International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307

Privacy Based Personalized Web Search

K. Anithagayathri[1], Dr. K. Palanivel[2]
[1]M.Phil, Research Scholar, AVC College (Autonomous), Mayiladuthurai, Tamil Nadu, India
[2]Associate Professor, AVC College (Autonomous), Mayiladuthurai, Tamil Nadu, India

ABSTRACT

Searching is one of the common task performed on the Internet. Search engines are the basic tool of the internet, from where one can collect related information and searched according to the specified keyword given by the user. The information on the web is growing dramatically. The user has to spend more time in the web in order to find the particular information they are interested in. Existing web search engines do not consider particular needs of user and serve each user equally. Moreover it also takes more time in searching a pertinent content. Privacy based Personalized Web Search Engine is considered as a promising solution to handle these problems, since different search results can be provided depending upon the choice and information needs of users. It exploits user information and search context to learning in which sense a query refer. In order to perform Personalized Web search it is important to model User's interest. User profiles are constructed to model user's need based on his/her web usage data. This Enhanced User Profile will help the user to retrieve concentrated information. It can be used for suggesting good web pages to the user based on his/her search query and background knowledge. And also implement the pruning algorithm to eliminate the user details from anonymous person for preserving the key word privacy. User privacy can be provided in the form of protection like without compromising the personalized search quality.

Keywords : Personalized Web Search, User Profile, Search Query, User's Interest, Privacy Risk

I. INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. As the name proposes, this is information gathered by mining the web. It makes utilization of automated apparatuses to reveal and extricate data from servers and web2 reports, and it permits organizations to get to both organized and unstructured information from browser activities, server logs, website and link structure, page content and different sources.

The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

Web structure mining can also have another direction -- discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure (schema) of Web pages, this would be good for navigation purpose and make it possible to compare/integrate Web page schemes. This type of structure mining will facilitate introducing database techniques for accessing

information in Web pages by providing a reference schema.Web mining can be divided into three different types – Web usage mining, Web content mining and Web structure mining.

Current web search engines are built to serve all users, independent of the special needs of any individual user. With the exponential growth of the available information on the World Wide Web, a traditional search engine, even if based on sophisticated document indexing algorithms, has difficulty meeting efficiency and effectiveness performance demanded by users searching for relevant information. Personalization of web search is to carry out retrieval for each user incorporating his/her interests. Personalized web search differs from generic web search, which returns identical results to all users for identical queries, regardless of varied user interests and information needs. When queries are issued to search engine, most return the same results to users. In fact, the vast majority of queries to search engines are short and ambiguous. Different users may have completely different information needs and goals when using precisely the same query. Personalized web search can be achieved by checking content similarity between web pages and user profiles. Some work has represented user interests with topical categories. User's topical interests are either explicitly specified by users themselves, or can be automatically learned by classifying implicit user data. Search results are filtered or re-ranked by checking the similarity of topics between search results and user profile.

II. RELATED WORK

Cockburn, et.al,…[1] investigated the enhancements to the temporal scheme that aim to improve the effectiveness of direct page access mechanisms. This work is motivated and informed by the findings of the studies reported earlier in the paper. Current web browsers support many tools for web page revisitation, including the Back and Forward buttons, bookmarks, and pop-up history lists. We believe that these tools can be beneficially integrated into one revisitation resource. We are also investigating the use of simple visualization techniques to aid web page identification within these tools. We have constructed several interfaces, collectively called Web View, to experiment with these concepts. The common features of all WebView prototypes are as follows. First, they interact with unaltered versions of Netscape Navigator. Figure 5, for example, shows an implementation of WebView, in which the complete temporal list of pages is added to the drop-down menu associated with the Back menu. WebView automatically adapts to navigational actions made in Netscape, and Netscape responds to user actions in WebView. Second, they provide a visual representation of the complete temporal list of previously visited pages, therefore integrating History and the temporal behaviour of the Back and Forward buttons. Third, they provide zooming thumbnail representations of all pages visited in the browser. The small thumbnails, which are automatically captured whenever a new page is displayed in the browser, expand whenever the user points the cursor over them. Fourth, the thumbnails are enhanced with additional identification cues including 'dog ears' that encode information about the number of visits to a page and whether the page is bookmarked.

J.Teevan, et.al,…[2] have many ramifications for search engine design, and potentially for browsers and search toolbars. Re-finding, or searching for previously found information, represents a significant fraction of user behavior. Traditionally, search engines have focused on returning search results without consideration of the user's past query history, but the results of the log study suggest it might be a good idea for them to do otherwise. Although finding and re-finding tasks may require different strategies, tools will need to seamlessly support both activities. As shown in the log analysis, people often clicked on both old and new results during the same search.

Because people repeat queries so frequently, search engines should assist their users by providing a means of keeping a record of individual users' search histories, perhaps via software installed on the user's own machine. A number of search history designs have been explored. The results of the log study indicate it is important to account for individual differences in how people repeat queries. For example, different users made use of repeat queries at different rates, and may benefit from having a different amount of screen real estate devoted to displaying their search history. This form of shortcut could be highly effective for many in terms of rapid access to information. While a user may simultaneously have a finding and re-finding intent when searching, satisfying both needs may be in conflict. Finding new information means being returned the best new information, while re-finding means being returned the previously viewed information. The searcher's ability to re-find was hampered. It is important to consider how the two search modalities can be reconciled so a user can interact with new and previously seen information. As Teevan has previously proposed, before information is allowed to change, it is important to understand which aspects of it that a person has already interacted with are memorable. Despite the personal nature of re-finding, it is possible that repeat queries from one user could benefit another.

C. E. Kulkarni, et.al,…[3] analyze the system to saw each page along with its associated topic-phrases, and rated the relevance of the phrases to the page on a scale of 1 to 7 (1: entirely irrelevant; 7: entirely relevant). Based on their ratings, four users were then chosen to participate in a semi-structured interview Users rated topic-phrases for 36 pages in all (users were shown less than 5 pages when the program couldn't find pages with non-intersecting topic-phrases). The median rating obtained was 4.5 (SD= 2.42). First, for some pages (such as technical documentation), page-content was remarkably more important than the page-context, so topic-phrases that captured context were seen as irrelevant. Second, participants gave low ratings when they expected a particular topic-phrase and did not find an exact match (e.g. P7 was looking for "Cricket", but we found the page was about "Fantasy Cricket"; P5 was disappointed we found "Event Log" as a topic-phrase, but not the exact API call he was using for event logging). Lastly, we note that resulted in "entirely relevant" topic-phrases had at least one topic-phrase that was not present in the target page but was present in the page-context. Systems that ignore page-context would miss these completely.

T. Deng, et.al,…[4] has to mimic the retrieval and recall mechanism of human brain memory discovered by the life scientists, we develop a context-based information refinding approach. We build a link between the information and its previous accessed context instance, represented as a multidimensional vector. A context memory contains a large volume of associated context instances organized in clusters. To mimic the characteristic of human brain memory that some prominent events can last very long or even a life long, while the majority will gradually degrade and disappear in the end, we bind each context instance with a dynamic life-cycle decay policy. Memory reinforcement is also incorporated by adjusting the decay speeds of context instances. Based on the context memory, we build a recall-based query-by context model to support users' information refinding queries. We explore the use of context cluster and association to efficiently process context-based refinding queries. A system called Refiner has been implemented to assist users refinding Web pages or files based on their previous accessed context including time, place, and concurrent activity. Refiner provides a small translucent window at the right corner of the computer screen, by double clicking which users can annotate contextual information (time, place, and concurrent activity) for any opened file or viewed

Web page. Refiner also implements an IE browser plug-in and a desktop-based right-button pop-up plug-in to facilitate users in context annotation. When users want to refind his/her accessed files or Web pages, they only need to indicate associated access context to refiner, which will return the matching results.

T. Deng, et.al,…[5] evaluates our approach by conducting two sets of experiments with synthetic data and through a 6-week user study. In the synthetic data experiment, we simulate the contextual search method You Pivot and use it as a baseline. The comparisons of revisit precision and recall show our method outperforms You Pivot, as our method can adapt to the user's revisitation habit. In the user study, the revisit recall rate of our revisit prototype is over 90%. On average, 16.25 seconds are needed to complete a web revisitation task with our method and 38.66 seconds with popular methods like bookmark, browse history, search engines, etc. The experimental results show that our prototype provides a complementary effective solution in facilitating user's web revisitation through contextual keywords. To evaluate our approach, we implemented a prototype called ReVisit and conducted two sets of experiments with synthetic data and through a user study.

As the amount of information on the web continuously grows, it has become increasingly difficult for web search engines to find information that satisfies users' individual needs. Personalized search is a promising way to improve search quality by customizing search results for people with different information goals. Most of them could be categorized into two general approaches: Re-ranking query results returned by search engines locally using personal information; or sending personal information and queries together to the search engine. A good personalization algorithm relies on rich user profiles and web corpus. Therefore, most of personalized search services online like Google Personalized Search and Yahoo! My Web adopts the second approach to tailor results on the server by analyzing collected personal information, e.g. personal interests, and search histories. The web search engine has long become the most important portal for ordinary people looking for useful information on the web. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts. Personalized web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query.

Limitations :

- The existing methods do not take into account the customization of privacy requirements.
- All the sensitive topics are detected using an absolute metric called surprised based on the information theory.

III. PROPOSED FRAMEWORK

The propose a privacy-preserving personalized web search framework UPS(User Privacy-Preserving Search) which can generalize profiles for each query according to user-specified privacy requirements. Relying on the definition of two conflicting metrics, namely personalization utility and privacy risk, for hierarchical user profile, we formulate the problem of privacy-preserving personalized search as δ-Risk profile generation, with its NP-hardness proved. The develop two simple but effective generalization algorithms, Greedy DP and Greedy IL algorithm to support runtime profiling. While the former tries to

maximize the discriminating power(DP) the latter attempts to minimize the information loss(IL).By exploiting a number of heuristics, GreedyIL out-per forms Greedy Significantly. In this proposed work provide an inexpensive mechanism for the client to decide whether to personalize a query in UPS. This decision can be made before each runtime profiling to enhance the stability of the profile. Our extensive experiments demonstrate the efficiency and effectiveness of our UPS framework. A greedy algorithm is an algorithmic paradigm that follows the problem solving heuristic of making the locally optimal choice at each stage[1] with the intent of finding a global optimum. In many problems, a greedy strategy does not usually produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a globally optimal solution in a reasonable amount of time.The algorithm spends less of its time selecting the most similar profile for the current cluster in each iteration. Assume the original user profile set has n profiles and the algorithm generates clusters which mean it has l iterations. Since it searches at most n profiles at each iteration, the overall time complexity is O(ln).The step of our approach is to compute a representative for all member profiles in each profile cluster. The cluster centroid or union (based on the original user profiles not the augmented ones) is computed and used as the group representative. We generated a set of user profiles from the AOL search query log2. The log was collected over three months. Each entry of the query log contains user ID, query terms, query time, item rank, clicked URL's. We only extracted the query terms for each user id and they are assigned with the same initial weight value 1.0. In the pre-processing, terms that do not exist in WordNet dictionary are considered as typos and invalid terms and removed. We also discard stop words which are language words that have no significance meaning in a keyword based search system. We use the stop words set from Lucene3 which includes words like "a", "an", "and", "are" etc. Each query session keyword vector is generated from query session which is represented as follows query session=(input query,(clicked URLs/Page)+) where clicked URLs are those URLs which user clicked in the search results of the input query before submitting another query ; „+" indicates only those sessions are considered which have at least one clicked Page associated with the input query.

Based on the hardness of the problem, we use a greedy algorithm. In the beginning, a user profile is randomly selected as the seed of a new cluster. The closest user profile is continuously selected and combined with the seed until the cluster satisfies p-link ability or the size of the cluster |Gi| satisfies the constraint $|Gi| \geq |U|avgp$. At next step, a user profile with the longest distance to the previous seed is selected as the seed of the new cluster.

result ← ∅
C ← ∅
seed ← a randomly picked user profile from S
while |S| > 0 do
seed ← the furthest user profile(with the min similarity value) to seed
while C does NOT satisfy p-linkability AND |S|>0 do
add the closest user profile (with the max similarity value) to C
end while
if C does satisfy p-linkability then
result ← result ∪ C;
C ← ∅
end if
end while
for each user profile in C do
assign it to the closest cluster end for.

The proposed work is shown in fig 1.

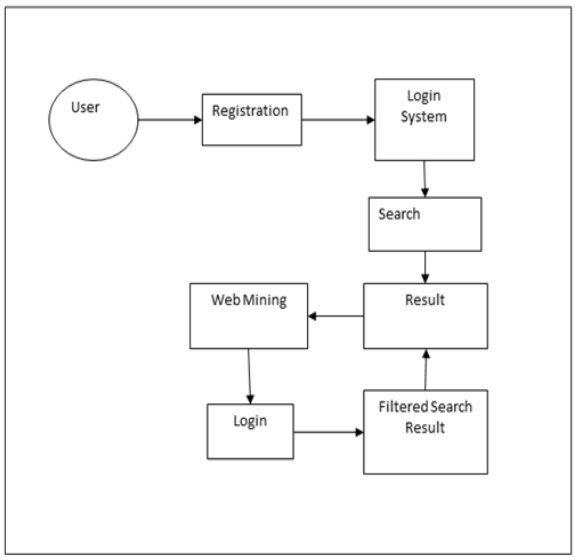

Figure 1 : Proposed Work

IV. EXPERIMENT AND DISCUSSION

The remarkable development of information on the Web has forced new challenges for the construction of effective search engines. The only input to the search engine is the keyword and it searches the whole WWW to provide the relevant information to the users. However, most of users are likely to use only a few keywords to convey their information requirements, and thus the search queries usually do not correspond to what the users want specifically. In addition, with the huge development of the information presented on the Web, it is very complicated for Web search engines to satisfy the user information requirement only with a short ambiguous query. To overcome such a basic difficulty of information retrieval, personalized search, which is to provide the customized search results to each user, is a very promising solution.

Algorithm	True positive	True negative	False positive	False negative
K-means	5	10	20	30
Page rank	10	8	15	20
Greedy algorithm	20	5	10	10

Table 1 : Performance Measurement

Fundamentally, in studying how a search can be personalized, the most significant thing is to accurately identify users' information. During web search the results by the search engine are converted to feature vectors using the same pre-processing techniques that were used for the user profile. Thus each search result is represented by a feature vector. These feature vectors are then passed to Similarity Scorer which assigns them scores based on their similarity to interest vectors. Each result is assigned a score equal to the maximum of the similarity scores with each interest vector. The results along with their scores are passed to Re- Ranker which sorts the results based on the scores assigned and the modifies the ordering that is ultimately presented to the user. For web search personalization we conducted various experiments differing in the way a user's profile was created and search results represented. We found that the best performance was obtained when the short term history was used for disambiguation and long term history was used for constructing the feature space. We can evaluate the performance using accuracy metrics. The accuracy metric is evaluated as

Implement the system using C#.NET as front end and SQL Server as back end. The personalized search results can be shown in fig 1.

Figure 2 : Search Results of Proposal Method

Evaluate the performance using accuracy metrics. The accuracy metric is evaluated as

$$\textbf{Accuracy} = \frac{\textbf{TP+TN}}{\textbf{TP+TN+FP+FN}} \textbf{*100}$$

The proposed algorithm provide improved accuracy rate than the machine learning algorithms.
Accuracy table shown in table 2.

Algorithm	Accuracy (%)
K-means	23
Page rank	34
Greedy algorithm	55

Table 2 : Accuracy Table

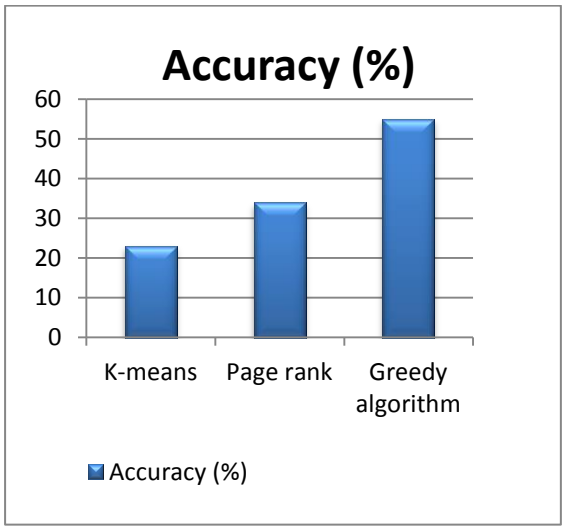

Figure 3 : Accuracy Analysis Chart

From the performance chart, Greedy algorithm provide high level accuracy than the existing machine learning algorithms.

V. CONCLUSSION

Personalized web search customizes the search results to improve the search quality for web users. However, user's personal information might be exposed in the user profile which is the basis in personalized web search. In this project, proposed a grouping approach for anonym zing user profiles with likability notion to bound the probability of linking a potentially sensitive term to a user by presented a greedy clustering technique with novel semantic similarity metric based on augmented user profiles in order to address the user profiles and take into account semantic relationships between user profiles. Personalized search is a promising way to improve search quality. However, this approach requires users to grant the server full access to personal information on Internet, which violates users' privacy. In this work, we investigated the feasibility of achieving a balance between users' privacy and search quality. First, an algorithm is provided to the user for collecting, summarizing, and organizing their personal information into a hierarchical user profile, where general terms are ranked to higher levels than specific terms. For Future work will try to resist adversaries with broader background knowledge such as richer relationship among topics (e.g., exclusiveness, sequentiality, and so on), or capability to capture a series of queries (relaxing the second constraint of the adversary) from the victim.

VI. REFERENCES

[1]. A. Cockburn, S. Greenberg, S. Jones, B. Mckenzie, and M. Moyle. Improving web page revisitation: analysis, design and evaluation.IT &Society, 1(3):159-183, 2003.

[2]. J. Teevan, E. Adar, R. Jones, and M. Potts. Information re-retrieval: repeat queries in yahoo's logs. In SIGIR, pages 151-158, 2007.

[3]. C. E. Kulkarni, S. Raju, and R. Udupa. Memento: unifying content and context to aid webpage re-visitation. In UIST, pages 435-436, 2010.

[4]. T. Deng, L. Zhao, H. Wang, Q. Liu, and L. Feng.Refinder: a context-based information re-finding system. IEEE TKDE, 25(9):2119-2132, 2013.

[5]. T. Deng, L. Zhao, and L. Feng.Enhancing web revisitation by contextual keywords. In ICWE, pages 323-337, 2013.

[6]. J. A. Gamez, J. L. Mateo, and J. M. Puerta. Improving revisitation browsers capability by using a dynamic bookmarks personal toolbar. In WISE, pages 643-652, 2007.

[7]. R. Kawase, G. Papadakis, E. Herder, and W. Nejdl. Beyond theusual suspects: context-aware revisitation support. In HT, pages 27-36, 2011

[8]. D. Morris, M. R. Morris, and G. Venolia. Searchbar: a searchcentric web history for task resumption and information refinding. In CHI, pages 1207-1216, 2008.

[9]. L. Tauscher and S. Greenberg.Revisitation patterns in world wide web navigation. In CHI, pages 399-406, 1997.

[10]. S. S. Won, J. Jin, and J. I. Hong. Contextual web history: using visual and contextual cues to improve web browser history. In CHI, pages 1457-1466, 2009.