ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

Proficient Search Results for Users with Generalization Algorithm

K.Surya[1], D. Asir Antony Gnana Singh[2], E. Jebamalar Leavline[3]

Department of CSE, Bharathidasan Institute of Technology, Anna University, Tiruchirappalli, India [1&2]

Department of ECE, Bharathidasan Institute of Technology, Anna University, Tiruchirappalli, India [3]

ABSTRACT: Web search engines (e.g. Google, Yahoo, Microsoft Live investigate etc.) are widely used to find certain data among a huge amount of information in a minimal amount of time. These useful tools also pose a privacy threat to the users: web search engines profile their users by storing and analyzing past searches submitted by them. In the proposed system, fuzzy-c means clustering algorithm is implemented for improving the search results or to reduce the search time. Personalized search is a promising way to improve the accuracy of web search, and has been attracting much concentration recently. However, effective personalized search require collecting and aggregating user information, which often raises serious concerns of privacy infringement for many user applications. Indeed, these concerns have become one of the main barriers for deploying personalized search to address the privacy threat, new solutions propose new mechanisms that introduce a high cost in terms of computation and communication. This work proposes l-diversity algorithms to improve the privacy of user profiles in personalized web search.

KEYWORDS: Personalized web search, fuzzy c-means clustering algorithm, l-diversity technique, privacy preserving.

I. INTRODUCTION

The search engines are attempting to satisfy the user's needs by ranking the web pages with respect to queries. The recent improvements and extensive use of high-throughput technology are produced day by day in using query and click log data in web search. The explosion of the information obtainable on the internet and its heterogeneity present a challenge for keyboard based search technologies to find useful information for users. These technologies have a systematic behavior of showing the same result for all users query at a certain time. Sometimes, there is ambiguity on user's query; it reduces the performance of these technologies. Our investigation shows that the main reason for this issue is that they do not consider the user perspective in the retrieval process.

As search engines reach the limits inherent in selecting data and are hungry for more data in a competitive environment, mining cursors movements, suspended, and scrolling becomes important. A series of interactions of a process of querying, knowledge, and reformulating between users and search engine have to satisfy a single info need. The proposed system has a better method to convert user query to user search goal that is known as feedback gathering, pseudo-document, and clustering technique. And they have used criteria to evaluate the performance and number of clustering techniques called Classified Average Precision (CAP). When compared to existing methods of extracting the information about the user click through data as from direct click-through logs, in earlier literature they use a feedback session which will reduce the noise by using the information from direct click through the pseudo-documents are mapped from the feedback gathering, also the binary vector method is used which will not provide the exact details to extract the keywords [1].

For clustering, in existing methods they use K-Means clustering algorithm which is simple and efficient. But it is computationally difficult and the similarity between the pseudo terms is also important while clustering. The keywords are called as rifle text. Generally, web search session contains the series of successive queries to satisfy a single need and some clicked search results. User rifle goals are assumed for a exact query. Therefore, the single session introduces only one query, which differentiates from the predictable session. The feedback session is based on a single session and the feedback session will play a major role. The feedback session defines both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. It was motivated that before the previous click, all the URLs

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

have been scanned and evaluated by users. Therefore, further the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks. For a query, users will usually have some unclear keywords representative their interests in their minds. They utilize these keywords to determine whether a document can satisfy their needs. However, while search texts can reflect user needs of information, they are underlying and not uttered explicitly. Therefore, the feedback session can be represented as pseudo-documents as substitutes to approximate rifle texts. pseudo-documents can be used to assume user rifle goals.

II. RELATED WORKS

There are several earlier attempts on personalizing web search. One advance is to ask users to specify general interests. The user interests are then used to filter search results by checking content similarity between returned web pages and user interests [2]. X. Li et al. have proposed a semi-supervised learning approach with the use of click graphs to query set on classification. They aim at radically expanding the training data as an attempt to improve performance of classification. They serially performed graph-based learning and content-based learning in a unified framework. They completely investigate an orthogonal approach instead of enriching feature representation. Specifically, they assume class memberships of unlabeled queries from those of labeled ones according to their proximities in a click graph.
In [3], authors presented a system to automatically build a hierarchical user profile using browsing history and standard emails. The profile represents the user's implicit personal safety. In this, general terms with superior frequency are at higher levels and exact terms with lower frequency are at lower levels. The user specify a threshold, which represents the least amount number of documents where a frequent term is required to occur. Next, a new partial profile is build using the threshold and the absolute profile, this limited profile has the maximum levels. The authors inhabited a search engine wrapper on the server side that incorporates the partial user profile with the results retrieved from the web search engine. This method was tested using the MSN search engine. Nevertheless, this solution does not prevent the search engine from building user's profiles because users submit their queries directly to the search engine [4].

Kim and Chan [5] also built user profiles from the equal source, however they use clustering to create a user interest steps. The collected Web pages are then assigned to the appropriate cluster. The fact that a user has visited a page is an indication of user interest in that page's content. Extending this idea, Chan describes a metric to estimate the level of user interest; for example the percentage of links visited on a page or URL presented in bookmarks.

According to Gruber [6], ontology is a specification of a conceptualization. Ontologies are capable of defining in dissimilar ways but they all represent taxonomy of concepts along with the relations between them. In the situation of the World Wide Web, ontologies are significant as they formally define terms shared between any type of agents without ambiguity, allowing in sequence to be processed automatically and accurately. OntoSeek [7] is an example of system based on ontologies. It utilizes the sources one by one such as product catalogs and yellow pages it applies conceptual graphs to represent both queries and resources.

To achieve effective personalization, profiles should decide between long-term and short-term interests and include a model of the user's context, i.e., the mission in which the client is currently engaged and the environment in which they are situated [8]. Several systems have attempted to provide personalized search that are tailored based upon user profiles that capture one or more of these aspects [9]. In this OBIWAN project [10], search results starting a conventional search engine are classified with respect to a reference ontology based upon the snippets shortening the retrieved papers are re-ranked based upon how well their concepts match those that appear highly weighted in the user profile.
Competitive Intelligence Spider and Meta Spider [11] are part of a client-based applications that collect and organize web documents on the user's machine. Spiders may group information directly from web sites or through search engines. Collected papers are then analyzed and noun phrases are extracted to create a personal dictionary for the user to guide future searches. The noun phrases are also used to organize the documents and a graphical map of the results is generated. Users can personalize the search plainly by selecting specific Web sites, the number of trap pages to collect, and the noun phrases used in the ultimate map of results.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

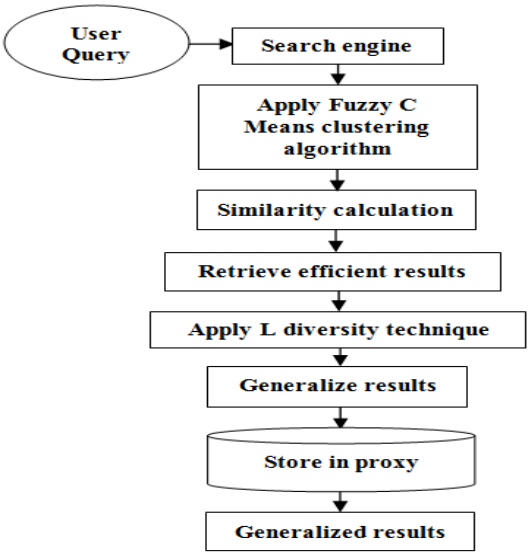
The Personal Search Assistant [12] [13] is an application in which a background process collects information on behalf of a user by submitting queries to various search engines. Results are stored on the local machine and are analyzed so that they can be organized conceptually. The user manually creates a conceptual database that is input to a personal agent responsible for building a user profile. The profile is used to filter the results of later searches.

Privacy concerns are natural and important especially on the internet. Some previous studies on private information retrieval (PIR) [14] [15] focus on the problem of allowing a consumer to retrieve information while keeping the query confidential. Instead, this study targets preserving privacy of the user profile, while still benefiting from selective permit to general information that the user agrees to discharge To our knowledge, this trouble has not been considered in the context of personalized search. One promising reason for this is that personal information, i.e. browsing describing and emails, is mostly shapeless data, for which privacy is difficult to measure and quantify. [3] proposed a privacy protection solution for personal web server (PWS) based on hierarchical outline. Using a user-testing access, a generalized outline is obtained in effect as a rooted subtree of the total profile. Seriously, this work does not concentrate on the query utility, which is dangerous for the examine quality of PWS.

The solutions in class two do not require third-party assistance or collaborations between social network entries. The users only convict themselves and cannot accept the exposure of their complete profiles an secrecy server. In [16], Krause and Horvitz employed statistical techniques to learn a probabilistic style, and then utilize this form to generate the near-optimal partial profile. One main constraint in this effort is that it builds the user profile as a finite set of attributes, and the probabilistic imitation is qualified through predefined frequent queries. These assumptions are discomfited in the context of PWS according to Xu et al. The authors also concluded that personalization significantly increased output quality for uncertain and semi-ambiguous queries, but for noticeable queries,one should prefer general web search. In [17], queries were divided and captivated to fresh queries and returning queries. The authors found that recent history tends to be much more useful than remote history especially for fresh queries while the complete history was helpful for improving the search accuracy of recurring queries. This also give us a sense that not all queries should be personalized in the same way [18]. These conclusions stirred our detailed analysis[19].

III. PROPOSED METHOD

Fig. 1 shows the architecture of the proposed system. First, dataset pre-processing takes place. In that, we have to select the dataset and eliminate the null values in the dataset. After the dataset has been pre-processed, user login process is carried out. Then, the user query is given to the search engine.



,. Fig.3.1: Architecture of the proposed system

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Special Issue 6, May 2015

User's profiles are generated by using the l-diversity and sent to the server. Then, the server analyses the user query and retrieves results by using the fuzzy C-means algorithm by applying the similarities between the user given query and word in the database and retrieve the results and then store the generalized results into the history.

System description

Web search engines profile their users by storing and analyzing past searches submitted by them. In the proposed system, it can be implement the clustering algorithms for improving the better search quality results. To address this privacy, current solutions propose new mechanisms that introduce a high cost in terms of computation and communication. In this, it present a novel protocol specially designed to protect the users privacy in front of web search profiling. It can be proposed and try to resist adversaries with broader background knowledge, such as richer connection among topics. Richer relationship means it generalize the user profile results by using the background data which is going to store in history. Through this it can be hide the user search results.

IV. CONCLUSION

This paper proposed a client-side privacy protection framework called UPS for personalized web search. UPS could potentially adopted by any PWS that captures user profiles in a hierarchical taxonomy. This approach ields better efficiency results when compared with existing system. It will provide privacy mechanism when adversaries recover the results by using background knowledge. In this work similarity measures are taken using the fuzzy c-means clustering algorithm.

REFERENCES

[1] Shen, X., Tan, B., and Zhai, C., 2005, "Context-Sensitive Information Retrieval Using Implicit Feedback", Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval, pp. 43-50.
[2] Breese, J.S., Heckerman, D., and Kadie, C.M., 1998, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence, pp. 43-52.
[3] Xu, Y., Wang, K., Zhang, B., and Chen, Z., 2007, "Privacy-Enhancing Personalized Web Search," Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 591-600.
[4] Zhu, Y., Xiong, L., and Verdery, C., "Anonymizing User Profiles for Personalized Web Search" Proc. 19th Int'l Conf. World Wide Web (WWW), pp. 1225-1226.
[5] Kim, H.R., Chan., P.K., 2003, "Learning implicit user interest hierarchy for context in personalization", Proc. 8th international conference on Intelligent user interfaces, Miami, Florida, USA, pp. 101 - 108.
[6] Gruber, T. R, 1993, "A translation approach to portable ontologies", Knowl Acquis., 5(2), pp. 199 – 220.
[7] Guarino, N., Masolo, C., Vetere. G., 1999, "OntoSeek: Content-Based Access to the web", IEEE Intell Syst., 14 (3), pp. 70 – 80.
[8] Mizzaro, S., Tasso, C., 2002, "Ephemeral and persistent personalization in adaptive information access to scholarly publications on the web", Proc. Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp 306 – 316.
[9] Pitkow, J., Shutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., and Breuel, T., 2002, "Personalized Search",. Comm. ACM, 45 (9), pp. 50-55.
[10] Shen, X., Tan, B., and Zhai, C., 2005, "Implicit User Modeling for Personalized Search", Proc. 14th ACM Int'l Conf. Information and Knowledge Management, pp. 824-831.
[11] Chau, M., Zeng, D., Chen, H., 2001, "Personalized spiders for websearch and analysis", Proc. 1st ACM-IEEE Joint Conference on Digital Libraries, pp. 79 - 87.
[12] Gauch, S., Chafee, J., and Pretschner, A., 2004, "Ontology-Based Personalized Search and Browsing", Web Intelligence and Agent Systems. 1 (3-4), pp.219-234.
[13] Kaushik, P.R., Narayana Murthy, K., 1999, "Personal Search Assistant: a configurable personal meta search engine", Fifth Australian World Wide Web Conference, Southern Cross University, Lismore, Australia.
[14] Gasarch, W., 2004, "A survey on private information retrieval", Bull. Eur. Assoc. Theor. Comput. Sci., 82, pp.72—107.
[15] Shen, X., Tan, B., and Zhai, C., 2007, "Privacy Protection in Personalized Search", SIGIR Forum, 41(1), pp. 4-17.
[16] Krause A., and Horvitz, E., "A Utility-Theoretic Approach to Privacy in Online Services," J Artif Intell Res., 39, 633-662.
[17] Xu, Y., Wang, K., Yang, G., and Fu, A.W.C., "Online Anonymity for Personalized Web Services", Proc. 18th ACM Conf. Information and Knowledge Management, pp. 1497-1500.
[18] Tan, B., Shen, X., and Zhai. C., 2006, "Mining long-term searching history to improve search accuracy", Proc. of KDD, pp. 718–723.
[19] Shou, Lidan, et al. "Supporting Privacy Protection in Personalized Web Search." Knowledge and Data Engineering, IEEE Transactions on 26.2 (2014): 453-467.