Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-5, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

Analytical Study on Personalized Information Retrieval System

Sneha A. Taksande[1], Prof. A. V. Deorankar[2]
PG Scholar, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India.[1]
Associate Professor, Department of Information Technology, Government College of Engineering Amravati, Maharashtra, India.[2]

Abstract- *World Wide Web has become inimitable as it is a vast resource of information, but such a vast information available over the internet has made web search a time consuming and a very complex process. Users represent their information needs as series of queries based on their search intentions when they use existing search engines. Though there is a rapid advancement in the field of information retrieval and especially in the field of search engine but still there is a lot of scope for development of personalized search engine. This paper addresses this question of how to build a personalized search system which can exploit the context of the query by using users search history to gather additional information present in user history. Significant information can be retrieve from such a vast available resource using various Information Retrieval methods. This paper would review few of these methods for intelligent information retrieval systems based on the Personalization technique for information retrieval. Since Personalized Web Search is a method of searching to improve the quality and accuracy of web search thus it has gained much attention recently.*

Keywords- *Information Retrieval, Search Engine, Personalization.*

1. Introduction

Today, search engines such as Google, Bing, Yahoo etc. that are used by people searches the information using keyword to answer the queries from the users. These search engines searches unnecessary pages because the main focus of these search engines to solve queries close to accurate result and most of the time user did not get the required information. Searching and surfing documents over internet is becoming an integral part of people's life. Internet access, such as World Wide Web (WWW), becomes one of the most important platform as collection of documents with increasing pace of technologies and the need of digital data. There is an increasing demand of retrieving such documents from a large resource of information that are relevant to the information asked for. Thus, Information Retrieval is gaining more importance day by day to cope up with this demand and supply paradigm. Retrieving documents online is of interest in the information retrieval (IR) community. Document retrieval actually refers to finding such documents which are similar for a given user's query.

At present, people use search engines as a way to search any kind of information. Generally existing search engines that are used, place heavy burdens on users when they try to represent their information needs as queries, because existing search engines only accept a set of keywords as a query and output the list of retrieved pages. Therefore, users often need to submit queries repeatedly until they find an answer.

Accuracy and speed are two fundamental needs of effective retrieval methodologies. Thus the advanced search engines which enable users to search more efficiently by restricting target domains have been studied to improve search efficiency. A user can specifies his query as full described sentence or just a few keywords. Among the broadly used retrieval methods used by different search engines are keywords based searching methods, like Google, where unskilled users provide just a few keywords to the search engine and in return Google provide a list of relevant documents available online. Sometimes this search cannot fulfill the user's requirements. So apart from the problems of scaling traditional search techniques to deal with the ambiguous query, there are new technical challenges involved with using the additional information present in user search history to produce better search results. This paper presenting the different intelligent information retrieval techniques based on personalization by considering various aspects of users profile and users interests.

2. Basic Concepts of Search Engine

Search is a technique used to find relevant information for the user on internet. Search engine is computer software designed to search for desired information on WWW. A demand for technology is growing day by day as a result we use search engine

frequently. Web search engines are composed of three main elements: the crawler recovers documents from the Web. According to the URL address of the web pages Search engine crawls the web pages. The indexer indexes the documents collected by the crawler which contains the keyword that matches with user query of any specific web page and shows the result. The searcher solves user queries by using the generated index and other components required to achieve efficient performance. As internet growing at an exponential rate search engine regularly updated index.

Figure 1 shows the relationship among these three components. In this article, we focus on how to simulate the searcher to evaluate the performance using personalization tool. In the searcher, users submit queries composed of keywords through search interface and, in return, they receive a list of pointers to Web documents ordered in accordance with a relevance metric function on the query keywords. Here searcher refers as the search engine.

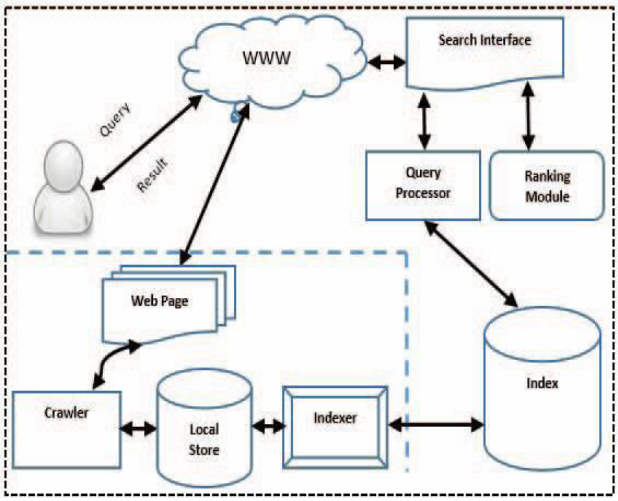

Fig. 1. Architecture of web search engine

The ranking module then assigns the ranks to these web pages in order of their relevance and importance with the help of ranking algorithm. There are different ranking algorithms to rank these web pages by considering different aspects of users search intentions. These web pages are then displayed in search engine interface in decreasing order of relevance and importance.

A search engine is usually built as a collection of services deployed on a large cluster of processors, wherein each service is distributed onto a set of processors. The processors and the communication network are expected to be constructed from commodity hardware. Each processor is expected to be a multicore processor, enabling efficient multithreading on shared data structures. Message passing is performed among processors to compute on the distributed memory supported by the processors.

3. Information Retrieval Methods based on Personalization

Various researches have been done on personalized web search and user profile. This section contains the literature review of profile-based personalization and privacy protection in PWS system. The users' data can be obtained from many resources explicitly, but this approach provides the user's interface through which users can give his details using which the system produces the results.

3.1 Personalizing search based on user search histories

User profiles is constructed on the basis of user interactions with a particular search engine is used. A Google Wrapper i.e. wrapper around the Google search engine was used for monitoring user search activities. This wrapper logs the queries search results and clicks performed per user. Then the user profiles are created using this information. Here user profiles are represented as a weighted topic hierarchy. The weights represent the amount of user interest in the topic. The user profiles are constructed from web pages browsed by the user, but focus was given to user's search history rather than their browsing history. When a user query was submitted, the top ten results obtained were re-ranked on the basis of their original rank and conceptual similarity to the user profile. A document profile was constructed by the classification of search results and summaries. The document profile was similar to the user profile. Then a comparison between the document profile and the user profile was conducted to find how similar each document and the user interests are. The cosine similarity function was used for the comparison. Then the conceptual rank of the documents was calculated by re-ranking the documents by their conceptual similarity. The combination of the conceptual rank with Google's original rank was used for calculating the final document rank. The concept hierarchy was static. The amount of user interest can vary, so a dynamically adaptive hierarchy construction was needed.

3.2 Automatic identification of user interest for personalized search

A search engine learns user preferences based on his past click history data and personalize based on user preference. As most of the users are reluctant to provide any explicit feedback on search results and interests, an automatic learning of the

user preference was of great advantage. Here, the user preferences are represented as a topic preference vector defined as an m-tuple T= [T(1),.....T(m)], m is the number of topics under consideration. A user's page preference vector was defined as an n-tuple, P= [P(1),....P[n)], n is the total number of web pages. A topic-driven random surfer model was used to learn the topic preference vector of a user from his past click history. Average Rank metric was used to measure performance with reduced human involvement. To measure the overall improvement of personalization in the quality of search results metrics like page rank, topic-sensitive page rank, personalized page rank and query-based personalized page rank were used.

3.3 Personalized privacy preservation

Personalized discretion was used for privacy preservation. By personalized anonymity, the user was given freedom to specify how much privacy protection he wants for his sensitive data. The user can specify the degree of privacy protection needed by specifying guarding nodes in the taxonomy of the sensitive attribute. A generalization framework was developed for customization of privacy needs. Personalized anonymity provides better protection to individual users than the previous techniques k-anonymity and l-diversity. In personalized anonymity a third person cannot directly associate an individual user with his sensitive data.

3.4 Large-scale evaluation and analysis of personalized search strategies

A large-scale evaluation framework performed personalized search on the basis of query logs. The user clicks were recorded in query logs for evaluating search accuracy and simulating various personalized re-ranking strategies. The work concluded that personalization has different effect on different queries, users and search contexts. Personalization has more effect on the queries with large click entropy and little effect on the queries with small click entropy.

3.5 Privacy protection in personalized search

Different levels of privacy protection were provided to different users. Privacy was treated as the identification of an individual. Four levels of privacy protection are defined. Level I - Pseudo-identity: A PWS system has pseudo-identity if the user identity is replaced by pseudo identity, which contains less identifiable information than user identity. This is the lowest level of privacy protection. The descriptions of user information needs are aggregated according to pseudo identity.

Level II- Group Identity: A PWS system has group identity if a group of users share a single user identity. The description of user information needs was aggregated at the group level according to user identity. Level II has higher privacy protection than Level I. Individual user profile cannot be constructed at Level II. Only an aggregate group profile can be constructed. So it is difficult to extract the information needs of an individual user.

Level-III- No Identity: A PWS system has no identity if the user identity is not available to the search engine. The description of user information needs cannot be aggregated on the search engine. Level III has privacy protection than Level- I.

Level-IV- No Personal Information: A PWS system has Level IV privacy protection if neither the user identity nor the description of user information is available to the search engine.

3.6 Preserving users privacy in web search engine

The Useless User Profile (UUP) protocol was used to protect privacy of a user in web search profiling. For preserving user's privacy the user profile was not given directly to the search engine. The protocol was used to shuffle queries among a group of users who issue them, providing a distorted user profile to the search engine. The UUP was based on the El Gamal encryption and re-masking cryptographic tools .No entity can profile a particular individual because of UUP. This scheme constructed a reliable profile and privacy was met to certain level. But the delay required for cryptographic operations and network communications was there. The existence of a third party anonymizer, which was not readily available over the Internet, was assumed.

3.7 Profile based personalized search

Personalization is the process of providing right information to the right person at the right time. For personalization user interests are to be studied. This requires collection, analysis and accumulation of user data both general and personal. But generally users are reluctant to disclose their personal data as it may reveal their personal behavior and affect privacy. In profile based PWS, user interests and data are modeled as user profiles. Previous works built user profile statically, only once offline. But such “one profile fits all” methods suffers from drawbacks. It affects the search quality for ad-hoc queries. The existing methods do not consider the customization of user privacy requirements. This may cause overprotection of some user private data while exposing some privacy sufficiently. Many existing personalization techniques need iterative user interaction for creating personalized search

results. This method is infeasible for runtime profiling. In profile based personalized search system different parameters of users can be taken into consideration such as user's profession, his search history, preferences given by the other users, etc. The re-ranking can be performed using these parameters. After re-ranking the most relevant data will be up ranked by the system and the less relevant below it. This leads to improve the accuracy of search system.

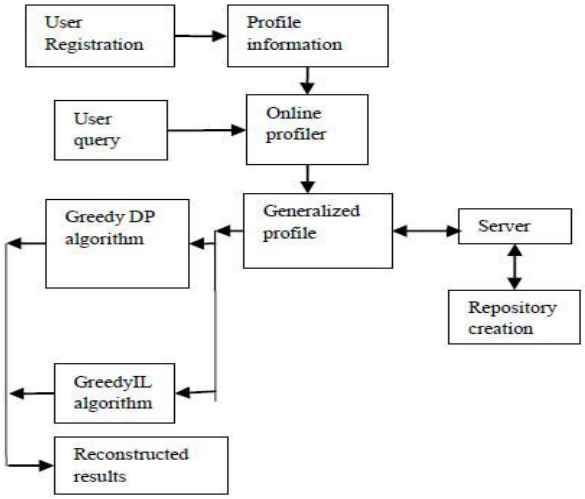

Fig. 1 Block diagram of Generalized Profile Construction

4. Conclusion

The huge mass of information on web has forced to develop efficient information retrieval system for web search engine. For given small query to search engine it has to searches whole World Wide Web to provide relevant information to user. So this paper presents various techniques based in personalization as Personalized Web Search (PWS) is one of the efficient technique that improves the quality of search services on the Internet. Privacy preservation methods are used in PWS to prevent leakage of personal information on the Internet. The need for privacy and privacy risks related to the different approaches of personalization were studied. So it also addresses the issue of privacy preservation.

References

[1] Ramitha A T and Dr. Jayasudha J S, "Personalization And Privacy In Profile-Based Web Search," International Conference on Research Advances in Integrated Navigation Systems (RAINS), IEEE, April 06-07, 2016.

[2] Sanjib Kumar Sahu, D. P. Mahapatra and R. C. Balabantaray, "Analytical Study on Intelligent Information Retrieval System Using Semantic Network," International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2016.

[3] Shogo Kori, Yanjun Zhu, Koichi Yamaguchi, Satoru Takiguchi and Yasufumi Takama, "Analysis of User's Behavior Based on Search Intentions for Information Retrieval Using Search Engines," IEEE, 2015.

[4] Prakasha S, Shashidhar HR and Dr. G T Raju, "Structured Intelligent Search Engine for Effective Information Retrieval using Query Clustering Technique and Semantic Web," IEEE, 2014.

[5] Dr. Daya Gupta and Devika Singh, "User Preference Based Page Ranking Algorithm," International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2016.

[6] Anoj Kumar and Mohd. Ashraf, "Efficient Technique for personalized web search using users browsing history," International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2015.

[7] Fabrizio Lamberti, Andrea Sanna, and Claudio Demartini, "A Relation-Based Page Rank Algorithm for Semantic Web Search Engines," IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 1, January 2009.

[8] Radheshyam Prajapati and Suresh Kumar, "Enhanced Weighted PageRank Algorithm based on Contents and Links Visits," IEEE, 2016.