International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

# Implementation of Ontology based Personalized Search Filtering (OBPSF) on Smartphone

Devayani Phadke
PG Student,
Department of Computer Engineering,
Sinhgad Technical Education Society's,
Smt. Kashibai Navale college of Engineering
Pune, Maharashtra, India

Jyoti Nandimath
Asst. Prof.
Department of Computer Engineering,
Sinhgad Technical Education Society's,
Smt. Kashibai Navale college of Engineering
Pune, Maharashtra, India

***Abstract*** **- Nowadays in common people, mobile technology and internet are becoming an integral part of daily life. In mobile search the interactions between the users and search engines are limited by the small form factors of the mobile devices. Ontology Based Personalized Search Filtering(OBPSF),it is client-server model. Basically Ontology Based Personalized Search Filtering is personalize mobile search engine. In a ontology based personalized search filtering (OBPSF) on smart phone 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, OBPSF classifies these concepts into content concepts and location concepts. 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. To characterize the diversity of the concepts associated with a query and their relevance's to the user's need, four entropies are introduced to balance the weights between the content and location facets. In our design, the client collects and stores locally the clickthrough data to protect privacy, whereas heavy tasks such as concept extraction, training, and reranking are performed at the OBPSF server. We prototype OBPSF on the Google Android platform.Association rule is used to findout frequent query and laocation pattern.Experimental results show that OBPSF significantly improves the precision comparing to the baseline.**

## 1. INTRODUCTION

Due to advances in mobile technologies and internet access users life become very comfortable and convenient to do the work very intelligently. User's enter queries, click some of the links in the results, click on ads, spend time on website pages, reconstruct their queries, and perform many actions. These interactions can serve as a significant source of information for improving web search result ranking.

In mobile search[1] the interactions between the users and search engines are limited due smartphone devices. As a result, mobile users tend to submit shorter and ambiguous queries compared to their web search counterparts. Also search engine does not gives personalized result, it gives the results globally which are same for all users.

In OBPSF, backend is on one of the commercial search engines, such as Google to perform the main search. The client is responsible for entering the user's requests, submitting the requests to the server, displaying the returned results, and collecting user clickthrough in order to derive user personal preferences.

- In this system[1] client server architecture is used.
- In this system unique characteristics of content and location concepts, and provides a consistent strategy using a client-server architecture to integrate them into a uniform solution for the mobile environment.
- System incorporates a user's physical locations in the personalization process. The influence of a user's GPS locations in personalization. The GPS locations help improve retrieval effectiveness for location queries (i.e., queries that retrieve lots of location information).
- The proposed system is an new approach for personalizing web search results. By mining content and location concepts for different user profiling, it utilizes both the content and location preferences to personalize search results for a user.
- Proposed system facilitates good ranking quality and smooth privacy preserving control.
- Proposed system show that the ontology-based user profiles can successfully capture users' content and location preferences and utilize the preferences to produce relevant results for the users.

## 2. COMPARATIVE WORK

Clickthrough data have been used in determining the users preferences on their search results. Many existing personalized web search systems [3], evaluating user preferences of web search results is crucial for search engine development, deployment, and maintenance. A real-world study of modelling the behavior of web search users to predict web search result preferences. Accurate interpretation and modelling of user behavior has important applications to ranking, web search personalization, click spam detection and other tasks. Key insight of this work to improving robustness of interpreting implicit feedback is to model query-dependent deviations from the expected noisy user behaviour. In this work shows that model of clickthrough interpretation improves prediction accuracy over state-of-the-art clickthrough methods. Generalize this approach to model user behaviour beyond clickthrough, which results in higher preference guess accuracy than models based on clickthrough information alone.

In [4], Geographic web search engines allow users to constrain and order search results in an intuitive manner by focusing a query on a particular geographic area. Geographic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

search technology, also called local search, has recently received key interest from major search engine companies. Academic research in this part has focused primarily on techniques for extracting geographic knowledge from the web. In this paper, study of problem of efficient query processing in scalable geographic search engines. Query processing is a major bottleneck in standard web search engines, and the most important cause for the thousands of machines used by the major engines. Geographic search engine query processing is different in that it requires a combination of text and spatial data processing techniques. They propose several algorithms for efficient query processing in geographic search engines, combine them into an existing web search query processor, and estimate them on large sets of real data and query traces.

In [12], addresses search engine personalization. They present a new approach to mining a user's preferences on the search results from clickthrough data and using the discovered preferences to adapt the search engine's ranking function for improving search quality. They develop a new preference mining technique called SpyNB, which is founded on the practical supposition that the search results clicked on by the user reset the user's preferences, but it does not draw any conclusions about the results that the user did not click on. As such, SpyNB is still applicable even if the user does not follow any order in reading the search results or does not click on all relevant results. Their extensive online experiments demonstrate that SpyNB discovers many more accurate preferences than existing algorithms do. The interactive online experimentation further confirms that SpyNB and our personalization approach are effective in practice. They also show that the efficiency of SpyNB is comparable to existing simple preference mining algorithms.

## 3. PROPOSED ARCHITECTURE

For providing good personalized search, proposed architecture uses Client-server model .

As shown in Fig. 1, proposed system architecture is providing

1) An application is on android smart phone where user is going to do login and enter a query.
2) Server will rerank the results. Backend for this server is global search engine and sends response to the application on an mobile.
3) RSVM is used for reranking.

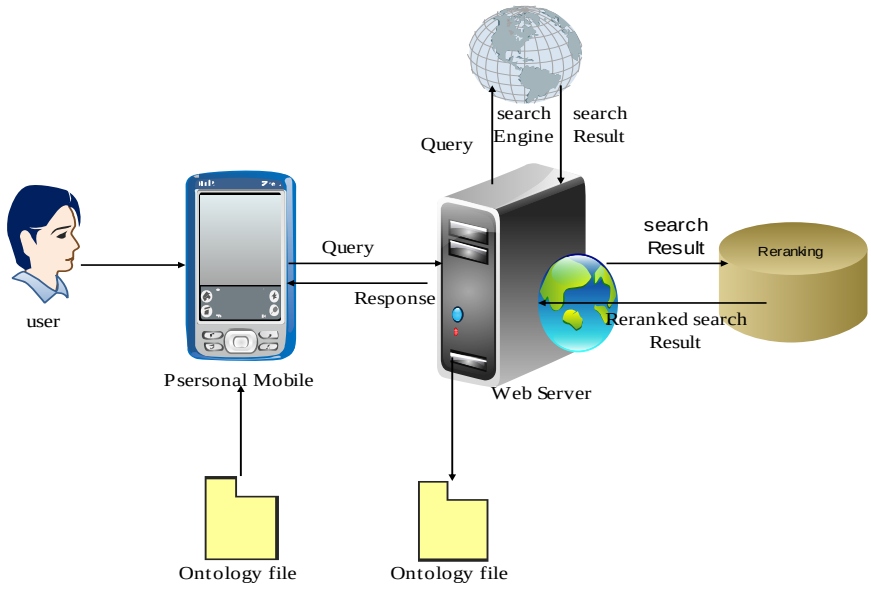



Fig.1: System Architecture

In the client-server architecture, clients are responsible for receiving clickthrogh , showing reranked search results are handled by the clients with some degree of computational power. On the other hand, heavy tasks, such as RSVM training and reranking of search results, are handled by the server. Furthermore, in order to reduce the data transmission between client and server, the client would only need to submit a query together to the 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. Design addressed the issues: 1) limited computational power on mobile devices, and 2) data transmission minimization.

System consists of two major activities[1]:

1. Reranking the search results at server. When a user submits a query on the client, the query together with the feature vectors containing the user's content and location

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

preferences (i.e., filtered ontologies according to the user's privacy setting) are forwarded to the 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 do ontology extraction for its speed. The feature vectors are then used in RSVM training to get 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 at server and clickthrough collection at client.* The ontologies update atte 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 server.When the user clicks on a search result, the clickthrough data jointly with the associated content and location concepts are stored in the clickthrough database on the client. The clickthroughs are stored on the clients, so the server does not know the exact set of documents that the user has clicked on. This design allows user privacy to be preserved to some extent. Two privacy parameters, minDistance and expRatio, are proposed to control the amount of personal preferences exposed to the server. If the user is concerned with 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 server for the personalization. On the other hand, if a user wants more accurate results according to preferences, the privacy level can be set to low so that the server can use the full feature vectors to maximize the personalization effect.

## 4. ASSOCIATION RULE MINING BASED OBPSF AND PMSE

### *4.1 Association Rule Mining (ARM)*

Association Rule Mining (ARM) based OBPSF to explore for go target that is user concept consequences ,practical data mining and association rules method to investigate the association among travelers' profile and their transactions in the data.

Specified a set of user click through is measured as set of of items I= $\{i_1,i_2,i_3,\ldots i_m\}$ and a record of transactions with travel patterns DB= {t1,t2…tn} where ti={Ii1,Ii2,…Iip}, $p \leq m$ and, if $A \subseteq I$ with K = |A| is called a k-itemset or simply an itemset. Let a database D be a multi-set of subsets of I as shown. Each $T \in DB$ supports an itemset $A \subseteq I$ if $A \subseteq T$ holds. An association rule is an expression A => B, where A, B are item sets and $X \cap Y = \emptyset$ holds. Number of transactions T supporting an item A w.r.t DB is called support of A, $Supp\ (A) = |\{T \in DB \mid A \subseteq T\}| / |DB|$. The strength or confidence (c) for an association rule A => B is the ratio of the number of transactions that contain A ∪ B to the number of transactions that contain A, Conf (A → B) =Supp(A∪ B)/ Supp (A).

### *4.2 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 more time occurrence of the use search quert USQ in top documents measure the amount of a particular keyword/phrase Ci with value to UGQ

$$support(c_i)=\frac{sf(c_i)}{n}.|c_i| \qquad (1)$$

where $sf(c_i)$ 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 $c_i$ OBPSF($c_i$) is the snippet frequency containing the most related query patterns in the concepts Ci . After that find the relations among concepts for ontology formulation. Measure the contrast between two concepts which coexist a group on the search results might represent the same topical interest with query travel patterns.

If coexist ($C_i$,$C_j$ ) > $\delta_1$(is a threshold), then $C_i$ and $C_j$ are measured as comparable. If pr($C_j$ | $C_i$ )> $\delta_1$ (is a threshold), score $C_i$ and $C_j$ child.

### *4.3 Location Ontology*

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 OBSF 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

## 5 USER INTEREST PROFILING

OBPSF 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 the content ontology (in Section 4.2) and location ontology (in Section 4.3). 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 OBPSF, we adopt ontologies to model the concept space because they not only can represent concepts but also capture the relationships between concepts

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

*5.1 Diversity of Content and Location Information*

Different queries may be associated with different amount of content and location information. To formally 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. In information theory [14], entropy indicates the uncertainty associated with the information content of a message from the receiver's point of view. In the context of search engine, entropy can be employed in a similar manner to denote the uncertainty associated with the information content of the search results from the user's point of view. 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

$$H_C(q) = -\sum_{i=1}^{k} p(c_i)\log p(c_i) \quad (2)$$

Where k is the number of content concept $C=\{c_1,c_2\ldots\ldots\ldots\ldots\ldots,c_k\}$ extracted,

$|c_i|$ is the number of search result containing the concept $c_i$, $|C|=|c_1|+|c_2|+\ldots\ldots+|c_k|$, $p(c_i)=\frac{|c_i|}{|C|}$

$$H_L(q) = -\sum_{i=1}^{k} p(l_i)\log p(l_i) \quad (3)$$

Where m is the number of content concept $L=\{l_1,l_2\ldots\ldots\ldots\ldots\ldots,l_m\}$ extracted, $|l_i|$ is the number of search result containing the concept $l_i$,$|L|=|l_1|+|l_2|+\ldots\ldots+|l_m|$ , $p(l_i)=\frac{|l_i|}{|L|}$

*5.2 Diversity of User Interest*

There is two another entropy click content entropy and click location entropy $H_{\bar{C}}(q,u)$ and content entropy

$$H_{\bar{C}}(q,u) = -\sum_{i=1}^{t} p(\bar{c}_{iu})\log p(\bar{c}_{iu}) \qquad H_{\bar{L}}(q,u) = -\sum_{i=1}^{v} p(\bar{l}_{iu})\log p(\bar{l}_{iu})$$

.....(4)

Where t is the number of content concepts $\overline{C_u}=\{\bar{c}_{1u}, \bar{c}_{2u}\ldots\ldots\ldots\ldots\ldots,\bar{c}_{tu}\}$ extracted,

$|\bar{c}_{i_u}|$ is the number of times that the content concept $c_i$ has been clicked by

u,$|\bar{C}_u|=|\bar{c}_{1u}|+|\bar{c}_{2u}|+\ldots\ldots+|\bar{c}_{tu}|$, $p(\bar{c}_{i,u})=\frac{|\bar{c}_{i_u}|}{|\bar{c}_u|}$ ,v is the number of location concepts

$\overline{L_u}=\{\bar{l}_{1u}, \bar{l}_{2u}\ldots\ldots\ldots\ldots\ldots,\bar{l}_{vu}\}$ Clicked by u,Where$|\bar{l}_{i_u}|$ m is the number of times that the location concept $l_i$ has been clicked by u, $|\overline{L_u}|=|\bar{l}_{1u}|,|\bar{l}_{2u}|+\ldots\ldots\ldots\ldots\ldots+,|\bar{l}_{vu}|$, and $p(\bar{l}_i,u)=\frac{|\bar{l}_{i_u}|}{|\bar{L}_u|}$

*5.3 Personalization Effectiveness*

To estimate the personalization effectiveness using the extracted content and location concepts with respect to user u as following formulae:

$$e_L(q,u)=\frac{H_L(q)}{H_{\bar{L}}(q,u)}$$

$$e_L(q,u)=\frac{H_L(q)}{H_{\bar{L}}(q,u)} \quad (5)$$

*5.4 User Preferences Extraction and Privacy Preservation*

User preferences based query patterns results are Returned from location concepts and content concepts in the above step to make security in the user profile based results preference ,first mining the results with the set of feature in both content and location concepts related to query patterns alongside through prospect queries to the PMSE server for discover end result reranking. SpyNB it can be adapt with OBPSF to mining the query travel pattern QTP with user preference and after that converse how OBPSF preserve user privacy. The SpyNB method QTP is the positive set of query patterns, U the unlabeled set and QTPN the query predicted negative set obtained from original set.

$$d_I < d_j, \quad \forall l_i \in P \quad l_j \in PN. \quad (6)$$

The OBPSF clients deliver the user's clickthrough data from QTP .It make a feature vector based query pattern based
clickthrough data and the filtered ontology according to the privacy ideals at different expRatio. If it doesn't satisfy it forwards UGQ (User Given Query) to OBPSF server. OBPSF make use of mindistance to pass through a filter the concept in the ontology. Mindistance is defined by $D((c_{i-1},c_k)$ and concept $C_i$ will be prune back and it satisfy the subsequent situation.

$$\frac{D_{(c_{i-1},c_k)}}{D_{(root,c_{i-1})}+D_{(c_{i-1},c_k)}} < \text{minDistance} \quad (7)$$

Where $c_{i-1}$ is the direct parent of $c_i$ and $c_k$ is the leaf node of concept,

The concept entropy $H_C(Uq,p)$ of the user profiles can be compute using the following equation:

$$H_C(Uq,p) = -\sum_{ci\in Uq,p} pr(c_i)logpr(c_i)$$

.....(8)

$$expRatio_{q,p}=\frac{H_C(Uq,p)}{H_C(Uq,0)} \quad (9)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

Ranking SVM is working to learn a modified ranking purpose for examine consequences according to the user satisfied and position preferences. For a given query (UGQ), a set of content concepts and a set of location concepts are extracted on or subsequent the search result as the article features. To take out the concepts calculate similarity and parent-child relations of the concepts in the extracted concept
ontologies are also built-in in the preparation based on the dissimilar types of relations such as Similarity, Ancestor, Descendant and Sibling. The content feature vector $\phi_c(q, d_k)$ with the subsequent equation:

$$\forall c_i \in s_k , \phi_c(q, d_k)[c_i] = \phi_c(q, d_k)[c_i] + 1 \quad (10)$$

For supplementary content concepts Cj that are related to the content concept Ci

$$\forall c_i \in s_k , \phi_c(q, d_k)[c_i] = \phi_c(q, d_k)[c_j] + sim_R(c_i,c_j) + ancestor(c_i,c_j) + descendent(c_i,c_j) + sibling(c_i,c_j) \quad .....(11)$$

Location feature vector $l_i$ is extract from the web snippet and equivalent values are incremented in the location feature vector and incremented location feature vector $\phi_L(q, d_k)$ with the subsequent equation:

$$\forall l_i \in d_k , \phi_L(q, d_k)[l_i] = \phi_L(q, d_k)[l_i] + 1 \quad (12)$$

$$\forall l_i \in d_i , \phi_L(q, d_k)[l_i] = \phi_L(q, d_k)[l_j] + sim_R(l_i,l_j) + ancestor(l_i,l_j) + descendent(l_i,l_j) + sibling(l_i,l_j) \quad (13)$$

Best result optimize the search result in both content and location concepts in OBPSF to combine the two weight vectors and find the final weight vector for user U'0. s ranking. The two weight vectors of query patterns are first normalize previous to the mixture:

$$\overrightarrow{w_{q,u}} = \frac{e_c(q,u)}{e_c(q,u)+e_L(q,u)} . \overrightarrow{w_{C,q.u}} + \frac{e_L(q,u)}{e_c(q,u)+e_L(q,u)} . \overrightarrow{w_{L,q.u}} \quad (14)$$

Let

$$e(q,u) = \frac{e_c(q,u)}{e_c(q,u)+e_L(q,u)} \quad (15)$$

$$\overrightarrow{w_{q,u}} = e(q,u) . \overrightarrow{w_{C,q.u}} + (1 - e(q,u)) . \overrightarrow{w_{L,q.u}} \quad .....(16)$$

will rank the documents in the returned search according to the
following equation ,

$$f(q,d) = \overrightarrow{w_{C,q.u}} . \phi(q,d) .....((17)$$

## 6. QUERY AND QUERY CLASSES

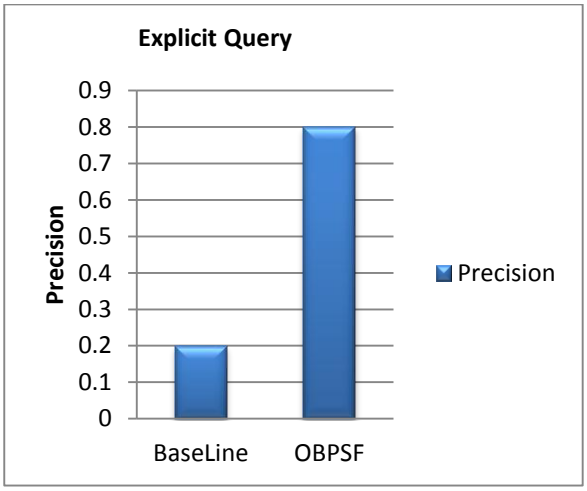
1. Explicit queries. Queries with low degree of ambiguity, i.e., $H_C(q) + H_L(q)$is small.
2. Content queries. Queries with $H_C(q) > H_L(q)$
3. Location queries. Queries with $H_L(q) > H_C(q)$.
4. Ambiguous queries. Queries with high degree of ambiguity, i.e., $H_C(q) + H_L(q)$ is large.

## 7. EXPERMENTAL RESULTS:

We compare all query classes with baseline (PMSE) i.e personalized mobile search engine

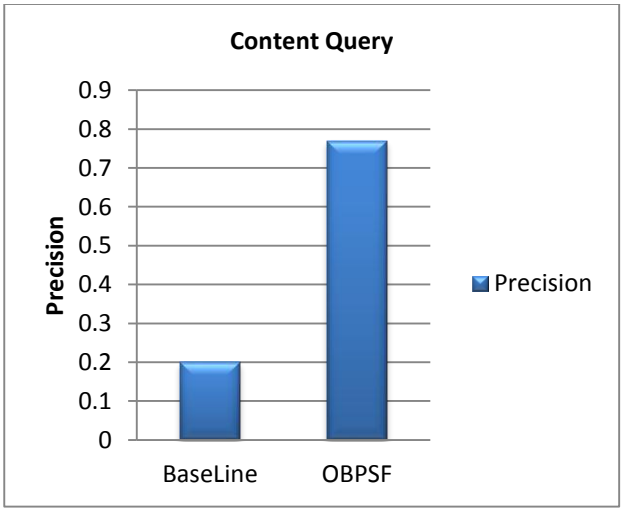
*1..Explicit Query:*

In explicit query baseline performance is 0.2 while OBPSF performance is 0.8.




1. *Content Query:*

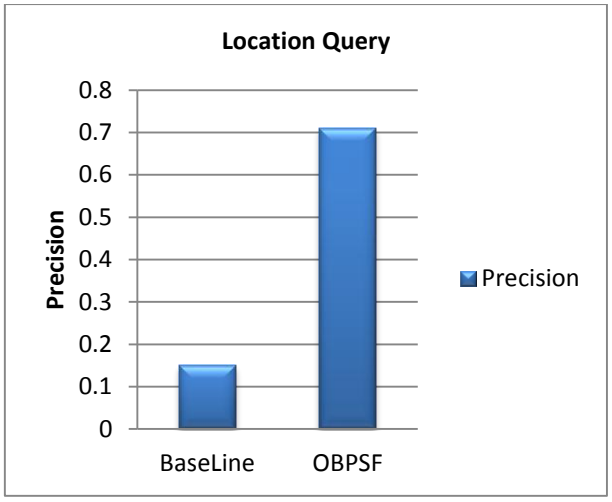
In content query baseline performance is 0.2 while OBPSF performance is 0.78.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 04, April-2017

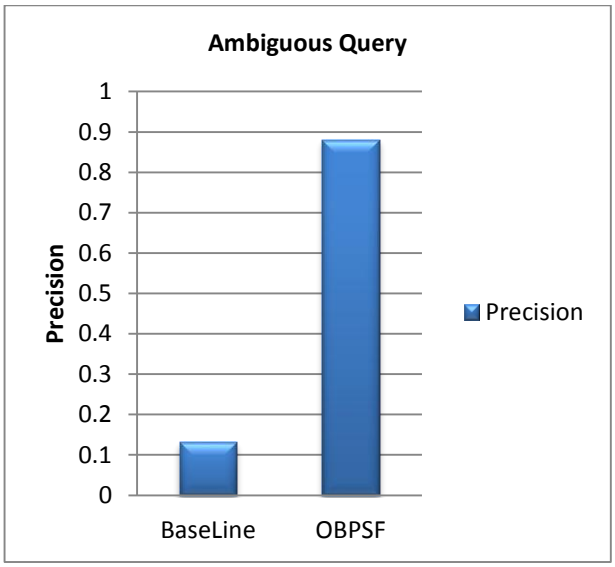
*3.Location Query*:

In location query baseline performance is 0.18 while OBPSF performance is 0.71.




*4.Ambiguous Query:*

In location query baseline performance is 0.14 while OBPSF performance is 0.88.




## 8. CONCLUSION

This Paper proposes system architecture, profile the users interests and personalize the search results according to the users profiles. The other global search engines are not giving the personalised result. For all the search, result is same. System represents different types of concept in different ontologies to include context information revealed by user mobility system also takes into account the visited physical location of users. Main computation task is distributed to the server so that it gives effective performance. Result shows OBPSF performance is better than baseline(PMSE) i.e personalized mobile search engine. It gives user frequent query on top based on location.

## REFERENCES

[1] Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee, PMSE: A Personalized Mobile Search Engine, IEEE Trans .On Knowledge and Data Engineering ,Vol .25 No.4,April 2013

[2] E. Agichtein, E. Brill, and S. Dumais,ImprovingWeb Search Ranking by Incorporating User Behavior Information, Proc. 29th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.

[3] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, Learning User Interaction Models for Predicting Web Search Result Preferences, Proc. Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.

[4] Y.-Y. Chen, T. Suel, and A. Markowetz, Efficient Query Processing in Geographic Web Search Engines, Proc.Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006

[5] 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.

[6] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, Analysis of Geographic Queries in a Search Engine Log, Proc. First Intl Workshop Location and the Web (LocWeb), 2008.

[7] T. Joachims, Optimizing Search Engines Using Clickthrough Data, Proc. ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining, 2002.

[8] K.W.-T. Leung, D.L. Lee, and W.-C. Lee, Personalized Web Search with Location Preferences,Proc. IEEE Intl Conf. Data Mining (ICDE), 2010.

[9] 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.

[10] H. Li, Z. Li, W.-C. Lee, and D.L. Lee, A Probabilistic Topic-Based Ranking Framework for Location-Sensitive Domain Information Retrieval, Proc. Intl ACM SIGIR Conf. Research and De-velopment in Information Retrieval (SIGIR), 2009.

[11] B. Liu, W.S. Lee, P.S. Yu, and X. Li, Partially supervised Classi_cation of Text Documents, Proc. Intl Conf. Machine Learning (ICML), 2002.

[12] 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.

[13] 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 Classi_cation Methods in a Library Setting, J. Information Science,vol. 34, no. 2, pp. 213-230, 2008.

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

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