Optimized Mobile Search Engine Using Click-Through Data

M. Divya

Assistant Professor, Dept of IT
Christu Jyothi Institute of Technology & Science, Jangaon

Abstract—Data mining is a system employing for more computer learning technique to automatically analyse and extracting knowledge from data stored in the database. The goal of data mining is to extract hidden predictive information from database. This paper make use of data mining concept for collecting user's multiple preference from click through data. we propose a personalized mobile search engine (PMSE) that captures the users' preferences in the form of concepts by mining their click through 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 represent 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 relevance to the user's need. 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 click through data to protect privacy, whereas heavy tasks such as concept extraction, training, and reranking are performed at the PMSE server. Moreover, we prototype PMSE on the Google Android platform.

1. 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 counter parts. In order to return highly related 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. We present in this paper a personalized mobile search engine(PMSE) which represents different types of concepts in different ontologies.

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 click-through of the query "hotel," PMSE can learn the user's content preference (e.g., "room rate" and "facilities") and location preferences ("Japan"). Accordingly, PMSE will favour results that are concerned with hotel information in Japan for future queries on "hotel." The introduction of location preferences offers PMSE an additional dimension for capturing a user's interest and an opportunity to enhance search quality for users.

GPS locations play an important role in mobile web search. For example, if the user, who is searching for hotel information, is currently located in "Shinjuku, Tokyo," his/her position can be used to personalize the search results to favour information about nearby hotels. Here, we can see that the GPS locations (i.e., "Shinjuku, Tokyo") help reinforcing the user's location preferences (i.e., "Japan") derived from a user's search activities to provide the most relevant results. A realistic design for PMSE by adopting the meta search approach which relies on one of the commercial search engines, such as Google, Yahoo, or Bing, to perform an actual search. The client is responsible for receiving the user's requests, submitting the requests to the PMSE server, displaying the returned results, and collecting his/her click through in order to derive his/her personal preferences. The PMSE server, on the other hand, is responsible for handling heavy tasks such as forwarding the requests to a commercial search engine, as well as training. And reranking of search results before they are returned to the client. The user profiles for specific users are stored on the PMSE

clients, thus preserving privacy to the users. PMSE has been prototyped with PMSE clients on the Google Android platform and the PMSE server on a PC server to validate the proposed ideas.

We also recognize that the same content or location concept may have different degrees of importance to different users and different queries. To formally characterize the diversity of the concepts associated with a query and their relevance to the user's need, we introduce the notion of content and location entropies to measure the amount of content and location information associated with a query. Similarly, to measure how much the user is interested in the content and/or location information in the results, we propose click content and location entropies. Based on these entropies, we develop a method to estimate the personalization effectiveness for a particular query of a given user, which is then used to strike a balanced combination between the content and location preferences. The results are re -ranked according to the user's content and location preferences before returning to the client.
The main contributions of this paper are as follows:

- This paper studies the unique characteristics of content and location concepts, and provides a coherent strategy using a client-server architecture to integrate them into a uniform solution for the mobile environment.
- It studies the unique characteristics of content and location concepts, and provides a coherent strategy using a client-server architecture to integrate them into a uniform solution for the mobile environment.

2. Related Work

Click through data have been used in determining the users' preferences on their search results. Many existing personalized web search systems are based on click through data to determine users' preferences. Joachims [10] proposed to mine document preferences from click through 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 click through 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.

Later on, Chen et al. [7] studied the problem of efficient query processing in location-based search systems. A query 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. More recently, Li et al. 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.

The differences between existing works and ours are:

- We propose and implement a new and realistic design for PMSE. To train the user profiles quickly and efficiently, our design forwards user requests to the PMSE server to handle the training and reranking processes.
- Existing works on personalization do not address the issues of privacy preservation. PMSE addresses this issue by controlling the amount of information in the client's user profile being exposed to the PMSE server using two privacy parameters, which can control privacy smoothly, while maintaining good ranking quality.

3. System Design

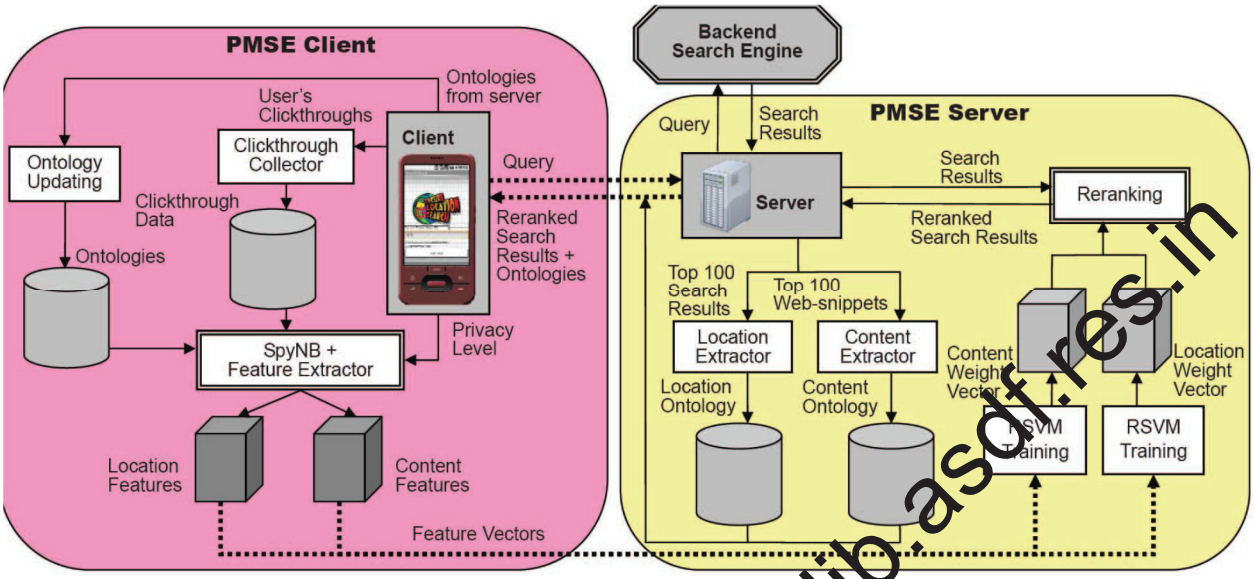


Fig. 1. The general process flow of PMSE

Fig. 1 shows PMSE's client-server architecture, which meets three important requirements. First, computation-intensive tasks, such as RSVM training for learning a linear weight vector (consisting both content and location features) to rank the search results. Second, data transmission between client and server should be minimized to ensure fast and efficient processing of the search. Third, click through 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 click through and the ontologies derived from the PMSE server. Simple tasks, such as updating click thoughts 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.

PMSE consists of two major activities:

1. Reranking the search results at PMSE server

When a user submits a query on the PMSE client, the query together with the feature vectors containing the user's content and location preferences (i.e., filtered ontologies according to the user's privacy setting) are forwarded to the PMSE server, which in turn obtains the search results from the back-end search engine (i.e., Google). The content and location concepts are extracted from the search results and organized into ontologies to capture the relationships between the concepts. The server is used to perform ontology extraction for its speed. The feature vectors from the client are then used in RSVM training to obtain a content weight vector and a location weight vector, representing the user interests based on the user's content and location preferences for the reranking. Again, the training process is performed on the server for its speed. The search results are then reranked according to the weight vectors obtained from the RSVM training. Finally, the reranked results and the extracted ontologies for the personalization.

2. Ontology update and clickthrough collection at PMSE client

Ontology formally represents knowledge as a set of concepts within a domain, and the relationships among those concepts. It can be used to reason about the entities within that domain and may be used to describe

the domain. Here, we are using the ontology concept to group the data as per the related domain. So that, if the user search the data, the data will displayed in domain they are requesting. Many geographical relationships among locations have already been captured as facts.The ontologies returned from the PMSE server contain the concept space that models the relationships between the concepts extracted from the search results. They are stored in the ontology database on the client. When the user clicks on a search result, the clickthrough data together with the associated content and location concepts are stored in the clickthrough database on the client. When the user clicks on a search result, the clickthrough data together with the associated content and location concepts are stored in the clickthrough database on the client. The clickthroughs are stored on the PMSE clients, so the PMSE server does not know the exact set of documents that the user has clicked on. This design allows user privacy to be preserved in certain degree. Two privacy parameters, min distance and expRatio, are proposed to control the amount of personal preferences exposed to the PMSE server. If the user is concerned with his/her own privacy, the privacy level can be set to high so that only limited personal information will be included in the feature vectors and passed along to the PMSE server for the personalization.

On the other hand, if a user wants more accurate results according to his/her preferences, the privacy level can be set to low so that the PMSE server can use the full feature vectors to maximize the personalization effect. To address privacy issues, clickthroughs are stored on the PMSE client, and the user could adjust the privacy parameters to control the amount of personal information to be included in the feature vectors, which are forwarded to the PMSE server for RSVM training to adapt personalized ranking functions for content and location preferences.

Content Ontology

Our content concept extraction method first extracts all the keywords and phrases (excluding the stop words) from the web-snippet sarising from q. If a keyword/phrase exists frequently in the web-snippets arising from the query q, we would treat it as an important concept related to the query, as it coexists in close proximity with the query in the top documents. The following support formula, which is inspired by the well-known problem of finding frequent item sets in data mining, is employed to measure the importance of a particular keyword/phrase p with respect to the query q.

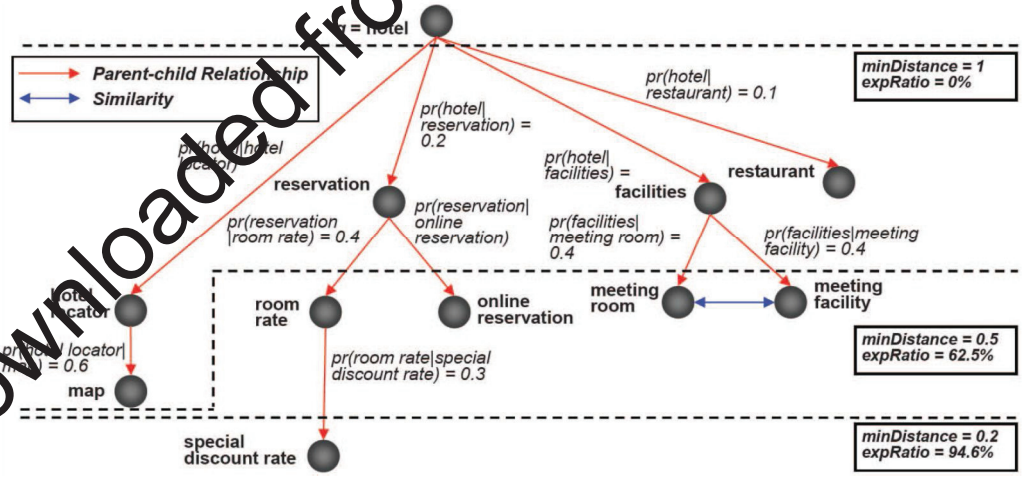


Fig. 2. Ontology for q ="hotel" with p=0.2,0.5,1.0

Fig. 2 shows an example content ontology created for the query "hotel," where content concepts linked with a one sided arrow (→) are parent-child concepts, and concepts linked with a double-sided arrow (←>) are similar concepts. Fig. 2 shows the possible concept space determined for the query "hotel," while the click through data determine the user preferences on the concept space.

Location Ontology

Our approach for extracting location concepts is different from that for extracting content concepts. 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. To alleviate this problem, we extract location concepts from the full documents. Second, the similarity and parent-child relationship cannot be accurately derived statistically because the limited number of location concepts embodied in documents. Furthermore, many geographical relationships among locations have already been captured as facts.

Table 1 Statistics of the Location Ontology

No. of Countries	8	Total no. of nodes	17899
No. of Regions	200	Country region nodes	200
No. of Provinces	5700	Region-Province nodes	1959
No. of Towns	10233	Province-City nodes	15897

We organize all the cities as children under their provinces, all the provinces as children under their regions, and all the regions as children under their countries. The statistics of our location ontology are provided in Table 1.The predefined location ontology is used to associate location information with the search results. All of the keywords and key-phrases from the documents returned for query q are extracted. If a keyword or key-phrase in a retrieved document d matches a location name in our predefined location ontology, it will be treated as a location concept of d. For example, assume that document d contains the keyword "Los Angeles." "Los Angeles" would then be matched against the location ontology. Since "Los Angeles" is a location in our location ontology, it is treated as a location concept related to d. Furthermore, we would explore the predefined location hierarchy, which would identify "Los Angeles" as a city under the state "California." Thus, the location "/United States/California/Los Angeles/" is associated with document d. If a concept matches several nodes in the location ontology, all matched locations will be associated with the document.

4. Experimental Evaluation

In this section, we evaluate the effectiveness of PMSE. We describe the experimental setup in the following section then, we evaluate the ranking quality of PMSE with different user profiles. We study the effect of noise clicks on the personalization quality.

Experiment Methodology

The experiment aims to answer the following question: Given that a user is only interested in some specific aspects of a query, can PMSE generate a ranking function personalized to the user's interest from the user's clickthroughs?

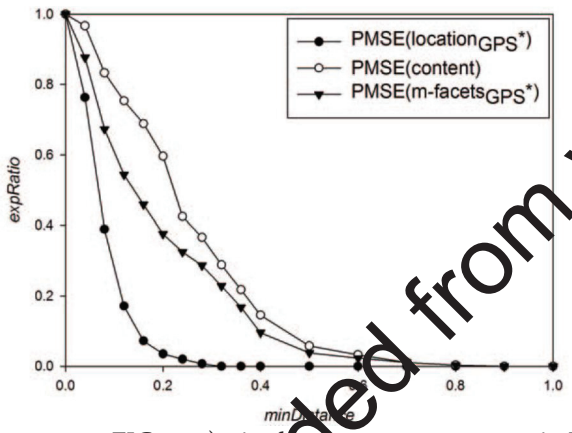
To answer this question, we need to evaluate the search results before and after personalization. The difficulty of the evaluation is that only the user who conducted the search can tell which of the results are relevant to his/her search intent. This is in contrast to the evaluation of traditional information retrieval systems [20], where expert judges are employed to judge the relevance of a set of documents (e.g., TREC) based on a detailed description of the information need. The relevance judgment is then considered the standard to judge the quality of the search results. This evaluation method clearly cannot be applied to personalized search, because what an expert judge considered as relevant to a query needs not be relevant from another user's point of view because the same query issued by two different users may have different goals behind it. Instead of using a small number of users each searching a large number of queries we use a large number of users each searching a small number of queries to prevent the users from overly adapted to the system.

For example, when the topical category is "photography" and the query is "canon," the user will look for information about "canon" digital cameras but not "canon" laser printers or "canon" as a location name. Yet, within the "photography" category, the user can decide what to look for, e.g., specific products, photo gallery, etc.

5 Analysis & Results

Privacy versus Ranking Quality

We evaluate PMSE's privacy parameters, min Distance and expRatio, against the ranking quality. We plot expRatio(the amount of private information exposed) against min Distance for a number of PMSE methods in Fig. 9a. The expRatio of PMSE(content), which employs content ontology only, decreases uniformly from 1 to nearly zero when min Distance increases from 0 to 0.7.minDistance measures the distance of a concept away from the root (i.e., too specific).Since the heights of the trees in the content ontology are mostly less than 0.7, most of the concepts are pruned when min Distance>0:7 in PMSE (content). On the other hand, the expRatio of PMSE (location GPS_), which employs location ontology only, decreases uniformly from 1 to nearly zero when min Distance increases from 0 to 0.3. The heights of the trees in the location ontology are mostly less than 0.3. We observe that a node in the location ontology can associate many children (e.g., a country has many provinces or states, a province/state has many cities). Once a node is pruned in the location ontology, all the children will also be pruned, thus expratio decreases much faster than that in PMSE (content). Finally, the expRatio of PMSE (m-facets GPS_), which employs both content and location ontologies, decreases faster than PMSE (content), but slower than PMSE (location GPS_).



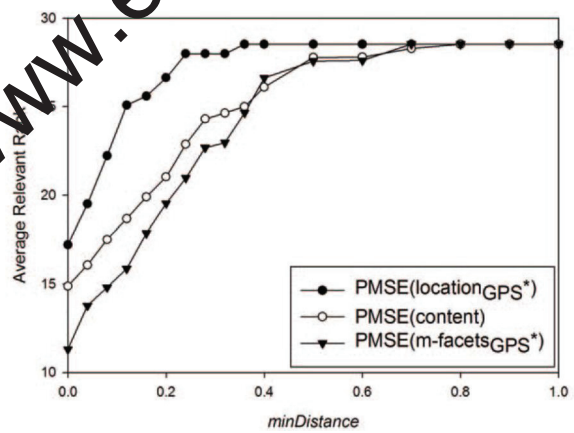


FIG 9 a)min distance versus exp ratio FIG 9 b)min distance vs average relevant rank

We plot the ARR of the search results against minDistance in Fig. 9b. As discussed before, the amount of private information exposed (expRatio) in PMSE (content) drops uniformly when minDistance increases from 0 to 0.7. Thus, the ARR of PMSE (content) increases uniformly when minDistance increases from 0 to 0.7. Similarly, the ARR of PMSE (location GPS_) increases uniformly when minDistance increases from 0 to 0.3, and the ARR of PMSE (m-facets GPS_)increases uniformly when minDistance increases from 0 to 0.6.

Conclusion

We proposed PMSE to extract and learn a user's content and location preferences based on the user's clickthrough.To adapt to the user mobility, we incorporated the user's GPS locations in the personalization process.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.

References

1. Appendix, http://www.cse.ust.hk/faculty/dlee/tkde-pmse/appendix.pdf, 2012.
2. Nat'l geospatial, http://earth-info.nga.mil/, 2012.
3. svmlight, http://svmlight.joachims.org/, 2012.
4. World gazetteer, http://www.world-gazetteer.com/, 2012.
5. E. Agichtein, E. Brill, and S. Dumais, "Improving Web SearchRanking by Incorporating User Behavior Information," Proc. 29thAnn. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.
6. E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning User Interaction Models for Predicting Web Search Result Preferences,"Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.