An Optimized Personalized Mobile Search Engine towards Secured Data Analysis

B. Bhanu Prakash 1, V. Kesav Kumar 2
1 M.Tech student, 2 Associate prof, Sri Mittapalli College of Engineering, India

ABSTRACT: As we proposed an personalized mobile search engine, PMSE where its role is to capture all the users preferences that is in the form of concepts by greatly minimizes clickthrough data. Now because of location information in the mobile search PMSE divides these concepts in to their content concepts as well as location concepts. Moreover to this user’s location were used to provide location concepts in the PMSE. So user preferences are formulated in to ontology based, multi facet user profile that are helpful for the personalized ranking functions for the rank adaptation of the feature search results. Also to characterize the diversity of particular concepts those are linked with the query and their relevance’s to the users needful. Finally we demonstrated detailed architecture and design implementation of PMSE based on client server model. Here in this system client accumulate and stores locally click through data to protect privacy also heavy tasks like concept extraction, training and also reranking have to be at PMSE server. we also prototyped PMSE based on Google Android platform.

Keywords - Clickthrough data, ontology, personalisation, reranking, user profiling

I. INTRODUCTION

As we seen the big issue in the field of mobile search is that the interactions between the existed users and also search engine that are limited by the short form factors of the mobile devices. Due to this effect, mobile users have to submit shorter, in similar way more dilemma queries as compared to their web search counterparts. So therefore to return highly relevant results to the users, mobile search engines should be capable to profile the users’ interests only and also personalize their search results accordingly to the all users’ profiles. In personalized mobile search engine (PMSE) that captures the users preferences in the form of concepts by mining their click through data. Knowing the importance of the location information in mobile search, this search engine captures users preferences in the form of concepts viz., content concept and location concept. Location information are supplement to the location concept. User can also submit the location by simply typing it on a particular column or GPS helps. 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 relevance’s to the users need, four entropies are introduced to balance the weights between the content and location facets.

Observing the need for different types of concepts, we present in this paper an ontology based personalized mobile search engine (OBPMSE) 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. For example, a user who is planning to visit Japan may issue the query “hotel,” and click on the search results about hotels in Japan. From the clickthroughs of the query “hotel,” OBPMSE can learn the user’s content preference (e.g., “room rate” and “facilities”) and location preferences (“Japan”). Accordingly, OBPMSE will favor results that into location concepts and content concepts. For example, a user who wishes to visit Tourists places in India may submit query as Tourists places. From that query keyword “Tourists place”, OBPMSE understand user’s content preference (“India”). That all results will show again if user submit “Tourist”. If user is searching for Shopping mall whose location is Delhi. This gives location of all shopping malls nearby Delhi to the user. The introduction of location preferences offers OBPMSE an additional dimension for capturing a user’s interest and an opportunity to enhance search quality for users.

I. LITERATURE SURVEY

As discussed in [3], Gan et al. postulated a classifier that role is to classify geo as well as non-geo queries. They observed that a crucial number
of queries were seems to be location aware queries basically focuses on particular location information. So to tackle this queries that focus on location information, and a number of location-based search systems designed for location queries also been proposed by them.

As discussed in Yokoji [7] formulated an location-based search system for the requirement of web documents. In this location information have to be extracted from the existed web documents, that work is to be convert into the latitude-longitude pairs only. So basically when a user submits immediately a query together with a latitude-longitude pair, then the system creates job search circles that is centered at the predetermined latitude-longitude pair and restore all the documents that contains location information within the confined search circle.

Later on, Chen et al. [2] studied the problem of efficient query processing in location-based search systems. A query is assigned with a query footprint that specifies the geographical area of interest to the user. Several algorithms are employed to rank the search results as a combination of a textual and a geographic score.

Later, Ng et al. [6] proposed to combine a spying technique together with a novel voting procedure to determine user preferences.

More recently, Leung et al. [4] introduced an effective approach to predict users’ conceptual preferences from clickthrough data for personalized query suggestions.

More recently, Li et al. [5] proposed a probabilistic topic-based framework for location-sensitive domain information retrieval. Instead of modeling locations in latitude-longitude pairs, the model assumes that users can be interested in a set of location sensitive topics. It recognizes the geographical influence distributions of topics, and models it using probabilistic Gaussian Process classifiers.

II. PROPOSED SYSTEM:

As depicted in fig. 1 the job of OBPMSE’s client-server architecture is to meets three important requirements compulsory. As we seen first, is to computation-intensive tasks, that is like RSVM training, must be handle by the OBPMSE server because of very limited computational power on mobile devices. Second, one is the data transmission between the client and server must be minimized to make sure fast and efficient processing of the search. Third, one is click through data, that work is to precise user preferences on the search results, should be stored on the OBPMSE clients in order to restore user privacy.

Fig1. OBPMSE Client-Server Architecture

As we seen that OBPMSE’s client-server architecture, here the OBPMSE clients are employed for storing the user clickthroughs as well as ontologies obtained basically from the OBPMSE server. Now the Simple tasks like, updating clickthroughs and ontologies, and creating feature vectors, and also displaying reranked search results that were to be handled by the particular OBPMSE clients with less system computational power. So therefore at the other end, huge tasks, such as RSVM training and reranking of search results, are tackled by the OBPMSE server. Further in order to reduce the data transmission between client and also server, the OBPMSE client could only need to submit a particular query all together with the feature vectors with that of OBPMSE server, as well as the server would in this same case.
III. USER INTEREST PROFILING

As discussed earlier PMSE uses “concepts” to formulate the interests and also preferences of a user. As we know location information is important in mobile search, the concepts are in turn divided into two different methods, namely, content concepts and location concepts. These concepts are modeled as ontologies, so in order to capture the relationships between the concepts. We also observe that the characteristics of the content concepts and location concepts are different. Therefore, we wish to propose two different techniques for building the content ontology and location ontology.

A. Content Ontology

Here content concept extraction method extracts all the keywords and phrases (excluding the stop words) from the web-snippets that are generated from q. If a keyword/phrase exists in many cases in the web-snippets arising from the query q, we could treat it as an important concept related to the query, as it coexists in close proximity with the query in the top documents. As given following support formula that is inspired by the well-known problem of finding frequent item sets in data mining [8] is used to measure the importance of a particular keyword/phrase ci with respect to the query q:

\[
\text{support}(C_i) = \frac{s_{fci}}{n} \cdot |C_i| \quad \text{...........1}
\]

Where \(s_{fci}\) is the snippet frequency of the keyword/phrase ci (i.e., the number of web-snippets containing ci), n is the number of web-snippets returned and \(|C_i|\) is the number of terms in the keyword/phrase ci.

We adopt the following two propositions to determine the relationships between concepts for ontology formulation:

Similarity. Two concepts which coexist a lot on the search results might represent the same topical interest.

Parent-child relationship. More specific concepts often appear with general terms, while the reverse is not true.

Fig. 2 shows an example content ontology created for the query “hotel,” where content concepts linked with a one-sided arrow (\(\rightarrow\)) are parent-child concepts, and concepts linked with a double-sided arrow (\(\leftrightarrow\)) are similar concepts.

![Fig. 2. Ontology for q hotel with P = 0.2; 0.5; 1.0.](image)

In general, the ontology covers more than what the user actually wants. The concept space for the query “hotel” consists of “map,” “reservation,” “room rate,”..., etc. If the user is indeed interested in information about hotel rates and clicks on pages containing “room rate” and “special discount rate” concepts, the captured clickthrough favors the two clicked concepts. Feature vectors containing the concepts “room rate” and “special discount rate” as positive preferences will be created corresponding to the query “hotel.” As indicated in Fig. 2, when the query is issued again later, these feature vectors will be transmitted to the PMSE server and transformed into a content weight vector to rank the search results according to the user’s content preferences.
B. Location Ontology

We observe two important issues in location ontology formulation.

- First, a document usually embodies only a few location concepts, and thus only very few of them co-occur with the query terms in web-snippets.
- Second, the similarity and parent-child relationship cannot be accurately derived statistically because the limited number of location concepts embodied in documents.

IV. DIVERSITY AND CONCEPT_ENTROPY

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.

A. Diversity of Content and Location Information

To characterize the content and location properties of the query, we use entropy to estimate the amount of content and location information retrieved by a query.

Since we are concerned with content and location information only in this paper, we define two entropies, namely, content entropy \( H_c(q) \) and location entropy \( H_l(q) \), to measure, respectively, the uncertainty associated with the content and location information of the search results

\[
H_c(q) = - \sum_{i=1}^{k} P(C_i) \log P(C_i) \\
H_l(q) = - \sum_{i=1}^{m} P(L_i) \log P(L_i)
\]

Where \( k \) is the number of content concepts.

B. Diversity of User Interests

We also introduce click content entropy and click location entropy to indicate, respectively, the diversity of a user’s interest on the content and location information returned from a query. The entropy equations for click content and location concepts are similar to (2), but only the clicked pages, and hence the clicked concepts, are considered in the formula. Since the click entropies reflect the user’s actions in response to the search results, they can be used as an indication of the diversity of the user’s interests.

C. Personalization Effectiveness

For click entropies, we expect that the higher the click content/location entropies, the worse the personalization effectiveness, because high click content/location entropies indicate that the user is clicking on the search results with high uncertainty, meaning that the user is interested in a diversity of information in the search results. When the user’s interests are very broad (or the clickthroughs could be “noisy” due to irrelevant concepts existing in the clicked documents), it is difficult to 1) find out the user’s actual needs and 2) personalize the search results toward the user’s interest. On the other hand, if the click content/ location entropies are low, the personalization effectiveness would be high because the user has a focus on certain precise topic in the search results (only a small set of content/ location concepts has been clicked by the user).

Hence, the profiling process can identify the user’s information needs and the personalization process can personalize the results to meet those needs.

V. USER PREFERENCES EXTRACTION AND PRIVACY PRESERVATION

Instead of transmitting all the detailed personal preference information to the server, PMSE allows the users to control the amount of personal information exposed.

To control the amount of personal information exposed out of users’ mobile devices, PMSE filters
the ontologies according to the user’s privacy level setting, which are specified with two privacy parameters, minDistance and expRatio. The privacy preserving technique in PMSE aims at filtering concepts that are too specific. Thus, minDistance is used to measure whether a concept is far away from the root (i.e., too specific) in the ontology-based user profiles.

VI. PERSONALIZED RANKING FUNCTIONS

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.

The extraction of content feature vector and location feature vector are defined formally as follows:

1. Content feature vector. If content concept \( C \) is in a web-snippet \( S \), their values are incremented in the content feature vector \( C(q, d_k) \) with the following equation:

\[
\forall c_j \in s_k, \phi_C(q, d_k)[c_j] = \phi_C(q, d_k)[c_j] + 1.
\]

2. Location feature vector. If location concept \( L \) is in a web-snippet \( d_k \), its value is incremented in the location feature vector \( L(q, d_k) \) with the following equation

\[
\forall l_i \in d_k, \phi_L(q, d_k)[l_i] = \phi_L(q, d_k)[l_i] + 1.
\]

VII. EXPERIMENTAL RESULTS

We study the effect of noise clicks on the personalization quality. The accuracy of the estimated facet combination threshold. Figures show the flow of the test and evaluation processes.
CONCLUSION

The proposed OBPMSE uses content concept and location concepts which are modeled as ontologies. To adapt to the user mobility user's
GPS locations are used in the personalization process which helps to improve retrieval effectiveness, especially for location queries. Privacy is addressed by allowing users to control the amount of personal information exposed to the OBPMSE server. 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.

REFERENCES


BIODATA

Author

B. Bhanu Prakash presently pursuing M.Tech software engineering from Sri Mittapalli College of Engineering, India. He completed his B.Tech in Information Technology in 2007-2011 from E.V.M College of engineering and technology, India.

Co-Author

V. Kesav Kumar working as Associate Professor in CSE Department in Sri Mittapalli College of Engineering, India.