Neurocomputing 000 (2017) 1–5

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Query completion in community-based Question Answering search

Xian-Ling Mao, Yi-Jing Hao, Dan Wang, Heyan Huang*

Department of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

ARTICLE INFO

Article history:
Received 2 March 2016
Revised 3 June 2016
Accepted 27 June 2016
Available online xxx

Keywords:
Query completion
Auto-completion
cQA search

ABSTRACT

Query completion has long been proved useful to help a user explore and express his information need. In general search, such completions can be generated from a large scale query log and other accessory information. However, without query log, how to generate query completion for community-based Question Answering (cQA) search remains a challenging problem.

In this work, we propose a novel query completion algorithm based on ranking cQA questions with entity and phrase information for cQA search, and a demonstration system has been developed. Without involvement of query log, this method clearly helps users complete their queries. Empirical experiments on a large scale cQA dataset show that the proposed algorithm can successfully improve user experience.

1. Introduction

Query completion aims to help users complete their queries. Different techniques have been exploited to mine web search query logs [1–3] and page contents [4,5] for query completion. Major search engines have also launched the suggestion service that recommends relevant phrases for query completion. For community-based Question Answering (cQA) search, the query completion remains untouch.

There is a gap between the queries submitted by users and the requirement of cQA ranking function. On the one hand, when users search the cQA collection, usually they want to obtain the similar questions that can satisfy their information need with high probability, which requires question-type queries (i.e. a query is a question). On the other hand, the retrieval models for cQA are tended to assume that queries are questions [6,7], which is essentially to rank cQA documents by comparing the similarity between the query and questions [8–10]. However, users are tended to submit short queries, not questions [11]. We did a survey in the AOL query log, and our observations show that the proportion of questions in queries is not more than 2%, and that the average number of terms in a query is 2.14. At the same time, the average number of words in a question in our crawled data from Yahoo! Answer is 10.61. The large gap between the two above-mentioned shows the fact that a query is usually a part of a complete question, and it is why we should pay attention to the query completion.

Furthermore, a question can provide more information than its corresponding short query for retrieval models. For example, a query: *[the latest version iphone]*, and corresponding question is *[what is the latest version of iphone?]*, the result from Yahoo! Answer Search are presented in Fig. 1, which shows if a query is a question (not only keywords), cQA system can return results better. Thus, how to help users to submit question-type queries which are convenient for cQA retrieval models is a challenging problem. Here, query completion can be used to bridge the gap, however, we have no query log. Many web search queries do not look for web pages, but instead are seeking information about named entities like products, people, locations and organizations [12,13]. By analyzing the 36 million queries of the AOL data set, it was found that 20–40% of web search queries consist solely of single entities, while 70–87.5% queries contain entities as parts of them. These findings demonstrate that a very high percentage of all web queries recognizably target entities or have entities associated with them [14,15]. Thus, how the entities in search queries should be preprocessed and used is very important for query completion.

We make the following contributions in this paper:

- We propose a novel query completion algorithm based on ranking cQA questions with entity and phrase information for cQA search, and a demonstration system has been developed.
- Experimental results indicate that our method can effectively generate question-type query suggestions.

In this paper, we first identify the requirements of query completion in cQA search, and then propose a new approach to satisfy these requirements by considering the local word order of phrases and entities in queries.

* Corresponding author.
E-mail addresses: maoxl@bit.edu.cn (X.-L. Mao), 2220150504@bit.edu.cn (Y.-J. Hao), wangdan12856@bit.edu.cn (D. Wang), hhy63@bit.edu.cn (H. Huang).

http://dx.doi.org/10.1016/j.neucom.2016.06.096
0925-2312/

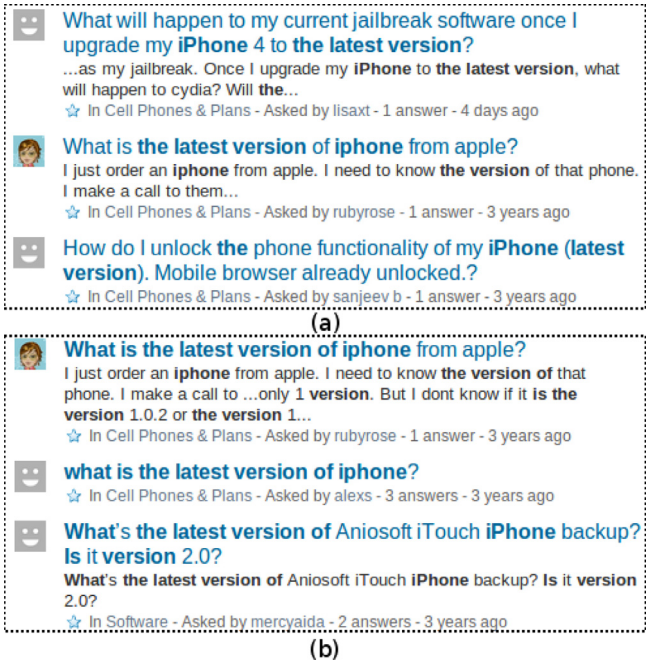

Fig. 1. Results of Yahoo! Answer Search: (a) query is "the latest version iphone"; (b) Query is "what is the latest version of iphone?".

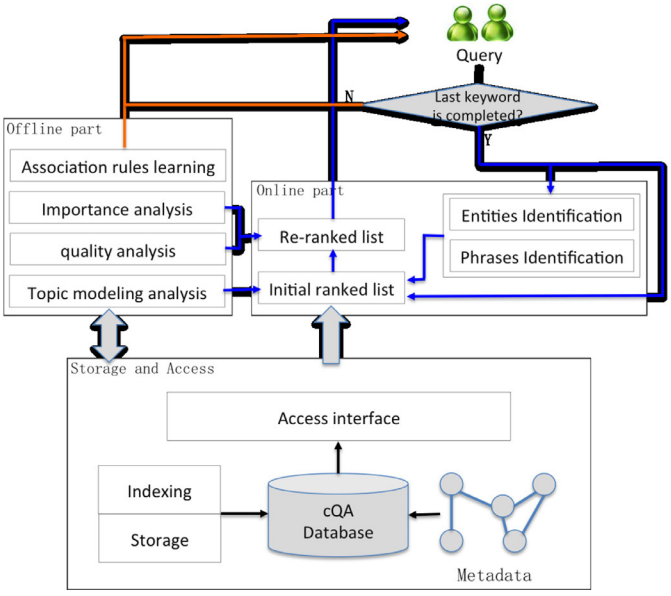

Fig. 3. Architecture of query completion in cQA search.

2. The requirements of query completion in cQA search

Given an uncompleted query (UQ) and a cQA archive, a ranked list of questions that are candidates to complete the UQ is returned during the input stage of cQA search, called *query completion in cQA search*. Besides the short response time, below we first identify the special requirements of query completion in cQA search.

- *The divergence.* As known, there are many similar questions in cQA. For example, query is "iphone 3g iphone 3gs", the top 6 suggestions are shown in Fig. 2(a), all six completed questions are similar questions, which leads to bad user experience. Thus, the divergence of completed queries is an important factor for promoting user experience.
- *The size-moderation.* As shown in Fig. 2(a), the last completed question is so long that its length excesses that of suggestion frame, and users have to take too much time to scan it, which downgrades user experience. Considering the position of query completion function in a search interface, it is better that the length of completed questions does not excess that of search box. For example, the completions in Fig. 2(b) appear better.
- *The local order-preserving.* The word order in query is very important, especially, entity order and phrase order. Query completion methods in major search engines are usually based on query log, which constrain that the suggested queries must contain the uncompleted query sequentially. For example, the results from Google and Yahoo are showed in Fig. 4(a) and (b). For query completion in cQA search, because the completed queries are questions, it is impossible to constrain that the completed questions must contain the user's query sequentially. In our implementation, it is reasonable that only keep the local word order of entity and phrase in a completed query.
- *The fidelity.* Any modification for a query maybe lead to obtaining different meaning query, which is disastrous. Thus, query completion should try to keep the fidelity of the original query.

3. Technical specification

Fig. 3 shows the architecture of our approach to the query completion problem in cQA search. In general, it consists of three components: "storage and access", "offline part" and "online part".

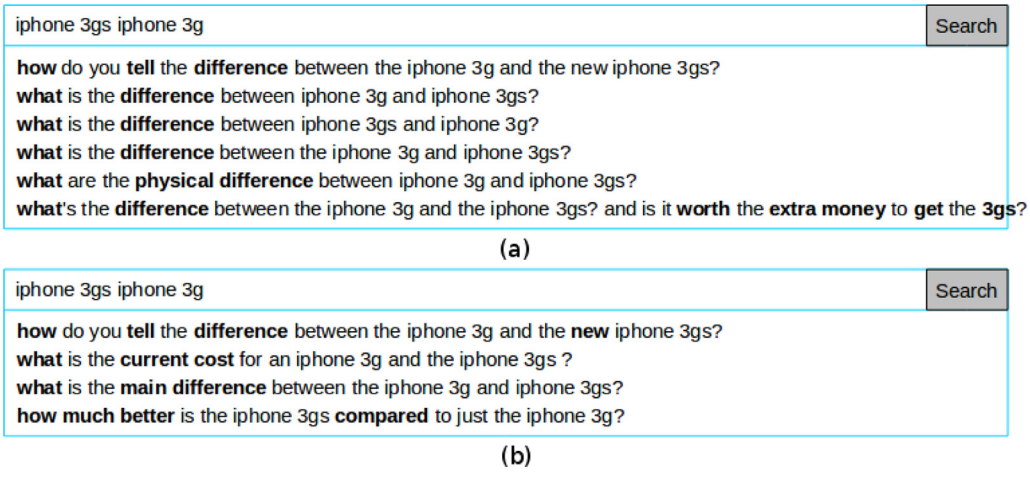

Fig. 2. Query completion examples. (a) Query completions of low quality; (b) Query completions of high quality.

Please cite this article as: X.-L. Mao et al., Query completion in community-based Question Answering search, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.06.096