

A Survey on Design and Implementation of Mobile Search Engine Using Ontology

Abhimanyu S. Dutonde, N.Umare

M.Tech Student, Dept. of CSE, AGPCE, RTMNU, Nagpur, India

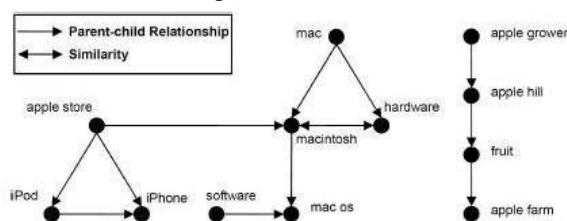
Assistant Professor, Dept. of CSE, AGPCE, RTMNU, Nagpur, India

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: Click through 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 click through 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 personalized mobile search engine (PMSE) 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 click throughs of the query "hotel," PMSE 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.



Example Content Ontology Extracted for the Query "apple".

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

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. To incorporate context information we also take into account the visited physical locations of users in the PMSE. GPS locations play an important role in mobile web search.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

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 responsible for receiving the user's requests, submitting the requests to the PMSE server, displaying the returned results, and collecting his/her click throughs 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 re-ranking 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. Here the same content or location concept may have different degrees of importance to different users and different queries. We introduce the notion of content and location entropies to measure the amount of content and location information associated with a query, to measure how much the user is interested in the content 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.

II. RELATED WORK

Click through data have been used in determining the user's 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 centered at the specified latitude-longitude pair and retrieves documents containing location information within the search circle.

Inefficient 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. In a probabilistic topic-based framework a location-sensitive domain information retrieval system is used. 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 mine are.

The main point of this paper is:-

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.

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. PMSE 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 PMSE. To train the user profiles quickly and efficiently, the design forwards user requests to the PMSE server to handle the training and reranking processes.

The 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.

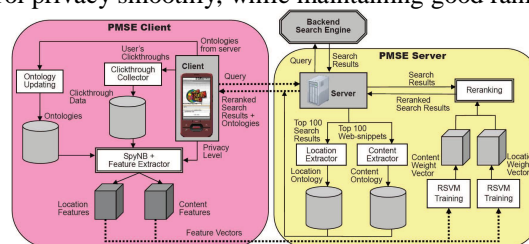


Fig 2:- The general process flow of PMSE.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Fig. 2 shows PMSE's client-server architecture, which meet 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 clickthroughs 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. Moreover, in order to minimize the data transmission between client and server, the PMSE client would only need to submit a query together with the feature vectors to the PMSE server, and the server would automatically return a set of reranked search results according to the preferences stated in the feature vectors.

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. PMSE's design addressed the issues: 1) limited computational power on mobile devices, and 2) data transmission minimization.

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 of future queries are returned to the client.

2. Ontology Update And Clickthrough Collection At PMSE Client.

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. 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, *minDistance* 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. Since the ontologies can be derived online at the PMSE server, an alternative system design is for the user to pass only the clickthrough data to the PMSE server, and to perform both feature extraction and RSVM training on the PMSE server to train the weight vectors for reranking. However, if all clickthroughs are exposed to the PMSE server, the server would know exactly what the user has clicked. 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.

2.1 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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

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 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 a 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 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.

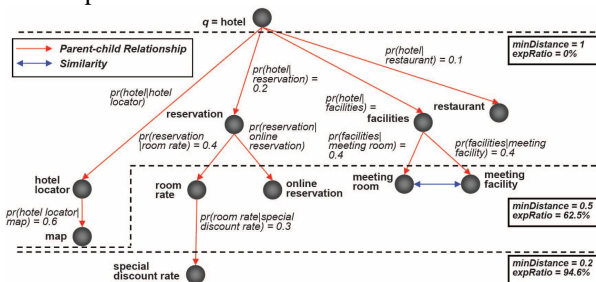


Fig 3:- Ontology for q $\frac{1}{4}$ 00hotel00 with p $\frac{1}{4}$ 0:2; 0:5; 1:0

2.1.1. Location Ontology

The approach for extracting location concepts is different from that for extracting content concepts. The 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 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. 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. Similar to the content ontology, the location ontology together with clickthrough data are used to create feature vectors containing the user location preferences. They will then be transformed into a location weight vector to rank the search results according to the user’s location preferences. 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 entire part of the keywords and key-phrases from the Query patterns documents (QPD) returned for query (UGQ) are extracted with exact matches of the results in location concept.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

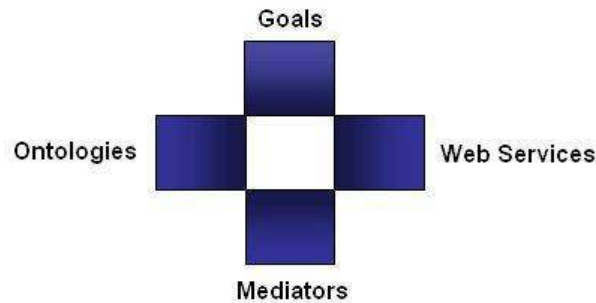


Fig 4:- Model Of The Personalized Search

2.1.2. Content Ontology

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 C_i () is the snippet frequency containing the most related information.

III. 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 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 Search Ranking 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-Line Resources 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 Clickthrough Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 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-Based Clustering of Search Engine Queries," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp. 1505-1518, Nov. 2008. H. Li, Z. Li, W.-C. Lee, and D.L. Lee, "A Probabilistic Topic-Based Ranking Framework for Location-Sensitive Domain Information Retrieval," Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2009.
13. B. Liu, W.S. Lee, P.S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," Proc. Int'l Conf. Machine Learning (ICML), 2002.
14. W. Ng, L. Deng, and D.L. Lee, "Mining User Preference Using Spy Voting for Search Engine Personalization," ACM Trans. Internet Technology, vol. 7, no. 4, article 19, 2007.
15. 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.



ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

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

BIOGRAPHY

Abhimanyu Shrikrishna Dutonde is a Student in Computer Science and Engineering, Abha Gaikwad College of Engineering, Nagpur University. He received Master of Technology (M-Tech) degree in 2016 from RTMNU, Nagpur, MS, India.