query reformulation to show that considering related data (e.g., search log information, user's social networks information) in initial query would improve IR performance as it helps search engines to better understand the users' original query.

2.1 search log in query reformulation

Many works have been conducted to take advantage of users' search log to understand users' search intent. Chen et al. [4] proposed the Personalized Query Suggestion Diversification (PQSD) model, using the AOL data set [18]. This model combines personalization and diversification. The proposed model ensures that suggested queries are close to a user's specific search intent. Ahmad et al. [1] identified users' search context to improve query suggestion and document ranking. To identify the search context, they consider users' previous queries and document clicks (users feedback) in the search log. To do this, they consider previous queries related to a search task in the search log and consider it as a search session. Using AOL data set, Fiorini et al. [7] proposed new approaches in neural Query Auto-Completion (QAC) to address the limitations of previous methods. This paper integrated personalization and time-sensitive information in neural QAC to address the inability of the previous methods to predict unseen queries. Dehghani et al. [6] used the sequence-to-sequence(seq2seq) models and proposed a new architecture to generate query suggestions. The author detects two main problems in existing models. First, directly employing the seq2seq model would lead to losing information at a query level because this model considers a query like a sequence of words. These models are unable to handle out-of-vocabulary words (OOV: terms that are not part of the normal lexicon found in a natural language processing environment). Moreover, these models are less

likely to generate very low-frequency words. The proposed model overcome previous models short-comes.

2.2 social information in query reformulation

With the advent of social networks, users could communicate with each other and share data in form of text, photos, videos on social networks. Moreover, users could share their opinion by retweets, likes, comments. This sharing of information and opinions results in producing a huge amount of data. Researches take advantage of this information in query suggestions to help users to address their information needs [12, 14]. Flora Amato et al. [2] proposed a state-of-the-art method for big data applications to make multimedia suggestions (suggests: photo, video, text) based on different aspects of one or more social media networks of users. They used a subset of the Yahoo Flickr [1] Creative Commons 100 Million Data (YFCC100M) multimedia collection (containing about 500,000 images), provided by Yahoo in 2014. The results show that the proposed SOS model by Flora Amato et al. [2] outperforms the previous algorithms when the rate of sparsity matrix increases. Dionisis et al. [15] proposed techniques for query suggestion on user's social information on social networks. This method considers both a user's items choice like browsing and rating and the information from who have an influence on the user on social networks. The results show that in high values, query suggestions can be improved by social information.

2.3 Query types in query suggestion

Considering query types would help search engines to better understand the user's search intent. with the am of understanding the need behind the query, Border [3] divided web search queries into three categories, namely: navigational, informational, and transactional queries. He claimed that search engines need to handle all three types ,and an understanding of these categories is essential to the successful web search development. Santos et al. [21] considering informational and navigational query types using the TREC 2009 data set [5] to infer the intent of the query. They introduced an intent-aware approach and focused on selecting suitable retrieval models for different query types to improve the diversification in results. Their approach significantly improves the diversification in web search results. Researchers try to define different types of queries in different domains to increase IR performance and also increase diversity in retrieving results. Vallet et al. [22] proposed the method to predict important types based on informational queries. They introduce the top three query types in the informational query category, namely: country, organization, events. They claim that this would increase the performance of the informational query search. Grbovic et al. [8] classified users' queries into several predefined interest categories, such as "finance", "sport", or technology. This allows search engines to better understand user intentions and provide relevant results. They trained the model using more than 12 billion search sessions collected on Yahoo servers. The proposed system achieves %86 precision with high recall.

In this research proposal, we consider query types in query refinement to improve IR performance. We use the TREC 2009 data set [5] and use different query refinement methods for all three query types in the data set. Using metrics (e.g. NDCG) we would evaluate the IR performance by our approach. Based on the evaluation, the appropriate query refinement method is considered for each query type. We expected using an appropriate query refinement method for each query type improves the IR performance.

[1]https://webscope.sandbox.yahoo.com

REFERENCES

[1] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 385–394.

[2] Flora Amato, Vincenzo Moscato, Antonio Picariello, and Francesco Piccialli. 2019. SOS: A multimedia recommender System for Online Social networks. *Future generation computer systems* 93 (2019), 914–923.

[3] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.

[4] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2017. Personalized query suggestion diversification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 817–820.

[5] Charles L Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the trec 2009 web track*. Technical Report. WATERLOO UNIV (ONTARIO).

[6] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session-based query suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1747–1756.

[7] Nicolas Fiorini and Zhiyong Lu. 2018. Personalized neural language models for real-world query auto completion. *arXiv preprint arXiv:1804.06439* (2018).

[8] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Narayan Bhamidipati, Jordan Hawker, and Caleb Johnson. 2015. querycategorizr: A large-scale semi-supervised system for categorization of web search queries. In *Proceedings of the 24th International Conference on World Wide Web*. 199–202.

[9] Chien-Kang Huang, Lee-Feng Chien, and Yen-Jen Oyang. 2003. Relevant term suggestion in interactive web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology* 54, 7 (2003), 638–649.

[10] Nick Jardine and Cornelis Joost van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information storage and retrieval* 7, 5 (1971), 217–240.

[11] Karen Sparck Jones. 1971. Automatic keyword classification for information retrieval. (1971).

[12] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M Jose. 2009. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 195–202.

[13] Matti Lankinen, Hannes Heikinheimo, Pyry Takala, Tapani Raiko, and Juha Karhunen. 2016. A character-word compositional neural language model for finnish. *arXiv preprint arXiv:1612.03266* (2016).

[14] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. 2011. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*. 287–296.

[15] Dionisis Margaris, Costas Vassilakis, and Panagiotis Georgiadis. 2018. Query personalization using social network information and collaborative filtering techniques. *Future Generation Computer Systems* 78 (2018), 440–450.

[16] Melvin Earl Maron and John Larry Kuhns. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)* 7, 3 (1960), 216–244.

[17] Dipasree Pal, Mandar Mitra, and Samar Bhattacharya. 2015. Exploring query categorisation for query expansion: A study. *arXiv preprint arXiv:1509.05567* (2015).

[18] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*. 1–es.

[19] Joseph Rocchio. 1971. Relevance feedback in information retrieval. *The Smart retrieval system-experiments in automatic document processing* (1971), 313–323.

[20] Gerard Salton. 1989. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley* 169 (1989).

[21] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2011. Intent-aware search result diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 595–604.

[22] David Vallet and Hugo Zaragoza. 2008. Inferring the most important types of a query: a semantic approach. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*. 857–858.