Customized Mobile Search Engine With Privacy Preservation and Truth Discovery (CMSE)

Jisha Joseph, Namitha Jacob

Abstract— **A customized mobile search engine (CMSE) is proposed that captures the users' preferences in the form of concepts by mining their clickhistory data which is obtained by meta searching ie, searching in three main well known search engines. Location information is very important while searching in mobile phones, CMSE classifies these concepts into content concepts and location concepts. In addition, users' locations (positioned by GPS) are also used to supplement the location concepts in CMSE. The user preferences are collected in an ontology-based, multiple faced user profile, which are used to produce a personalized ranking function for rank adaptation for future search results. To characterize the diversity of the concepts associated with a query and their relevances to the user's need, entropies are used to balance the weights between the content and location concepts. A detailed architecture and design for implementation of CMSE is proposed based on the client server model. In this 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 CMSE server through metasearching in three main search engines. Also, the privacy issue is faced by restricting the information in the user profile exposed to the CMSE server with two privacy parameters. CMSE is prototyped on the Google Android platform. Experimental results show that CMSE significantly improves the precision comparing to the baseline.**

Index Terms— **Clickthrough data, concept, location search, mobile search engine, ontology, customization, user profiling, metasearching, truthdiscovery.**

I. INTRODUCTION

The interaction between the user and search engine is affected by the small size of mobile phones which is a major problem. Therefore, mobile users tend to submit shorter, and more ambiguous queries compared to their web search counterparts that is desktop computers.

Manuscript received April, 2014.

Jisha Joseph, *Department of Computer Science, Calicut University/ KMCT College of Engineering, Kerala, India, +919446251360*

Namitha Jacob, *Department of Information Technology, SRM University,Kerala,India,,+919946449833*

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 to the users' profiles. A practical approach to capturing a user's interests for personalization is to analyze the user's click history data [5]. Accordingly, CMSE will favor results that are concerned with hotel information in India for future queries on "hotel." The introduction of location preferences offers CMSE an additional dimension for capturing a user's interest and an opportunity to enhance search quality for users. To integrate context information revealed by user mobility, the visited physical locations of user is also taken into account in the CMSE. Since this information can be conveniently obtained by GPS devices, it is hence referred to as GPS locations.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 "Kochi ,Kerala," his/her position can be used to personalize the search results to favor information about nearby hotels. Here, we can see that the GPS locations (i.e., "Kochi ,Kerala") help reinforcing the user's location preferences (i.e., "India") derived from a user's search

activities to provide the most relevant results. The proposed framework is capable of combining a user's GPS locations and location preferences into the personalization process. In this paper, a realistic design for CMSE is proposed by adopting the metasearch approach which replies on from some of the major commercial search engines, such as Google, Yahoo, and Bing, to perform an actual search. The searched results are reranked based upon truth discovery,ie common links present in the three search engines are given higher rank. The client is responsible for receiving the user's requests, submitting the requests to the CMSE server, displaying the returned results, and collecting his/her clickthroughs in order to derive his/her personal preferences. The CMSE server, on the other hand, is responsible for handling heavy tasks such as forwarding the requests to a commercial search engines, 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 CMSE clients, thus preserving privacy to the users. CMSE has been prototyped with CMSE clients on the Google Android platform and the CMSE server on a PC server to validate the proposed ideas.

The same content or location concept may have different degrees of importance to different users and different queries. Therfore content and location entropies are introduced to

ISSN: 2278 – 7798

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, click content and location entropies are proposed. Based on these entropies, a method is developed 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 reranked 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. 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. CMSE incorporates a user's physical locations in the personalization process.

One of the challenging issues in CMSE is privacy preseervation, where users send their user profiles along with queries to the CMSE server to obtain personalized search results. CMSE addresses the privacy issue by allowing users to control their privacy levels with a privacy parameters, and expRatio.

A. *Motivation*

Clickthrough data is used in determining the users' preferences in the search results. Many existing personalized web search systems [5], [8] are based click history data to determine users' preferences. Ng et al proposed to combine a spying technique together with a novel voting procedure to determine user preferences. Later, Joachims [6] proposed to mine document preferences from clickthrough data. More recently, Leung et al. introduced an approach to predict users' conceptual preferences from clickthrough 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 "Indian hotels," "museums in China," and "African historical sites." In [7], Gan et al. developed a classifier to classify geo and non-geo queries. Where it was found that 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. Yokoji proposed a location-based search system for web documents. Location information 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.

Later on, Chen et al. [7] studied the problem of efficient query processing in location-based search systems. A query is assigned with a query footprint that specifies the geographical area of interest to the user. More recently, Li et al. [1] 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.

Most existing location-based search systems, such as [2], require users to manually define their location preferences (with latitude-longitude pairs or text form), or to manually prepare a set of locationsensitive topics. Existing works on personalization do not address the issues of privacy preservation.

II. SYSTEM DESIGN

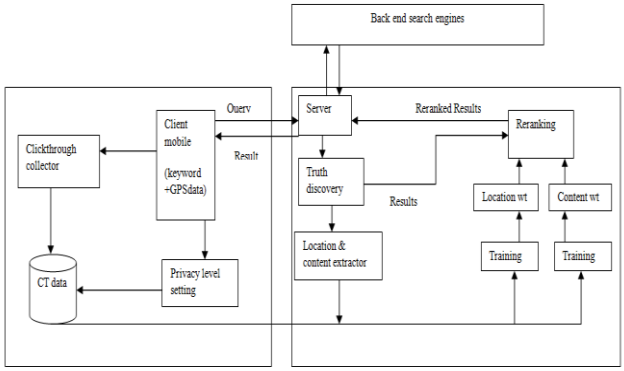


Fig 1:Process diagram

Fig. 1 shows CMSE's client-server architecture, which meets three important requirements. First, clickthrough data, representing precise user preferences on the search results, should be stored on the CMSE clients in order to preserve user privacy. Second, data transmission between client and server should be minimized to ensure fast and efficient processing of the search. Third, computation-intensive tasks, such as RSVM training, should be handled by the CMSE server due to the limited computational power on mobile devices. In the CMSE's client-server architecture, CMSE clients are responsible for storing the user clickthroughs and the ontologies derived from the CMSE server. Simple tasks, such as updating clickthoughs and ontologies, creating feature vectors, and displaying reranked search results are handled by the CMSE clients with limited computational power. On the other hand, heavy tasks, such as RSVM training and reranking of search results, are handled by the CMSE serverAlso to minimize the amount of data transmission between the user and server only the keywords along with the feature vector is sent 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. CMSE's design addressed the issues: 1) limited computational power on mobile devices, and 2) data transmission minimization.

A. Reranking the search results at PMSE server.

When a user submits a query on the CMSE client, the query

ISSN: 2278 – 7798

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 CMSE server, which in turn obtains the search results from the back-end search engines (i.e., Google,Yahoo and Bing).The results from these three SE are compared to find the similar links and they are given higher preference. The content and location concepts are extracted from the search results and organized into ontologies to capture the relationships between the concepts.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. The search results are then reranked according to the weight vectors obtained from the RSVM training.

III. USER PROFILE CREATION

CMSE 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.

Content Concept

The content concept extraction method extracts all the keywords and phrases (excluding the stop words) from the web-snippets arising from query. 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.

Location Concept

A document usually contains a few location concepts, and thus only very few of them occur with the query terms in web-snippets. To overcome this problem tha locations from the document is extracted. All cities are organized as children of there countries thus providing parent child relationship. 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 database, it will be treated as a location concept of d. For example, assume that document d contains the keyword "Kochi." "Kochi" would then be matched against the location ontology. Since "Kochi" is a location in our location ontology, it is treated as a location concept related to d. Moreover, we would explore the predefined location hierarchy, which would identify "Kochi" as a city under the state "Kerala." Thus, the location "/India/Kerala/Kochi/ " is associated with document d. If a concept matches several nodes in the location ontology, all matched locations will be associated with the document. Tthe location concept together with clickthrough data are used to create feature vectors containing the user location preferences similar to what is done in content concept extraction. They will then be transformed into a location weight vector to rank the search results according to the user's location preferences.

IV. PRIVACY PRESERVATION

User's preference can be learned from the past clicks . These search preferences, are to be submitted along with future queries to theCMSE server for search result reranking. Instead of transmitting all the detailed personal preference information to the server, PMSE allows the users to control the amount of personal information exposed.

The preference pairs together with the extracted concepts are used to derive a set of feature vectors on the CMSE client for submission along with future queries to the PMSE server which in turn finds a linear ranking function that best describes the user preferences using RSVM. In our client-server model, the click histories are entirely stored on the CMSE clients . The back-end search engine has no knowledge of a user's click history. Therefore, the user's privacy is ensured. The CMSE server is a trusted server, which would not store all the clickthrough data. It is aware of the user's preferences, but the how much it knows is controlled by the privacy settings set by the client. The client stores the user's click histories and has control on the privacy setting. It would create a feature vector based on its click history data and the filtered ontology according to the privacy settings at different expRatio. The feature vector is then forwarded to the CMSE server for the personalization. Thus, the CMSE server only knows about the filtered concepts that the client prefers in the form of a feature vector. To control the amount of personal information exposed out of users' mobile devices, CMSE filters the ontologies according to the user's privacy level setting,which is set by user The lower the privacy level (the more information being provided to theCMSE server for the personalization), the better the personalization results. Thus, there is a tradeoff between them. If the user is concerned with his/her own privacy, the privacy level can be set to high to provide only limited personal information to the PMSE server. Otherwise, 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 CMSE server can use the full user profile for the personalization process, and provide better results.

V. GPS INFORMATION

GPS locations are important information that can be useful in personalizing the search results. For example, a user may use his/her mobile phone to find nearby hotels providing a particular dish. Thus, CMSE incorporates the GPS locations into the personalization process by tracking the visited locations. This function is provided by the GPS part integrated in the software. We believe that users are possibly interested in locations where they have visited. Thus, our goal is to integrate the factor of GPS locations in the user profile those which they have visited earlier

VI. EXPERIMENTAL EVALUATION

ISSN: 2278 – 7798

google url	: yahurl :http	Bing url	common
://www.mango.com/	://mango.com	://mango.com	://en.wikipedia.org/wiki/Mango ://mango.com
://shop.mango.com/GB/mango	://shop.mango.com/index.faces?tienda=she	://shop.mango.com/index.faces?tienda=she	://shop.mango.com/IN/mango ://dictionary.reference.com/browse/mango
://shop.mango.com/home.faces?state=she_006_IN	://en.wikipedia.org/wiki/Mango	://en.wikipedia.org/wiki/Mango	full values ://www.mango.com/ ://shop.mango.com/GB/mango ://shop.mango.com/home.faces?state=she_006_IN
://www.flymango.com/	://shop.mango.com/IN/mango	://shop.mango.com/IN/mango	://www.flymango.com/
://en.wikipedia.org/wiki/Mango	://www.hort.purdue.edu/newcrop/morton/mango_ars.html	://www.hort.purdue.edu/newcrop/morton/mango_ars.html	://en.wikipedia.org/wiki/Mango_(clothing) ://shop.mango.com/index.faces?tienda=she ://www.hort.purdue.edu/newcrop/morton/mango_ars.html ://shop.mango.com/index.faces?tienda=she ://www.hort.purdue.edu/newcrop/morton/mango_ars.html
://en.wikipedia.org/wiki/Mango_(clothing)	://dictionary.reference.com/browse/mango	://dictionary.reference.com/browse/mango	

Fig 2: comparison of results before and after metasearch

The user submits the keyword and sets the privacy level. The output after metasearch and truth discovery can be examined by comparing the results of the three search engines and the proposed final result. The results after ranking, i.e., when a user clicks on a particular link, then the whole results will be ranked.

Before ranking	After ranking
://en.wikipedia.org/wiki/Jam	://en.wikipedia.org/wiki/Fruit_preserves
://www.iitb.ac.in/~pge/2k12/jam	://en.wikipedia.org/wiki/Jam
://en.wikipedia.org/wiki/Fruit_preserves	://en.wikipedia.org/wiki/Jam_session
://www.jamwithchrome.com/	://www.kilbanfoods.com/contact.html
://gate.iitk.ac.in/jam/	://kozhikode.infoisinfo.co.in/card/sweetvallery_bakery/2609521
://jamrestaurant.com/	://www.jamwithchrome.com/
://gate.iitd.ac.in/jam	://gate.iitk.ac.in/jam/
://jam.iitkgp.ac.in	://gate.iitd.ac.in/jam
://gate.iitd.ac.in/jam	://www.iitb.ac.in/~pge/2k12/jam
://jam.iitkgp.ac.in	://en.wikipedia.org/wiki/Fruit_preserves
://locations.westernunion.com/in/kerala/calicut/4358c8e8e83b4f2eb46bb73a7db59bb1	://locations.westernunion.com/in/kerala/calicut/4358c8e8e83b4f2eb46bb73a7db59bb1
://pincodes.info/in/Kerala/Kozhikode/Malayamma/Malayamma-Road-Kattangal-Kerala/	://pincodes.info/in/Kerala/Kozhikode/Malayamma/Malayamma-Road-Kattangal-Kerala/
://www.nitc.ac.in/	://www.nitc.ac.in/
://www.kilbanfoods.com/contact.html	://en.wikipedia.org/wiki/Kozhikode
://en.wikipedia.org/wiki/Kozhikode	://en.wikipedia.org/wiki/A.K._Premajam
://en.wikipedia.org/wiki/A.K._Premajam	://www.facebook.com/Echoes.IIMK
://kozhikode.infoisinfo.co.in/card/sweetvallery_bakery/2609521	://jamjoomsouk.com/
://www.facebook.com/Echoes.IIMK	
://jamjoomsouk.com/	

Fig 3: comparison of results before and after customizing

The above results show that after applying this application, the results will be ranked according to the users preference. To evaluate the ranking quality of PMSE, we compare the e_ectiveness of three alternative PMSE implementations, labeled as PMSE(content), PMSE(location), and PMSE(mfacets), against a baseline approach and the SpyNB method. PMSE (location) employs only the location-based features in customization, while PMSE (content) uses only the content-based features in customization. PMSE (m-facets) employs both the content-based and location based features, weighted by their customization effectiveness. The baseline composes of the ranked results returned by the back-end search engine (i.e., Google).The effectiveness of different customization methods is evaluated using average relevant ranks, which is the average rank of the documents rated as Relevant. Figure below shows the ARRs of different classes of queries grouped by location and content entropy. First, the ARR for the baseline method is low on explicit queries, which is expected to have good performance because they are very focused. Second, it has high ARR for ambiguous queries, showing that the general purpose search engines by design do not handle the ambiguity of queries well. Finally, the ARRs for content and location queries are slightly lower than the ARR on ambiguous queries. The observations show that the commercial search engines perform well for explicit queries, but suffer in various degrees for vague queries. It is observed that PMSE (location) method performs the best on location queries from figure, lowering the ARR from 26.28 to 15.11 (43 percent decrease in ARR). It also perform well on ambiguous queries, lowering the ARR from 30.65 to 19.77 (35 percent decrease in ARR). The performance of PMSE (location) method is not good for explicit and content queries, because only a limited amount of location information exists in them. On the other hand, PMSE (content) method performs the best on content queries, lowering the ARR from 25.77 to 10.85 (58 percent decrease in ARR). The ARR is also significantly lowered for ambiguous queries from 30.65 to 15.11 (51 percent decrease in ARR). PMSE (content) performs _ne on location queries, because location queries also contain a certain amount of content information. It lowered the ARR of location queries from 26.28 to 15.51 (41 percent decrease in ARR). Finally, as expected, the precisions are the best for explicit queries. However, the improvement is not as significant as in other query classes because the baseline method already performs reasonably well for explicit queries. PMSE (content) lowered the ARR of explicit queries from 22.86 to 16.20 (29 percent decrease in ARR). It is also observed that PMSE (content) performs better than PMSE (location) in general, showing that content information is an important factor in the customization. Although PMSE (location) by itself does not perform as well as PMSE (content), it does provide additional improvement for customization. By employing both the content-based and location-based features in the customization (i.e., PMSE (m-facets)), the ARRs is further lowered. The ARRs of explicit, content, location, and ambiguous queries using PMSE (m-facets) method can greatly reduced to 14.59, 13.19, 9.09, and 11.85, showing that both of the content and location information are useful in the customization process.

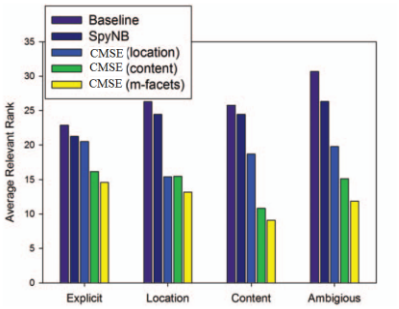
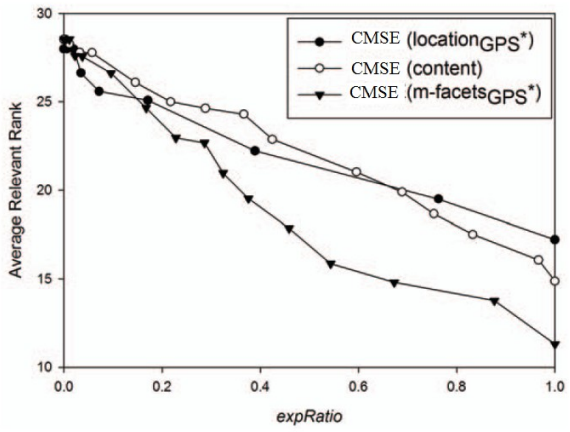


Fig 4:ARR for query classes

On the other hand, the expRatio of PMSE (locationGPS), which employs location ontology only, decreases uniformly from 1 to nearly zero when minDistance increases from 0 to 0.3.



ISSN: 2278 – 7798

Fig 5: Exposed ratio vs average value

The heights of the trees in the location ontology are mostly less than 0.3. It is observed 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-facetsGPS), which employs both content and location ontologies, decreases faster than PMSE (content), but slower than PMSE (locationGPS). The expRatio of PMSE (m-facetsGPS) decreases uniformly from 1 to nearly zero when minDistance increases from 0 to 0.6. Then studied the relationships between the privacy parameters and the ranking quality of the search results for PMSE (content), PMSE (locationGPS), and PMSE (m-facetsGPS). Thus, the ARR of PMSE (content) increases uniformly when minDistance increases from 0 to 0.7. Similarly, the ARR of PMSE (locationGPS) increases uniformly when minDistance increases from 0 to 0.3, and the ARR of PMSE (m-facetsGPS) increases uniformly when minDistance increases from 0 to 0.6. Finally, figure below shows the relationships between expRatio and ARR for di_erent PMSE methods. The more privacy information exposed (i.e., higher expRatio), the better the ranking quality. It is observed that PMSEs privacy parameters produce a smooth increase in ARR when minDistance increases, and a smooth decrease in ARR when expRatio decreases, and thus provide a smooth privacy settings for the users.

VII. CONCLUSIONS

Customized Mobile Search Engine is proposed to extract and learn a users content and location preferences based on the users clickthrough. To adapt to the user mobility, the users GPS locations is incorporated in the customization process.The GPS locations help to improve retrieval effectiveness, especially for location queries. Exposed ratio is the parameter used to address privacy issues in CMSE by allowing users to control the amount of personal information exposed to the CMSE server. The privacy parameters facilitate smooth control of privacy exposure while maintaining good ranking quality. For future work, investigate methods to exploit regular travel patterns and query patterns from the GPS and clickthrough data to further enhance the customization effectiveness of CMSE.

REFERENCES

1. Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee, *Personalized Mobile Search Engine*
2. Agichtein, E. Brill, and S. "Improving Web SearchRanking by Incorporating User Behavior Information," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.
3. T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDDTFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
4. W. Ng, L. Deng, and D.L. Lee, "Mining user preference using spy voting"
5. Q. Tan, X. Chai, W. Ng, and D. Lee, "Applying cotraining to click through data for search engine adaptation".
6. J. Teevan, M.R. Morris, and S. Bush" Discovering and Using Groups to Improve Personalized Search"
7. Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-Enhancing Personalized Web Search"
8. Nat'l geospatial, http://earth-info.nga.mil/, 2012.

Jisha Joseph Received Btech degree in Information Technology and MTech in Computer Science from Calicut University at KMCT College of Engineering. Also she has worked as a lecturer in the same college. Her research interest include Mobile computing, artificial intelligence and web search.

Namitha Jacob Received Btech degree in Information Technology at Rajagiri Institute of Science and Tchnology under MG university and MTech in Information Technology from SRM. Also she is currently working as Assistant Professor KMCT College of Engineering. Her research interest inat clude Mobile computing, artificial intelligence, web search ,information retrieval and search engines.

ISSN: 2278 – 7798