Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

Date of Publication: April 2013

PMSE: A Personalized Mobile Search Engine

Abstract

We propose a personalized mobile search engine, PMSE, that captures the users' preferences in the form of concepts by mining their clickthrough data. Due to the importance of location information in mobile search, PMSE classifies these concepts into content concepts and location concepts. In addition, users' locations (positioned by GPS) are used to supplement the location concepts in PMSE. The user preferences are organized in an ontology-based, multi-facet user profile, which are used to adapt a personalized ranking function for rank adaptation of future search results. To characterize the diversity of the concepts associated with a query and their relevances to the users need, four entropies are introduced to balance the weights between the content and location facets. Based on the client-server model, we also present a detailed architecture and design for implementation of PMSE. In our design, the client collects and stores locally the clickthrough data to protect privacy, whereas heavy tasks such as

Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

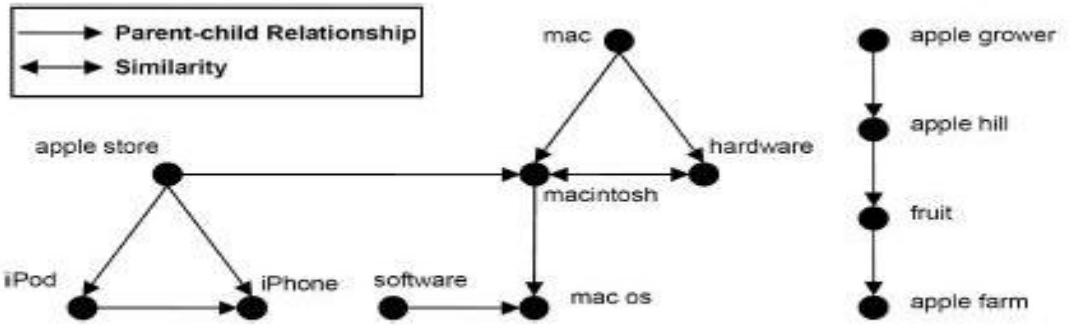
Date of Publication: April 2013

concept extraction, training and reranking are performed at the PMSE server. Moreover, we address the privacy issue by restricting the information in the user profile exposed to the PMSE server with two privacy parameters. We prototype PMSE on the Google Android platform. Experimental results show that PMSE significantly improves the precision comparing to the baseline.

INTRODUCTION

A major problem in mobile search is that the interactions between the users and search engines are limited by the small form factors of the mobile devices. As a result, mobile users tend to submit shorter, hence, more ambiguous queries compared to their web search counterparts. In order to return highly relevant results to the users, mobile search engines must be able to profile the users' interests and personalize the search results according to the users' profiles. A practical approach to capturing a user's interests for personalization is to analyze the user's clickthrough data [5], [10], [15], [18]. Leung, et. al., developed a search engine personalization method based on users' concept preferences and showed that it is more effective than methods that are based on

page preferences [12]. However, most of the previous work assumed that all concepts are of the same type. Observing the need for different types of concepts, we present in this paper a personalized mobile search engine, PMSE, which represents different types of concepts in different ontologies. In particular, recognizing the importance of location information in mobile search, we separate concepts into location concepts and content concepts.



Example Content Ontology Extracted for the Query "apple".

RELATED-WORK

Clickthrough data has been used in determining the users' preferences on their search results. Table 1, showing an example clickthrough data for the query "hotel", composes of the search results and the ones that the user clicked on (bolded search results in Table 1). As shown, ci's are the content concepts and li's are the location concepts extracted from

Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

Date of Publication: April 2013

the corresponding results. Many existing personalized web search systems [6], [10], [15], [18] are based clickthrough data to determine users' preferences. Joachims [10] proposed to mine document preferences from clickthrough data. Later, Ng, et. al. [15] proposed to combine a spying technique together with a novel voting procedure to determine user preferences. More recently, Leung, et. al. [12] introduced an effective approach to predict users' conceptual preferences from clickthrough data for personalized query suggestions. Search queries can be classified as content (i.e., non-geo) or location (i.e., geo) queries. Examples of location queries are "hong kong hotels", "museums in london" and "virginia historical sites". In [9], Gan, et. al., developed a classifier to classify geo and non-geo queries. It was found that a significant number of queries were location queries focusing on location information. In order to handle the queries that focus on location information, a number of location-based search systems designed for location queries have been proposed. Yokoji, et. al. [22] proposed a location-based search system for web documents. Location information were extracted from the web documents, which was converted into latitude-longitude pairs. When a user submits a query together with a latitudelongitude pair, the system creates a search

Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

Date of Publication: April 2013

circle centered at the specified latitude-longitude pair and retrieves documents containing location information within the search circle.

SYSTEM-DESIGN

Figure 1 shows PMSE's client-server architecture, which meets three important requirements. First, computation intensive tasks, such as RSVM training, should be handled by the PMSE server due to the limited computational power on mobile devices. Second, data transmission between client and server should be minimized to ensure fast and efficient processing of the search. Third, clickthrough data, representing precise user preferences on the search results, should be stored on the PMSE clients in order to preserve user privacy. In the PMSE's client-server architecture, PMSE clients are responsible for storing the user clickthroughs and the ontologies derived from the PMSE server. Simple tasks, such as updating clickthoughs and ontologies, creating feature vectors, and displaying reranked search results are handled by the PMSE clients with limited computational power. On the other hand, heavy tasks, such as RSVM training and reranking of search results, are handled by the PMSE server.

Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

Date of Publication: April 2013

USER-INTEREST-PROFILING

PMSE uses “concepts” to model the interests and preferences of a user. Since location information is important in mobile search, the concepts are further classified into two different types, namely, content concepts and location concepts. The concepts are modeled as ontologies, in order to capture the relationships between the concepts. We observe that the characteristics of the content concepts and location concepts are different. Thus, we propose two different techniques for building the content ontology (in Section 4.1) and location ontology (in Section 4.2). The ontologies indicate a possible concept space arising from a user’s queries, which are maintained along with the clickthrough data for future preference adaptation. In PMSE, we adopt ontologies to model the concept space because they not only can represent concepts but also capture the relationships between concepts. Due to the different characteristics of the content concepts and location concepts, In Section 4.1, we first discuss our method to mine and build the content ontology from the search results. In Section 4.2, we present our method to derive a location ontology from the search results.

Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

Date of Publication: April 2013

DIVERSITY-AND-CONCEPT-ENTROPY

PMSE consists of a content facet and a location facet. In order to seamlessly integrate the preferences in these two facets into one coherent personalization framework, an important issue we have to address is how to weigh the content preference and location preference in the integration step. To address this issue, we propose to adjust the weights of content preference and location preference based on their effectiveness in the personalization process. For a given query issued by a particular user, if the personalization based on preferences from the content facet is more effective than based on the preferences from the location facets, more weight should be put on the content-based preferences; and vice versa. The notion of personalization effectiveness is derived based on the diversity of the content and location information in the search results as discussed in Section 5.1, and the diversity of user interests the content and location information associated with a query as discussed in Section 5.2. We show that it can be used to effectively combine a user's content and location preferences for reranking the search results in Section 8.4.

Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

Date of Publication: April 2013

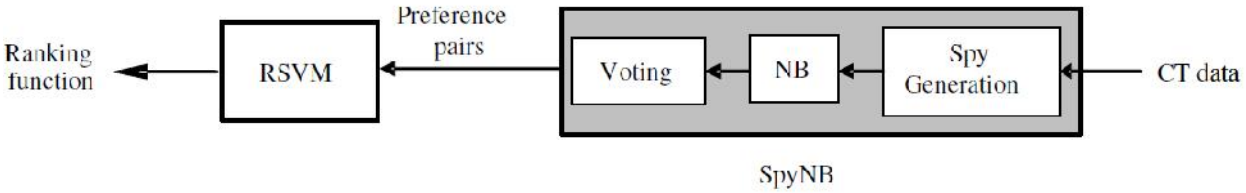
USER-PREFERENCES-EXTRACTION-AND-PRIVACY-PRESERVATION

Given that the concepts and clickthrough data are collected from past search activities, user's preference can be learned. These search preferences, inform of a set of feature vectors, are to be submitted along with future queries to the PMSE server for search result re-ranking. Instead of transmitting all the detailed personal preference information to the server, PMSE allows the users to control the amount of personal information exposed. In this section, we first review a preference mining algorithms, namely SpyNB Method, that we adopt in PMSE, and then discuss how PMSE preserves user privacy. SpyNB [15] learns user behavior models from preferences extracted from clickthrough data. Assuming that users only click on documents that are of interest to them, SpyNB treats the clicked documents as positive samples, and predict reliable negative documents from the unlabeled (i.e. unclicked) documents. To do the prediction, the "spy" technique incorporates a novel voting procedure into Na¨ıve Bayes classifier [14] to predict a negative set of documents from the unlabeled document set. The details of the SpyNB method can be found in [15]. Let P be the positive

Knowledge and Data Engineering, IEEE Transactions on (Volume:25 , Issue: 4)

Date of Publication: April 2013

set, U the unlabeled set and PN the predicted negative set (PN ⊂ U) obtained from the SpyNB method. SpyNB assumes that the user would always prefer the positive set over the predicted negative set.



PERSONALIZED-RANKING-FUNCTIONS

Upon reception of the user's preferences, Ranking SVM (RSVM) [10] is employed to learn a personalized ranking function for rank adaptation of the search results according to the user content and location preferences. For a given query, a set of content concepts and a set of location concepts are extracted from the search results as the document features. Since each document can be represented by a feature vector, it can be treated as a point in the feature space. Using the preference pairs as the input, RSVM aims at finding a linear ranking function, which holds for as many document preference pairs as possible. An adaptive implementation, SVMlight available at [3], is used in our experiments. In the following, we discuss two issues in the RSVM training process: 1) how to extract the feature

vectors for a document; 2) how to combine the content and location weight vectors into one integrated weight vector.

CONCLUSION

To adapt to the user mobility, we incorporated the user's GPS locations in the personalization process. We observed that GPS locations help to improve retrieval effectiveness, especially for location queries. We also proposed two privacy parameters, minDistance and expRatio, to address privacy issues in PMSE by allowing users to control the amount of personal information exposed to the PMSE server. The privacy parameters facilitate smooth control of privacy exposure while maintaining good ranking quality. For future work, we will investigate methods to exploit regular travel patterns and query patterns from the GPS and clickthrough data to further enhance the personalization effectiveness of PMSE.