International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

DESIGN AND IMPLEMENTATION OF MOBILE SEARCH ENGINE USING ONTOLOGY

Abhimanyu S. Dutonde[1], Nikita Umare[2]

[1]CSE, AGPCE, RTMNU, Nagpur, Maharashtra (india)
[2]Assistant prof.,CSE, AGPCE, RTMNU, Nagpur, Maharashtra (india)

ABSTRACT

A personalized mobile search engine (PMSE) captures the users' preferences by mining through click data. The importance of location information in mobile search, PMSE classifies the concepts into content and location concepts. The users' locations (positioned by GPS) are used to supplement the location concepts in PMSE. The user preferences are organized in an ontology-based, multifacet user profile, which are used to adapt a personalized ranking function for rank adaptation of future search results. Here 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. The Address of the privacy issue is by restricting the information in the user profile exposed to the PMSE server with two privacy parameters the prototype PMSE on the Google Android platform.

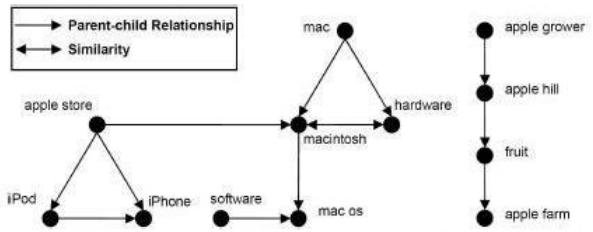
Keywords: Clickthrough data, concept, location search, mobile search engine, ontology, personalization, user profiling.

I. 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. The mobile users tend to submit shorter, hence, the ambiguous queries are 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. The approach is to capture a user's interests for personalization is to analyze the user's clickthrough data. In this method 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. The mobile search engine (MSE) represents different concepts in different ontologies. 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," MSE can learn the user's content preference and location preferences.The importance of location information in mobile search, leads to separate concepts into location concepts and content concepts.

The introduction of location preferences offers MSE an additional dimension for capturing a user's interest and an opportunity to enhance search quality for users. To incorporate context information we also take into account the visited physical locations of users in the MSE. GPS locations play an important role in mobile web search.

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354



Example Content Ontology Extracted for the Query "apple".

"Fig 1: Example content extracted for the query "apple""

GPS locations help reinforcing the user's location preferences derived from a user's search activities to provide the most relevant results. The framework is capable of combining a user's GPS locations and location preferences into the personalization process. The personalization framework is used that utilizes a user's content preferences and location preferences as well as the GPS locations in personalizing search results.

The client is responsiblefor receiving the user's requests, submitting the requests tothe MSE server, displaying the returned results, andcollecting his/her clickthroughs in order to derive his/herpersonal preferences. The MSE server, on the other hand, isresponsible for handling heavy tasks such as forwarding therequests to a commercial search engine, as well as trainingand reranking of search results before they are returned tothe client. The user profiles forspecific users are stored onthe MSE clients, thus preserving privacy to the users.Here the same content or locationconcept may havedifferent degrees of importance todifferent users and different queries. We introduce the notionof content and location entropies to measure the amount ofcontent and location information associated with a query, to measure how much the user is interested in thecontent and/or location information in the results, there is use click content and location entropies strategy. The method used here estimate the personalization for a particular query of a given user, which is then used to strike a balance combination between the content and location preferences.

The main points of this paper are:-

1) The unique characteristics of content and location concepts, and it provides a coherent strategy using a client-server architecture to integrate them into a uniform solution for the mobile environment.
2) By mining content and location concepts for user profiling, it utilizes both the content data.
3) Here the server-client model in which user queries are forwarded to a MSE server for processing the training and reranking quickly.
4) It incorporates a user's physical locations in the personalization process and location preferences to personalize search results for a user.
5) In privacy preservation the user send their profiles along with queries to the MSE server to obtain personalized search results. It addresses the privacy issue by allowing users to control their privacy levels with two privacy parameters, min distance and exp Ratio.

6) The ontology-based user profiles can successfully capture user's contents and location preferences and utilize the preferences to produce relevant results for the users.It significantly outperforms existing strategies which use either content or location preference only.

II. RELATED WORK

Clickthrough data have been used in determining the users'preferences on their search results.Search queries can be classified as content (i.e., non-geo) or location (i.e., geo) queries.

In another approach a classifier classifies geo and non-geo queries. 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 were designed for location queries. In location-based search system for web documents, Location information was extracted from the web documents, which was converted into latitude-longitude pairs. When a user submits a query together with a latitude-longitude pair, the system creates a search circle centred at the specified latitude-longitude pair and retrieves documents containing location information within the search circle.

In efficientquery processingin location-based search system. A query is assigned with a query footprint that specifies the geographical area of interest to the user. In a probabilistic topic-based framework a location-sensitive domain information retrieval system is used. Instead of modelling 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 difference between existing works and mine is.

The main points of this paper are:-

1) The unique characteristics of content and location concepts and it provide a coherent strategy using a client-server architecture to integrate them into a uniform solution for the mobile environment.

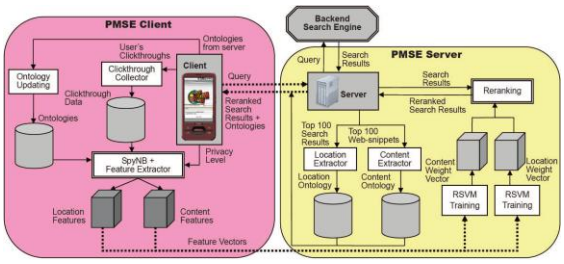
2) By mining content and location concepts for user profiling, it utilizes both the content data.

Most existing location-based search systems, require users to manually define their location preferences (with latitude-longitude pairs or text form), or to manually prepare a set of location sensitive. MSE profiles both of the user's content and location preferences in the ontology based user profiles, which are automatically learned from the clickthrough and GPS data without requiring extra efforts from the user.

In this new method we used a new and realistic design for MSE. To train the user profiles quickly and efficiently, the design forwards user requests to the MSE server to handle the training and reranking processes.

The Existing works on personalization do not address the issues of privacy preservation. MSE addresses this issue by controlling the amount of information in the client's user profile being exposed to the MSE server using two privacy parameters, which can control privacy smoothly, while maintaining good ranking quality.

International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 06 , June 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354



"Fi2: The general process flow of MSE."

Fig. 2 shows MSE's client-server architecture, which meet three important requirements. First, computation-intensive tasks, such as RSVM training, should be handled by the MSE 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 MSE clients in order to preserve user privacy.In the MSE's client-server architecture, MSE clients are responsible for storing the user clickthroughs and theontologies derived from the MSE server. Simple tasks, suchas updating clickthoughs and ontologies, creating feature vectors, and displaying reranked search results are handled by the MSE clients with limited computational power. On the other hand, heavy tasks, such as RSVM training and reranking of search results, are handled by the MSE server.

The data transmission cost is minimized, because only the essential data (i.e., query, feature vectors, ontologies and search results) are transmitted between client and server during the personalization process. MSE's design addressed the issues: 1) limited computational power on mobile devices, and 2) data transmission minimization.

MSE consists of two major activities:

1. Reranking The Search Results At PMSE Server.

When a user submits a query on the MSE 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 MSE 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 ontologyextraction 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 of future queries are returned to the client.

2. Ontology Update and click through Collection at MSE Client.

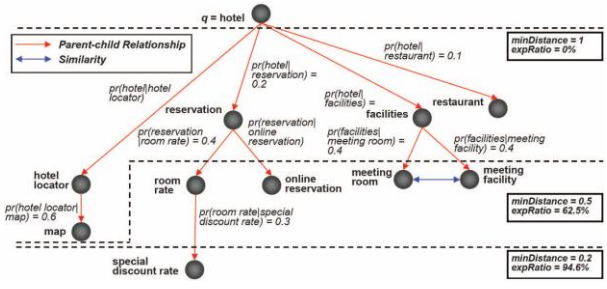
The ontologies returned from theMSE server contain the concept space that modelsthe relationships between the concepts extractedfrom the search results. They are stored in theontology database on the client. When the userclicks on a search result, the clickthrough datatogether with the associated content and locationconcepts are

stored in the clickthrough database onthe client. The clickthroughs are stored on the MSEclients, so the MSE server does not know the exactset of documents that the user has clicked on. Thisdesign allows user privacy to be preserved in certaindegree. Two privacy parameters, minDistance andexpRatio, are proposed to control the amount ofpersonal preferences exposed to the MSE server. Ifthe user is concerned with his/her own privacy, theprivacy level can be set to high so that only limitedpersonal information will be included in the featurevectors and passed along to the PMSE server for thepersonalization. On the other hand, if a user wantsmore accurate results according to his/her preferences;the privacy level can be set to low so that theMSE server can use the full feature vectors tomaximize the personalization effect.

III. 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. The characteristics of the content concepts and location concepts are different. Thus, here are two different techniques which are used for building the content ontology and location ontology. 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 MSE, we use 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, there is a method to derive location ontology from the search result.

Fig. 3 shows the possible concept space determined for the query "hotel," while the clickthrough data determine the user preferences on the concept space. 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 interestedin information about hotel rates and clicks on pages containing "room rate" and "special discount rate" concepts, the captured clickthroughfavors 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, this feature vectors will be transmitted to the PMSE server an transformed into a content weight vector to rank the search results according to the user's content preferences.



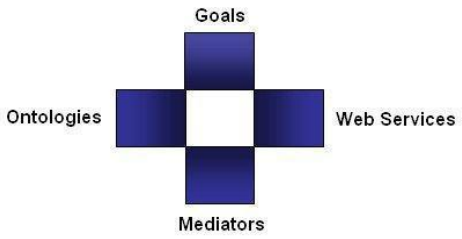
"Fig 3: Ontology for q ¼ 00hotel00 with p ¼ 0:2; 0:5; 1:0"

Location Ontology

The approach for extracting location concepts is different from that for extracting content concepts. The two important issues in location ontology formulate. 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 solve 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.

All the cities are organized as children under their provinces, all the provinces as children under their regions, and all the regions as children under their countries.

Extract location concepts are different from with the purpose of extracting content concepts with similar query travel patterns results from ARM. The predetermined location ontology with QTP is used to associate region information with the explore results. The entirepart of the keywords and key-phrases from the Query patternsdocuments (QPD) returned for query (UGQ) are extracted with exact matches of the results in location concept.



"Fig 4: model of the personalized search"

Content ontology method extracts all the keywords or terms and phrases from the web snippets and search engine results by user given query (UGQ). Here the most repeated UGQ based query patterns are analyzed after that it calculate the confidence value for moat time occurrence of the USQ in top documents measure the amount of a particular keyword/phrase $C_i$with value to UGQ where ( ) is the snippet frequency related to concepts $C_i$and n is the number of web-snippets from UGQ and | C_i | is the numeral of conditions in the keyword/phrase Ci ( ) is the snippet frequency containing the most related information.

Content Ontology

Our content concept extraction method first extracts all the keywords and phrases (excluding the stop words) from the web-snippets2 arising 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 ci with respect to the query

q: supportðciÞ¼

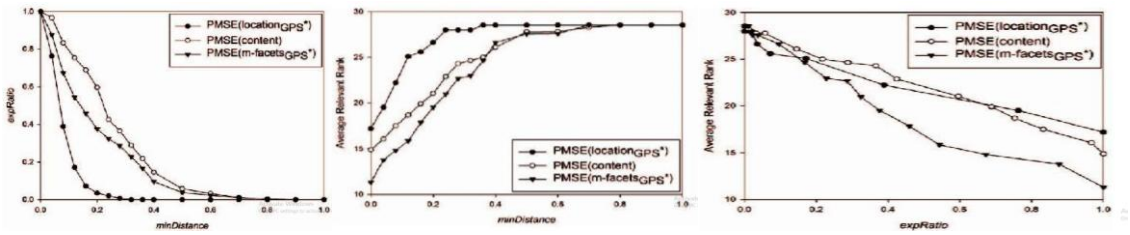
Fig. 3 shows an example content ontology created for the query "hotel," where content concepts linked with a one- sided arrow (!) are parent-child concepts, and concepts linked with a double-sided arrow ($) are similar concepts. Fig. 3 shows the possible concept space determined for the query "hotel," while the clickthrough data

determine the user preferences on the concept space. 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 clickthroughfavors the two clicked concepts. Feature vectors containing the concepts "room rate" and "special discount rate" as positive preferences will

IV. PROPOSED SYSTEM

Given that the concepts and click-through 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 MSE server for search result re-ranking. Instead oftransmitting all the detailed personal preference information to the server, MSE allows the users to control the amount of personal information exposed. In this section, we first review a preference mining algorithms, namely SpyNBMethod, that we adopt in PMSE, and then discuss how MSE preserves user privacy. SpyNB learns user behavior models from prefer-ences extracted from click-through 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., un-clicked) documents. To do the prediction, the "spy" technique incorporates a novel voting procedure into Naive Bayes classifier to predict a negative set of documents from the unlabeled document set. The details of the SpyNB method can be found in. Let P be the positive 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.

MSE 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 MSE aims at filtering concepts that is 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. If the user is concerned withhis/her own privacy, the privacy level can be set to high to provide only limited personal information to the MSE server. Nevertheless, the personalization effect will be less effective. On the other hand, if a user wants more accurate results according to his/her preferences, the privacy level can be set to low, such that the MSE server can use the full user profile for the personalization process, and provide better results.



"Fig5: Relationship between privacy parameters and ranking quality with different MSE methods"

V. CONCLUSION AND FUTURE WORK

The proposed personalized mobile search engine is an innovative approach for personalizing web search results. By mining content and location concepts for user profiling, it utilizes both the content and location preferences to personalize search results for a user.

The results show that GPS location helps improve retrieval effectiveness for location queries (i.e., queries that retrieve lots of location information).

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 MSE.

REFERENCES

Links

[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.

Conference

[5] E. Agichtein, E. Brill, and S. Dumais, "Improving Web SearchRanking by Incorporating User Behavior Information," Proc. 29th Ann. 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.

[7] Y.-Y. Chen, T. Suel, and A. Markowetz, "Efficient Query Processing in Geographic Web Search Engines," Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval(SIGIR), 2006.

[8] K.W. Church, W. Gale, P. Hanks, and D. Hindle, "Using Statistics in Lexical Analysis," Lexical Acquisition: Exploiting On-LineResources to Build a Lexicon, Psychology Press, 1991.

[9] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, "Analysis of Geographic Queries in a Search Engine Log," Proc.First Int'l Workshop Location and the Web (LocWeb), 2008.

[10] T. Joachims, "Optimizing Search Engines Using ClickthroughData," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and DataMining, 2002.

[11] K.W.-T. Leung, D.L. Lee, and W.-C. Lee, "Personalized Web Search with Location Preferences," Proc. IEEE Int'l Conf. Data Mining (ICDE), 2010.

[12] K.W.-T. Leung, W. Ng, and D.L. Lee, "Personalized Concept-BasedClustering of Search Engine Queries," IEEE Trans. Knowledge andData Eng., vol. 20, no. 11, pp. 1505-1518, Nov. 2008.

[13] H. Li, Z. Li, W.-C. Lee, and D.L. Lee, "A Probabilistic Topic-BasedRanking Framework for Location-Sensitive Domain InformationRetrieval," Proc. Int'l ACM SIGIR Conf. Research and Development inInformation Retrieval (SIGIR), 2009.

[14] B. Liu, W.S. Lee, P.S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," Proc. Int'l Conf. MachineLearning (ICML), 2002.

[15] W. Ng, L. Deng, and D.L. Lee, "Mining User Preference Using SpyVoting for Search Engine Personalization," ACM Trans. Internet Technology, vol. 7, no. 4, article 19, 2007.

[16] J.Y.-H. Pong, R.C.-W.Kwok, R.Y.-K.Lau, J.-X.Hao, and P.C.-C.Wong, "A Comparative Study of Two Automatic Document Classification Methods in a Library Setting," J. Information Science, vol. 34, no. 2, pp. 213-230, 2008.

[17] C.E. Shannon, "Prediction and Entropy of Printed English," Bell Systems Technical J., vol. 30, pp. 50-64, 1951.

[18] Q. Tan, X. Chai, W. Ng, and D. Lee, "Applying Co-Training to Clickthrough Data for Search Engine Adaptation," Proc. Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2004.

[19] J. Teevan, M.R. Morris, and S. Bush, "Discovering and Using Groups to Improve Personalized Search," Proc. ACM Int'l Conf. Web Search and Data Mining (WSDM), 2009.

[20] E. Voorhees and D. Harman, TREC Experiment and Evaluation in Information Retrieval.MIT Press, 2005.

[21] Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-Enhancing Personalized Web Search," Proc. Int'l Conf. World Wide Web (WWW), 2007.

[22] S. Yokoji, "Kokono Search: A Location Based Search Engine," Proc.Conf. World Wide Web (WWW), 2001