#### Knowledge-Based Systems 89 (2015) 553-568

Contents lists available at ScienceDirect

### **Knowledge-Based Systems**

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

# Kriniedije-Dee Svij 1245

## Query suggestion with diversification and personalization

Di Jiang<sup>a</sup>, Kenneth Wai-Ting Leung<sup>b</sup>, Lingxiao Yang<sup>c,\*</sup>, Wilfred Ng<sup>b</sup>

<sup>a</sup> Baidu.Inc, China

<sup>b</sup> Hong Kong University of Science and Technology, Hong Kong

<sup>c</sup> London School of Economics and Political Science, United Kingdom

#### ARTICLE INFO

Article history: Received 12 November 2014 Received in revised form 19 July 2015 Accepted 6 September 2015 Available online 15 September 2015

Keywords: Personalization Diversification Search engine

#### ABSTRACT

Web search query suggestion is an important functionality that facilitates information seeking of search engine users. In existing work, the concepts of diversification and personalization have been individually introduced to query suggestion systems. In this paper, we propose a new query suggestion paradigm, *Query Suggestion With Diversification and Personalization* (QS-DP) to effectively integrate diversification and personalization into one unified framework. In the QS-DP, the suggested queries are effectively diversified to cover different facets of the input query while the ranking of the suggested queries are personalized to ensure that the top ones are those that align with a user's personal preferences. We evaluate QS-DP on a commercial search engine query log against several existing query suggestion methods. The experimental results verify our hypothesis that diversification and personalization can be effectively integrated and they are able to enhance each other within the QS-DP framework, which significantly outperforms several strong baselines with respect to a series of metrics.

© 2015 Elsevier B.V. All rights reserved.

#### 1. Introduction

Web search query suggestion is an important functionality that helps users to formulate more effective search queries. Most of the existing query suggestion methods [1–5] belong to the category of relevance-oriented query suggestion, which focuses on maximizing the relevance of the suggested queries in response to an input query. However, since search queries are typically short and ambiguous [6], the relevance-oriented methods usually fail in the face of query uncertainty, which widely exists in the scenario of general web search.

Consider the following example of query uncertainty. When a search query "sun" is submitted to the search engine, the underlying information need can be related to at least one of the three facets: the star of the solar system, the computer manufacturer named Sun Microsystems or a newspaper in the United Kingdom. The relevance-oriented approaches usually generate suggestions that cover a few or even one facet, such as only suggesting queries about Sun Microsystems. To alleviate the problem of query uncertainty, the recent strands of query suggestion research can be broadly separated into two categories: introducing either *diversification* or *personalization* to the conventional query suggestion. Some researchers have introduced diversification to query

\* Corresponding author. *E-mail address:* audreyyoung@126.com (L. Yang). suggestion systems [6,7]. The logic of diversification is to cover as many facets of the input query as possible with a single query suggestion list. More specifically, diversification aims to minimize the number of the totally unsatisfied users, trading degrees of satisfaction in exchange for increasing the size of the satisfied population. The downside of diversification is that, for a specific user, the query suggestion lists may contain irrelevant suggestions and the irrelevant ones may even be ranked much higher than the relevant ones. For instance, when a user submits the query "sun" to search for some information about Sun Microsystems, presenting a suggestion list in which queries such as "solar energy" or "sun daily uk" are ranked higher than "oracle sun" or "sun solaris" is disappointing.

As an alternative approach to tackle query uncertainty, some researchers propose to apply personalization in order to identify the suggestion candidates that are most similar to the user's search history [8,9]. Personalization strives to get further knowledge through a user's search history, in order to reduce the uncertainty of the input query. Essentially, personalization narrows down the scope of the possible interpretations to those that only align with the a user's personal preference. However, web search is essentially dynamic and a user's preference changes over time. For example, when a computer scientist submits "sun" to search for information about a hot topic of solar energy, suggesting queries that are only focused on Sun Microsystems is rather unsatisfactory. Therefore, traditional personalization approaches risk over-personalization





and cannot effectively handle the dynamic change of a user's preference.

From the above discussion, we can see that both diversification and personalization have their own advantages and drawbacks. The result of diversification can be overly broad for a specific user but is flexible to handle the dynamic changes of a user's preference. In contrast, the result of personalization is too rigid to handle the changes of a user's preference but is effective for those who are looking for information relevant to his or her long-term preference. In order to achieve better query suggestion performance than only applying one of the two concepts, we propose a new query suggestion paradigm, Query Suggestion With Diversification and Personalization (QS-DP) to unify the two concepts which have been perceived to be incompatible so far. We hypothesize that a good query suggestion paradigm needs to integrally consider diversification and personalization, i.e., the suggested queries need to be diversified to cover different facets of the input query and the ranking of the suggested queries needs to be personalized to align with the user's preferences. Consider the case of "sun" again: we cannot aggressively deny the probability that a computer scientist may search for information about solar energy, thus, the suggestions should be diversified to cover different facets such as Sun Microsystems, the solar system, the solar energy, etc. Furthermore, to improve a specific user's searching experience, the ranking of the suggested queries should be further personalized to align with the user's preference, e.g., ranking the queries such as"sun oracle", "sun java" higher than "sun solar system" higher for a computer scientist, since we know that the queries about computer science have better chance to be relevant to the user's current information need. In this way, we can better capture the broadness of a user's information needs and facilitate the user's information seeking when the information need aligns with his or her long-term preference. While this seems to be a simple idea, we are not aware of any previous work that explores how to integrate both diversification and personalization to improve the performance of query suggestion.

The architecture of OS-DP is shown in Fig. 1 and it can be divided into three major components: the query log representation component, the diversification component and the personalization component. In the query log representation component, we propose a multi-bipartite-graph representation for the query log data and this representation significantly improves the richness of the information available for the downstream query suggestion. In the diversification component, we propose a two-phase method to obtain a list of diversified query suggestion candidates. In the personalization component, we employ a generative model named User Profiling Model (UPM) to integrate information such as search sessions, web dynamics and each individual user's preference of word usage and URL clicking. A technique of identifying topic significant query words and URLs and the MapReduce computing paradigm are collectively utilized to alleviate the efficiency bottleneck of UPM. Based on the user profiles obtained from the UPM, the ranking of the suggested queries is organized to align with the current information need as well as the user's personal preference. To evaluate the effectiveness of QS-DP, we conduct extensive experiments on a large-scale query log. Compared with several strong baselines, the proposed framework demonstrates superior performance with respect to a variety of metrics.

The contributions of this paper are summarized as follows:

• We propose a new representation for query log. The proposed multi-bipartite-graph representation comprehensively captures different kinds of relations between search queries in query log. A sophisticated mechanism is further designed to differentiate the importance of different relations between search queries.

- We design a novel two-phase method to iteratively obtain query suggestion candidates. In the first phase, we propose two different strategies to identify the first query suggestion candidate. The two strategies utilize different approaches to integrate the information in multi-bipartite-graph. In the second phase, we utilizes technique such as hitting time and the multi-bipartite-graph representation to enhance both the relevance and diversity of the set of query suggestion candidates.
- We develop the *User Profiling Model* (UPM), to profile search engine users through integrating search sessions, web dynamics as well each user's preference of word usage and URL clicking. The efficiency of UPM is significantly enhanced by a novel technique that identifies topic significant query words and URLs and the MapReduce computing paradigm.

The rest of this paper is organized as follows. In Section 2, we review the related work. In Section 3, we discuss the query log representation component. In Sections 4 and 5, we present the diversification component and the personalization component. In Section 6, we report the experimental results. Finally, we conclude the paper in Section 7.

#### 2. Related work

Relevance-oriented query suggestion attracts much attention in recent years. Most of the relevance-oriented methods rely on the click graph to represent the information in query log. Mei et al. [9] proposed an algorithm using hitting time to make query suggestion on a click graph. Cao et al. [2] proposed a query suggestion approach by using the concept sequence suffix tree. Youngho et al. [10] proposed a boolean query suggestion technique, which generates boolean queries by exploiting decision trees learned from pseudo-labeled documents. There also exists some work that focus on specific types of queries. For example, Szpektor et al. [5] proposed a method to extend the reach of query recommendation to long-tail queries by reasoning about rules between query emplates rather than individual query transitions. Kato et al. [4] proposed an approach to present query suggestions to the user and the method was designed to help query reformulation actions such as specialization and parallel movement. Markov random walk on the click graph was studied in [11,12] for ranking documents and discovering search tasks. Deng et al. [13] proposed a new framework for modeling the click graph, in which the various query-URL pairs are treated differently. Besides viewing query log as click graph, probabilistic topic modeling technique is gaining momentum in text mining [14,15] and researchers have developed different topic models [16,17] to analyze search engine query log. Xianghua et al. [18] utilized probabilistic topic models to analyze social views. Zhou et al. [19] proposed a topic-sensitive probabilistic model that is effectively in finding the community by incorporating link and user analysis into a unified framework. [20] proposes an approach for text classification using semi-supervised clustering to improve the performance of text classification.

Recently researchers have been aware of the limitations of relevance-oriented query suggestion and explored several strategies to incorporate either diversification or personalization into query suggestions. Ma et al. [6] proposed a method based on Markov random walk and hitting time to diversify the query suggestion results. Song et al. [7] introduced a query suggestion framework which works in a post-ranking fashion. This framework is essentially different from those proposed in [6,9] because it relies on external knowledge such as the search results from the search engine. Leung et al. [8] introduced an approach that captures the user's conceptual preferences in order to provide personalized query suggestions. Mei et al. [9] also proposed a Download English Version:

## https://daneshyari.com/en/article/402620

Download Persian Version:

https://daneshyari.com/article/402620

Daneshyari.com