Knowledge Based History And Location Based User Profile For Web Search Efficiency

Pranali Yewale
Department of Information Technology,
Caymets SCOE,
Pune, India.
pranaliyewale@gmail.com

Prof. Jyoti Pingalkar
Department of Information Technology,
Caymets SCOE,
Pune, India.
jyoti_pingalkar@rediffmail.com

***Abstract*—Internet contains massive information. Retrieving the desired information is aided by large amount of data available at search engines. Search engines are the mostly used sources of information over the internet, may it be textual data, images or locations which are difficult to obtain with just ambiguous queries. Now a day the web researchers are doing lots of experiments to improve the quality of the search engine. My proposed system keeps forth a novel idea and proposes the personalized method to improve the results being displayed over the search engines or information retrieved over the internet. In the proposed system, users search history or recently searched data is maintained as a source to guess the user's interest. The system makes use of the user history along with the domain knowledge. Once the user has searched something and enters the same query second time, the previous search results will be combined with the current search results and displayed to the user. And the very important thing apart from the history and domain knowledge is the location specific searches. Consider a user searches jaguar being in south Africa, then the location is appended to the search query , and as we know south Africa is famous for its dense forests and wildlife, the expected output here is description about the jaguar animal, not about the jaguar vehicle or jaguar hind ware. Thus the location helps us to trace the user search interests. The proposed system gives an immense help in retrieving relevant and faster results.**

***K*eywords– Personalized Search, Location specific search, Domain Knowledge, Ambiguous Queries, Search history.**

I. INTRODUCTION

Internet has been the one and only popular source of information or we can say the quickest and accessible at finger tips any time anywhere. But many times user gets wrong data while searching or he has to search on search engine to get his required data. Two different users can enter an ambiguous query in different locations with different intentions, but mostly both the users get the same results on searching over internet. For example if there are two users Alice and Bob. And Alice wants to search the query for jaguar car and the Bob wants to search the news about jaguar animal. They both enters same query in the two different search box i.e. jaguar. But these two different users get same search results giving them same results about the jaguar animal. Alice gets her expected results from search engine in first stroke, but Bob doesn't get his expectations in a single stroke. He has to search on two or three pages to get the information about jaguar animal.

This scenario motivated us to work out on this domain and develop such a system that will not only trace user interests but will also trace user location specific queries.

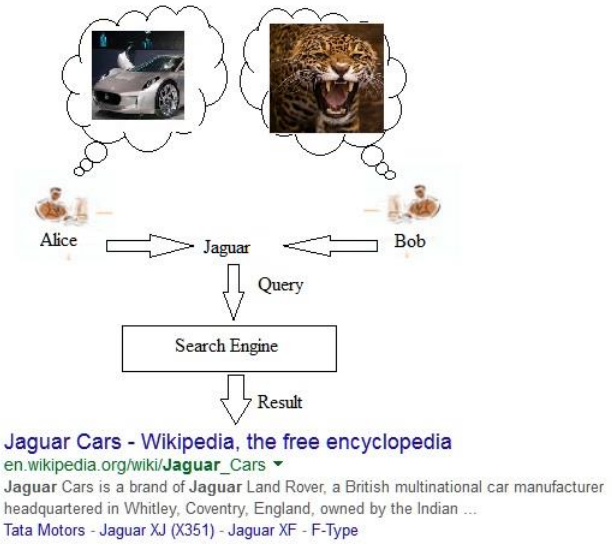

Fig. 1. Ambiguous Query results

II. LITERATURE REVIEW

There are so many researches have done on personalized search engine framework. It uses profile of user which is designed by evaluating user history or browsing application etc. For finding user interests system inspect its history.

K Wai-Ting Leung [1], highlighted most advanced methodology which comprised of a real time preferences such as location content and concept content which majorly provides highly relevant data for user queries. Due to the importance of location information in mobile search, PMSE classifies these concepts into content concepts and location concepts. In addition, users locations (positioned by GPS) are used to supplement the location concepts in PMSE. To characterize the diversity of the concepts associated with a query and their relevance to the user's need, four entropies are introduced to balance the weights between the content and location facets. Based on the client-server model, [1] also presented a detailed architecture and design for implementation of PMSE. But the major drawback is the user interests are not given any priority. User is not given any facility to provide user feedback for ambiguous results.

Zheng Lu, Xiaokang Yang, [2] presented an approach to predict the user search goals using just the user preference as its source of prediction. User clicked URLs are parsed to find the terms from the clicked terms, and then these clicked terms are mapped with unclicked URLs to find if any other URL is relevant to the user interested Data. But the major drawback of the system is that the preferences are only stored for single session. Once the search session is ended, the user clicks provided previously is not valid and need to provide the feedback newly for the further sessions.

Chunyan Liang [3] describe that various users shares various information by requesting same query, for solving their problem author introduces personalize search engine. Three approaches, k-Nearest method, Rocchio method and Support Vector Machines have been used in [3] to build user profile to present an individual user's preference and found that k-Nearest method is better than others in terms of its efficiency and robustness.

K. W. T. Leung et aI. [4] added location preferences into personalize search engine. Here author used two different concepts for location and contents and combine them to produce more accurate results. For maintaining relation between location and contents ontology is used. It make an ontology-based, multi-facet (OMF) user profile design basis of location and history of user search.

O. Shafiq et aI. [5] gives a search model that combines content based, community based and evidences based methods. On internet hundred terabytes of data being uploaded or downloaded per second. Large data is spending for searching web pages, news, blogs and social networking. Due to this massive amount of data it creates difficulties for user to search relevant data. The author addresses a model which produces results by evaluating preference and user interest.

Xuwei Pan et aI. [6] gives context based personalized method. At various situations personalize web search gives effective outcomes according to user's choice. It uses three main concepts modeling user context, semantic ranking of web resources and similarity matching between web resources and user context.

Micro Speretta et aI., [7] designed a framework designs search goals and retrieve relevant information from results. It creates user profile based on that goal. Depending on these profiles results are re ranked. This paper examines the order of re ranked result before applying user profile and after applying user profile. After inspection it found that newly re-ranked results has 34 percent relevancy than older one. Here user collection of user interest is done in a quick way and search data availability causes penalization. It does not force user to install or make use of proxy server for collection of user information.

Fang Liu et aI. [8] Figure out design of search engines are not makes any effect on user interests. So they derive method which depends upon history of user search. It also uses user profile for re-ranking of results. It combines search history, and profile of user to represent user's search interests and to remove non related words from query.

A. Two LLSF-based Algorithms:

It gives two matrices first is m-by-p document-category matrix and second is m-by-n document-term matrix. It consists of Linear Least Squares Fit (LLSF) technique which computes a p-by-n category-term matrix. For solving this problem it used the Singular Value Decomposition (SVD) which is categorized into the multiplication of three matrices,

$$M = DC^T * U * \sum^+ * V^T \text{ , Equation (1)}$$

It also calculates another method known as pseudo-LLSF (PLLSF) which is responsible for reducing the dimensions of Matrices. Initially the space is replaced by a k dimensional space. After the replacements, modified matrices can be calculated using the formula,

$$M = DC^T * U_k * \sum_k^+ * V_k^T$$

The dimension reduction technique is used to remove noise in the original document-term matrix. It is Latent Semantic Indexing method (LSI) which used successfully in IR.

B. Rocchio-based Algorithm

It is a feedback relevance method. Rocchio adopted in text categorization:

$$M(i, j) = \frac{1}{N_i} \sum_{k=1}^{m} DT(k, j) * DC(k, i)$$

Where M is the matrix which represents the user profile, Ni is the number of documents that are related to the i-th category, m is the number of documents in DT, DT(k,j) is the weight of the jth term in the kth document, DC(k,i) and is a binary value.

C. kNN:

On user profile, the k-Nearest Neighbor (kNN) method not depends. It depends upon the similarity between a user query and each category directly from DT and DC. To employ a hard classification approach in which we classify queries into categories and train a ranking model for each category be a straightforward approach to query dependent ranking. We think, however, that to achieve high performance it could be very difficult with this approach. When looking at the data, we observe that to draw clear boundaries between the queries in different categories, it is hard. Let us take example as the TREC 2004 web track data. In the dataset there are in total 225 queries, which have been manually classified into three categories: topic distillation, named page finding, and homepage finding. The queries are also associated with documents as well as the relevance labels of these documents. We represent the queries in a 27-dimensional query feature space and define features of queries. By using Principal Component Analysis (PCA) we next reduce the space to 2-dimensions.

D. Concept based query searching:

In this paper for user search queries the input queries are clustered to get a categorized output. The set of user click-through data uses this technique which is extracted from the Web-crawlers for building concept-based user profiles automatically:

STEP 1. For all possible pairs of query nodes using Equation, obtain the similarity scores in G.

STEP 2. Pair of most similar query nodes (qi, qi) then merges which does not contain the same query from different users. Assume that both query nodes qi and qi are connected to a concept node c with weight wi and wj, a new link is created between c and (qi, qi) with weight w = wi + wj.

STEP 3. For all possible pairs of concept nodes using Equation (1) obtain the similarity scores in G.

STEP 4. Having highest similarity score Merge the pair of concept nodes (ci, cj). Assume that both concept nodes ci and cj are connected to a query node q with weight wi and wj, a new link is created between q and (ci, cj) with weight w = wi + wj.

STEP 5. Repeat Steps 1-4 unless termination is reached.

Limitation of this technique is that, it searches only the queries which are inside the cluster. New queries which are other than clusters are not giving expected output.

III. Personalised Web Search Using Browsing History And Domain Knowledge With Location Preference

In proposed system we are improving the quality of search engine by suggesting some relevant pages of user interest. Here we profilise the user who is searching the information over the internet. The history and the domain knowledge of the user navigation are used to store the categories information. User entered query is transferred to the query optimizer and history of the system. Furthermore it is used with previous user profile to improve it. The domain knowledge is also used with them to produce new enhanced user query. This enhanced query and user interests are sent to the search engine. The obtained results from the search engine are then re ranked according to the user interests or feedback. Then the ultimate improved search results are displayed to the user. There are three main models in our project.

A. A Model of Domain Knowledge:

Domain knowledge is being used in the proposed system to improve the results obtained from search engine. Here we dont use the Domain Knowledge Directory (DKD) to store all the user search history. All results are extracted by using the web crawler from the internet itself. After the extracted results from the crawler are obtained matching keywords are again collected to form a vocabulary.

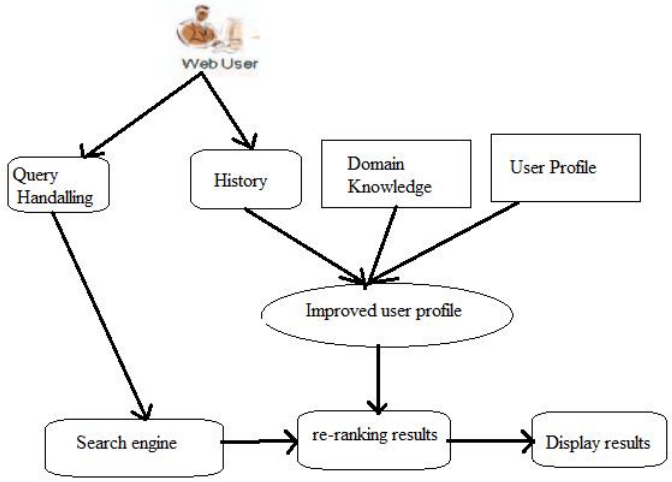

Fig. 2. System Architecture

B. A Model for user profile:

The web page is classified on basis of threshold value. This threshold value is generated by using the API. We are using the API and the DKD library together to form an enhanced user profile.

C. A Model of user interested results:(enhanced profile):

User feedback or search history is equally very important to domain knowledge or internet available information. The user interested information in matched with knowledge information over the internet. User selects his relevant information or URLS to add in the list. Then this URL is scanned throughout the list to find the relevant matches of result. Then in between that founded URLs only those URL are displayed which having the value above the average value.

D. Working of Proposed System:

For an instance consider a user Ui Enters Query Qi at time instance t, the user entered query is first passed on to the history checking procedure to check whether same query was been entered by the user previously or not.
If the same query was searched previously by the user, the user is suggested the results from the browsing history available in the history. If the information available in the history is not of user interest and the user is interested in searching some other information with same query, the user location is tracked and the query is enhanced with the below information and then passed to domain knowledge to get the interested information:
Qi=User I enters Query Q.
Hi= user is not interested in history Hi available in his / her browsing history.
Li= User Location is tracked before sending query to browser.

Now the users enhanced profile is created as follows;
Advanced Query:
AQi = Qi+(!Hi)+RQ(Li).

Now user gets the information from the browser according to the advanced query passed to Domain Knowledge.
The results being displayed have the facility to store the feedback. So user can select particular URLs or Results from all the displayed results and accordingly the selected results will be added to user history for further reference. When the user enters the Query Qi, the original results are processed and the relevant terms are extracted from the description of the results obtained and these terms help user to obtain the interested information from the internet with the help of TRIO (History, Location and Domain Knowledge).

IV. Procedure For Proposed System

The entire system is a concept of making the user comfortable and satisfies him/her with the expected results after searching over mobile browsers.

The below procedure is followed to achieve the desired output:

1. The user Ui initially places a Query Qi to the mobile browser.

2. The User profile is then enhanced by checking the user history for the same query.

3. If the user history does not have any records for the same query, then the user profile is further enhanced by finding user location.

4. Consider the user Location is Li, so the new user profile formed will be Ui at Location Li.

5. Now the final enhanced user profile and the query are concatenated to find the exact location specific results by making use of domain knowledge and user profile.

V. Expected Result

The proposed system should provide all the mobile users an efficient searching mechanism that works as an intelligent browsing system that provides you the search results with corresponding to user interest, Location and Domain knowledge. The expected output for the proposed system is that the entered query returns the user expected and relevant search results that have the concern of user preference, user location as well as domain knowledge.

VI. Conclusion

Here for improving search result, we have used the domain knowledge and the user profile along with history. This project builds the concept of enhanced user profile to make some suggestion to the user search. The output of the enhanced user profile is better than that of traditional user profile. This improves the performance of overall searching mechanism.

References

[1] Kenneth Wai-Ting Leung, Dik Lun Lee, Wang-Chien Lee, PMSE: A Personalized Mobile Search Engine, IEEE transactions on knowledge and data engineering, vol. 25, no. 4, april 2013.

[2] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, A New Algorithm for Inferring User Search Goals with Feedback Sessions, IEEE transactions on knowledge and data engineering, vol. 25, no. 3, march 2013.

[3] C Liang, "User Profile for Personalized Web Search", International Conference on Fuzzy Systems and Knowledge Discovery, pp. 1847-1850, 2011.

[4] K.W.T. Leung, D.L. Lee and Wang-Chien Lee, "Personalized Web search with location preferences", IEEE 26th International Conference on Data Engineering, pp. 70 I - 712, 2010.

[5] O. Shafiq, R. Alhajj and 1. G. Rokne, "Community Aware Personalized Web search", International Conference on Advances in Social Networks Analysis and Mining, pp. 3351 - 355,2010.

[6] X Pan, Z Wang and X Gu, "Context-Based Adaptive Personalized Web Search for Improving Information Retrieval Effectiveness", International Conference on Wireless Communications, Networking and Mobile1Computing, pp. 5427 - 5430, 2007.

[7] M Speretta and S Gauch, "Personalized Search Based on User Search Histories", Proceeding Of International Conference on Web Intelligence, pp. 622-628, 2005.

[8] F Liu, C Yu and W Meng, "Personalized Web Search for Improving Retrieval Effectiveness", IEEE Transactions On Knowledge And Data Engineering, pp. 28-40, Volume 16, 2004.