Volume 5, Issue 2, February 2015 **ISSN: 2277 128X**

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

PSBS: Preference & Sensor Based Search for Smartphone Users

Bobby Jasuja[*], Dhaval Doshi, Himey Jain, Nikhil Tiwari
Department of Information Technology
Medicaps Institute of Science and Technology, Indore, India

Abstract-Mobile search is a developing branch for data retrieval and analysis, centred on the use of mobile platforms. Compared to web based search there is huge scope of advancement in mobile based search through the use of sensors like GPS. They help in generating more user specific and relevant results. Preference and Sensor based mobile search (PSBS) takes into account users' choice by analysing their past history data .It also takes into account location based search through the use of GPS. GPS gives the accurate latitude and longitude location of smartphone over the entire globe. PSBS groups these detail into two parts-past data history (content based preference) and location based preference. The user choice are arranged in an ontology-based fashion. This arrangement is used to obtain a preferential ranking function for ranking of future search results. To recognize the diversity of the results returned through a query and their impact on the user's need, proportion ratio is calculated to balance the retrieved results between the location and content factor. In our architecture, the client' smartphone locally stores the historical data to protect privacy concerns, whereas heavy tasks such as data extraction, analyses and training and re-organizing of retrieved results are performed at the PSBS server.

Keywords-web-snippet, preference vector, proportioned ratio, re-rank, ontology, RSVM training

I. INTRODUCTION

One of the problem in mobile based search is that the interaction between the smartphone users and commercial search engines like Google,yahoo,Bing etc. are limited by few factors.Therefore, smartphone users tend to submit shorter and ambiguous queries compared to their counterparts (web based search). But in order to return relevant results to the users, mobile based search engines must be able to collect and analyse user's interests and re-rank the search results according to the users' past data history.A feasible approach to capturing a user's interests is to analyse the users past data history.

Leung et al had developed a web based search engine method based on user's preferences and proved that it is better than methods that are based on traditional methods. However, most of the previous researches assumed that all factors belong to the same categories. Thinking about the need for different types of factors, we present in this paper a Preference & Sensor Based Search (PSBS) which represents different types of factors according to ontologies. In particular, recognizing the importance of user location in mobile search, we separate factors into location factor and past data history factor.For example, a smartphone user who is going to visit USA may fire the query "hotel" and click on the search results related to hotels in USA.From the clicked results of the query "hotel"PSBS can learn the user's data factor (e.g. "room cost" and "services") and location factor("USA").Accordingly, PSBS will re-rank results that are related with hotel information in USA for queries on "hotel" in future.

Introduction of location factor offers PSBS an additional dimension for capturing a user's interest and an effective way to enhance search results for users.To incorporate location based search using smartphone mobility,we also make use of the current physical location of users in PSBS.This information can be easily obtained by GPS devices.Thus,GPS locations play a crucial role in mobile web search.For example- if smartphone user(searching for hotel information) is currently located in "San Diego,California" his/her position can be used to re-rank the search results to favour information related to nearby hotels.Here,we can see that the GPS location(i.e."San Diego, California")help reinforcing the user's place choice(i.e."USA") derived from a user's clicked activities in order to provide the most relevant results. Our idea is to form framework capable of combining a user's current location (given by GPS) and location choice into the re-ranking process.

To the best of our knowledge,this paper is the first to stipulate a re-ranking framework that utilizes a commercial web engine at the back end and the PSBS server to validate the proposed ideas.

We also understand that content or location factor might have different degrees of relevance to different users at different point of time.To classify the diversity of the factors associated with a query fired by user and their relevance'sto the user's current need, we have introduced the idea of past data history and location proportion ratios to measure the amount of content and location data associated with a query.Based on these proportionRatios, we hackneyed amethod to calculate the effectiveness for a particular query fired by a user,which is then used to obtain a more user proportioned combination between the content and location factors. The retrieved search results are re-ranked in accordance with the users past history data and location factors before returning to the client.

Table 1 past data history for the query “Hotel”

Doc	Search Results	Content Preference C_i	GPS Location	Location Preference L_i
D_1	Hotels.com	room cost, services	San Diego	International
D_2	US Hotels Guide	room rate, facilities	New York	USA,Chicago
D_3	Los AngelesBooking.com	casino, discount	New York	USA,LasVegas

Users send their profiles along with queries to the PSBS server to obtain preferred search results. PSBS handles the privacy issue by allowing users to regulate their privacy levels with two privacy limits-low and high.This model significantly overcomes the drawback of current strategies which are either content or location based only.

II. LITERATURE & SURVEY

There are many existing personalized web search systems based on past data history to determine user’s preferences.Joachim’s proposed a method to select document preferences from past data history.

Search queries can be categorized as non-geo data or location (i.e. geo) queries.Gann et al. developed a classifier to identify geo and non-geo queries. Several algorithms are employed to rank the search results as a combination of a textual and a geographic score.

Yokoji proposed a location(geographical)based search system for web data. When a client fires a location query containing latitude and longitude position, the system creates a search circle centred at the specified latitude and longitude position. Example of location based queries(geo) are “hotels in New York”, “museums in England” and “Indian culture” etc. Location data was extracted from the web documents and transformed into latitude and longitude positions and system retrieves documents containing location information within the search circle. Recently, Li et al. proposed a probabilistic topic-based design for location-sensitive domain information extraction.Instead of modelling locations in latitude and longitude pairs,this model assumes that users can be interested in a set of location sensitive topics.It recognizes the geographical diversity in topic distributions and models them using probabilistic Gaussian process classification.

Existing system can find web documents based on the distance between locations that are described in web documents and a location specified by a user. It consists of three modules-

- ❖ Artificial Intelligence-It contains a robot that gathers web documents from the Internet.
- ❖ Parser(Web extractor)-It extracts address strings from web documents and associates latitude-longitude information to the original document.
- ❖ Retrieval module-This module can retrieve location-related web pages that are overlooked by conventional keyword-based search engines.

Some of the problems in existing system are-

- ❖ Keyword-based search engines overlooked at least 25% of location-related web documents.
- ❖ Existing works on preference based search do not address issues related privacy preservation. When exposing the user personal data to server there is lot of possibilities to be leaked to vulnerabilities.
- ❖ Existing systemrequire users to manually define their location choices(with latitude-longitude pairs or text form) or to manually describe a set of location-sensitive topics.
- ❖ One of the problem in existing mobile search engine is that the communication between the users and search engines is limited by few factors of the mobile equipment and its environment.As a result, mobile users are forced to submit short and ambiguous queries compared to their web search counterparts.

The advantage of our proposed system over above enlightened systems is that-

PSBS profiles user’s content and location preferences in the ontology based fashion, which are automatically learned from the past data history and GPS location without requiring any extra manual efforts from the user.

III. PROPOSED WORK

The main postulates of this paper are as follows-:

- ❖ This paper studies the unique features of content and location factors on mobile based search and provides a logical strategy using a client-server model to integrate them into a coherent solution for the mobile platform.
- ❖ The proposed preference and sensor based mobile search engine is an unique approach for personalizing web search results obtained from commercial search engines like Google,Bing etc. By mining past data history and location factors for user profiling,PSBS utilizes both the content and location factors to optimize search results for a user.
- ❖ PSBS includes a current physical locations of a user in the personalization process.Use of GPS locations in general helps to improve retrieval data effectiveness for location based queries.
- ❖ We have proposed realistic architecture for PSBS.Our design incorporates server-client architecture in which client queries are re-directed to PSBS server for analysing,training and re-ranking the retrieved results from commercial

search engines quickly.We have adopted a meta search approach where client(smartphone device) has the responsibility for receiving the user's requests and submitting the same to the PSBS server, displaying the retrieved results, and maintaining his/her past data history in order to derive his/her preferences.On the other hand, PSBS server is responsible for performing heavy tasks like forwarding requests to a commercial search engine, training and re ranking of retrieved search results before they are displayed to the client.

Advantages of PSBS are as follows-

- ❖ Proposed system profiles both-user's content(past data history) and location preferences in the ontology-based fashion(Retrieves good result).
- ❖ It addresses privacy issue by controlling the amount of information in the client's user profile being exposed to the server(Security advantage).
- ❖ Avoids overlooking of irrelevant document.

IV. DESIGN

A. Architecture

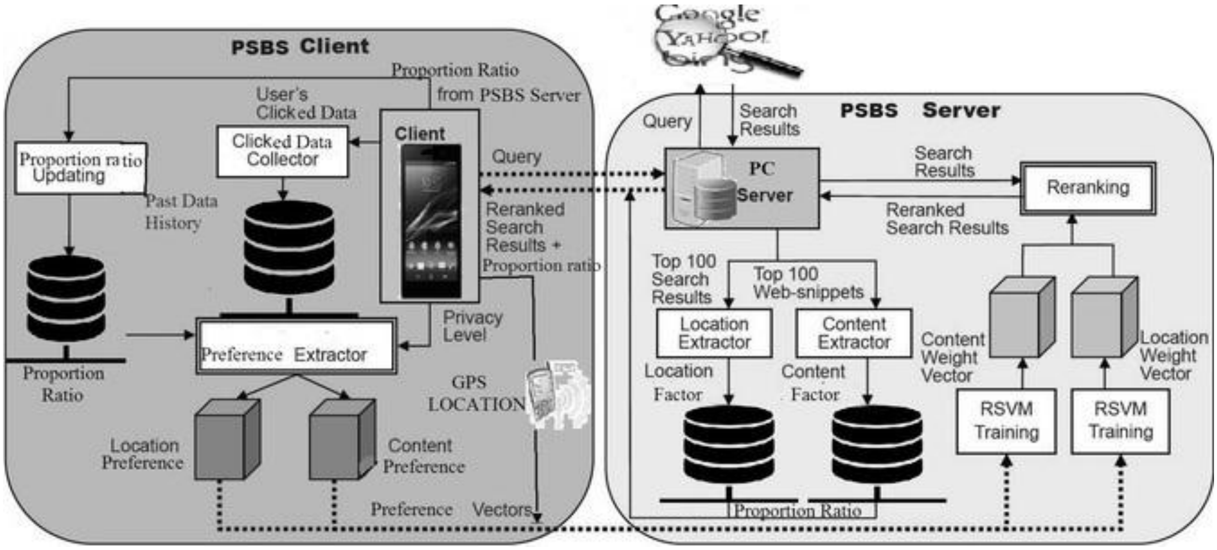

Fig 1-PSBS ArchitecturePSMS consists of following activities:

1. Re ranking the search results at PSMS server- When a user fires a query using PSMS client(smartphone device),the query together with the preference vector containing the user's location and content preferences (based on the privacy setting set by user)are forwarded to the PSMS server, which in turn obtains the unaltered search results from the back-end commercial search engine (i.e. Google, yahoo or Bing etc.). The content and location factors are extracted from the search results and proportion ratio are created to capture the relationships between the factors. The preference vectors forwarded by the client site are then used in RSVM re ranking process to obtain a location weight vector and content weight vector(representing user choice based on content and location preferences) for reranking search results.Search results are then re-ranked in accordance to the location and content weight vectors obtained using the RSVM training.Finally,the re-sequenced search results and the extracted proportion ratio for the re-ranking of future queries is returned to the client.

2. Proportion ratio update and past data history storing at PSMS client- Proportion ratio returned from the PSMS server contain the factor space that models the relationships between the factors analysed from the search results. They are stored in the past data history database on the client. When user clicks on a re-ranked search result, the past data history together with the content and location preference are stored in the past data database on the client. The past data history is stored on the PSMS clients, so the PSMS server has no idea of the exact set of details that the user has searched in past. This kind of design allows user's privacy to be kept intact up to a certain degree. Two privacy limits-low and high are proposed to control the amount of personal preferences exposed to the PSMS server. If the user is worried about its privacy issues, privacy level can be set high so that only limited personal data will be included in the preference vectors and passed to the PSMS server. On the other hand, if a smartphone user wants more precise results according to his/her preferences, privacy levels can be set low so that the PSMS server can use the full preference vectors to fully utilize the personal details.

Reasons for choosing client-server architecture-

- ❖ Computation-intensive tasks like RSVM selection and re-ranking of retrieved search results are handled by the PSMS server due to the limited processing power of mobile devices whereas easy tasks like updating past data history and proportion ratio, creating preference vectors and displaying reranked search results are done by the PSMS clients.
- ❖ Data transmission between client and PSMS server needs to be minimized to ensure fast, cheaper and efficient processing of the search.Therefore,PSMS client only submits a query along with the preference vectors to the PSMS server and the server has to provide a set of re-ranked search results in accordance with the preferences

stated in the preference vectors.Only the essential data (i.e. query, preference vectors, proportion ratio and re-ranked search results) are transmitted between the PSMS client and server during the personalization process.

- Past data history representing precise client preferences on the search results is stored on the PSMS clients in order to preserve user's privacy.

B. Content Factor

Content concept extraction method first gathers all the keywords and phrases from the web-snippets arising from query q. If a keyword or a phrase occurs frequently in the web-snippets arising from the query q, we treat it as an important factor related to the query,as it coexists in close relation with the query in the top 100 retrieved web documents from backend search engines. The following formula, is employed to measure the importance of a particular keyword/phrase K_i with respect to the query q:

$$\text{Support }(Ki_{)\,=}\ [\text{sf }(K_i)*|K_i|]/n$$

where sf (K_i) is the snippet frequency of the keyword/phrase K_i (i.e., the number of web-snippets containing K_i), n is the number of web-snippets returned and $|K_i|$ is the number of terms in the keyword/phrase K_i. If the support of a keyword/phrase K_i is higher than the threshold th, we treat K_i as an important factor for query q.

C. Location Factor

Most of the geographical locations over the entire globe have become known (global) over the internet. Thus, it is not an important task to gather data related to them (at least their name and relation with each other) in the form of a set. City,state, region name, and country names and relationship among these locations can be organized in the form of cities as children under their states, all the states as children under their country/region.

First, all the keywords and key phrases from the web documents returned for query q are collected. If a keyword/phrase in a retrieved web document d matches a field in our predefined location list, it will be treated as a location factor.

For example, let us assume that document d contains the keyword "New Delhi." "New Delhi" would then be matched against the predefined location list. Since "New Delhi" is a part of our location list, it will be treated as a location factor.We can also use the predefined location hierarchy, which would identify "New Delhi" as a capital of "India". If a concept matches several members of a location list, all matched locations will be linked with the document and query q.

D. Proportion Ratios

Content entropy HC (q) and location entropy HL (q) are used to measure the uncertainty associated with the content and location information respectively of the search results

$$HC\,(q) = -\sum_{i=1}^{n} p\,(K_i)\log p\,(K_i)$$

Where n is the number of content concepts K= K1, K_2…Kn; extracted, $|K_i|$is the number of search results containing the content concept c_i;

$$|K|=|K_1|+|K_2|+|K_3|+...|K_n|$$

$p\,(K_i)=|K_i|/|K|$

And $$HL\,(q) = -\sum_{j=1}^{m} p\,(l_j)\log p\,(l_j)$$

m is the number of location concepts L=l_1, l_2, ...l_m ; extracted,$|l_j|$ is the number of search results containing the location concept l_j,

$|L|=|l_1|+|l_2|+|l_3|+..|l_k|$

$p\,(l_j)=|l_j|/|L|$

1. Explicit queries-Queries with low degree of ambiguity i.e. HC (q), HL (q) is small.
2. Content based queries-Queries with HC (q)>HL (q).
3. Location based queries-Queries with HL (q) > HC (q).
4. Ambiguous queries-Queries with high degree of ambiguity i.e. HC (q), HL (q) is large.

E. Content and Location Preferences

We also introduce content preference entropy $HC_c(q)$ and location preference entropy $HL_c(q)$ to indicate, the diversity of a user's interest on the content and location information respectively returned from a query. The entropy equations for click preference and location preference are similar to HC (q) and HL (q) respectively but only the user's clicked pages are considered in the formula. Since the click entropies (past data history) reflects the user's response to the retrieved (re ranked)results, they can be used as an indication of the diversity of the user's interests for future queries.

Low click entropies- HCc(q) + $HL_c(q)$ is small.
Content seeking- $HC_c(q)>HL_c(q)$.
Location seeking- $HL_c(q) >HC_c(q)$.
High click entropies- $HC_c(q) + HL_c(q)$ is large.

F. Privacy Control

The back-end search engine has no knowledge of a user's past data history. Hence, the user's privacy is ensured. The

PSMS server is a trusted server, as it would not store all the past data history. It is aware of the user's choices, but the extendit knows is controlled by the privacy settings set through two levels by the client. The PSMS client stores the users past data history. It would generate preference vector based on its past data history and proportion ratio according to the privacy settings. The preference vector is then forwarded to the PSMS server for the re ranking the retrieved search results. Privacy settings are controlled using this preference vector.

There is a close relationship between privacy and personalization effectiveness. The lower the privacy level (more the personal information is provided to the PSMS server for reranking), the better the personalization is.If the user is worried about his/her own privacy, the privacy settings can be set to high level to provide only limited access of personal information to the PSMS server although then the personalization effect will be less effective.

All the location and content preference are passed along with GPS location in the form of preference vector, when privacy level are set low.But when privacy level is set high, only limited content and location information are passed depending on the $HC_c(q)$ and $HL_c(q)$ value of all keywords/key phrases of past data history. They have to be lower than certain value (say 0.5) in order to be passed in preference vector. This is done to ensure that a more general form of data with less privacy risk is passed on to server but at the same this data being part of clicked history means that it is acceptable/relevant to user requirement. Also GPS location are not used for this privacy level as they can extend the risk/concerns associated with the privacy exposure to the PSMS server. Thus, the PSMS server only knows about the filtered concepts that the client prefers in the form of a preference vector. For example, a user who searches for medical information may not want to reveal the specific medicine or disease she/he is trying to diagnose.

G. Re ranking Search Results and GPS Data

RSVM aims at finding a sequence of web pages by linearly ranking them all according to number of document preference pairs/vectors contained by a web page.GPS location of mobile device as well as all the possible nearest location to GPS location are also sent with preference vector with highest possible $HL_c(q)$ value for them. Nearest possible location to current GPS location are found using predefined location ontologies known to PSMS system.

V. EVALUATION

- Retrieved webpages for a query with high content/location factor indicate that they have a high degree of ambiguity and applying location/content preferences on the search results will help in retrieving user relevant information.
- On the other hand, when the content/location entropy is low (meaning that the retrieved webpages for a query are already very precise and user centric).Thus, personalization can do very little in further improving the precision of the result.
- For click location/content preferences, when the click content/location preferences are higher, effectiveness of preference entropies worsens because high click content/location preferences indicate that the user is clicking on the search results with a lot of uncertainty in his mind with regards to data he/she is searching for. When the user's interests are very scattered, it is difficult to find out the user's actual requirement.
- On the other hand, if the click content/ location preferences are low, the preference entropies effectiveness would be within acceptable limitsbecause the user has a focus on certain specific topic in the retrieved results (only a small number of content/location preferences have been clicked by the user).

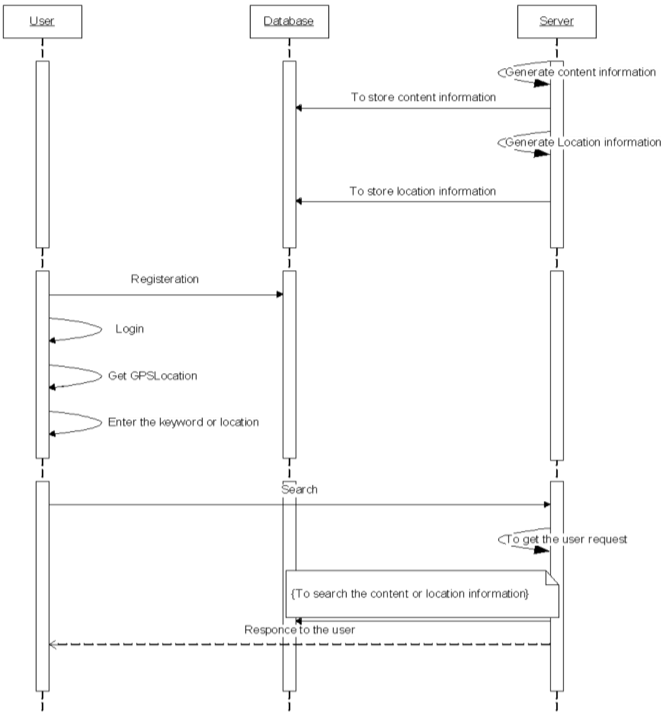
User
Database
Server
Generate content information
To store content information
Generate Location information
To store location information
Registeration
Login
Get GPSLocation
Enter the keyword or location
Search
To get the user request
{To search the content or location information}
Responce to the user

Fig 2.-Sequence Diagram for PSMS System

- One of the major limitation of PSMS is that it is highly dependent on back end search engine being used for a particular query. This is because on changing the search engine for future queries, retrieved web pages may change and the preference vector associated with the previous search engine may not give effective preferences in the re ranked results displayed to the client.

Based on the above reasoning, we propose to estimate the personalization effectiveness using the extracted content and location concepts with respect to user u as follows:

$$EC(q) = HC(q)/HC_c(q)$$

$$EL(q) = HL(q)/HL_c(q)$$

Queries with higher EC (q) and EL (q) would yield better user centric result.

VI. CONCLUSION

We proposed PSBS to collect and analyse user’s request based on location preferences and users past search history. To incorporate mobility feature, we used GPS locations in the personalization process. PS results provided effective and improved retrieval, especially for location based queries. We also used two privacy levels-high and low, to address privacy concerns in PSBS.This allowed users to regulate the personal information flow being exposed to the PSBS server. The privacy levels facilitated griped control over security while maintaining high ranking quality. In future work, we will try to investigate new methods to calculate query patterns from GPS and past history data in order to further enhance the effectiveness of PSBS.

ACKNOWLEDGMENT

We would like to express our gratitude and a special thanks to the editors and the reviewers for giving very insightful and encouraging comments.

REFERENCES

[1] E. Bill’s. Dumas and E. Agichtein - “Learning User Interaction Models for Predicting Web based Search Result Preference” and “Improving Web based Search Ranking by Including User Behaviour Information” 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2006.

[2] Fang Liu, Yu. C. and WeiyiMeng-”Personalized Web search for improving retrieval effectiveness” Knowledge and Data Engineering, IEEE (Volume: 16 and Issue: 1)

[3] Joachim’s-“Optimizing Search Engines Using Click through Data” ACM SIGKDD Conference on Knowledge Discovery

[4] K. T. Leung-“Personalized Web Search with Location Preferences” IEEE Conference on Data Mining (ICDE-2010).

[5] S. Yokoji, “Kokono Search- Location Based Search Engine” International Conference on World Wide Web (WWW 2001).

[6] http://www.computer.org/publications/dlib

[7] http://en.wikipedia.org/wiki/Personalized_search .