AN EFFICIENT RE-RANKING ALGORITHM BY REFINING CLIENT SIDE LOG TO PREDICT USER'S INTEREST BASED ON WEB PERSONALIZATION

Mercy Paul Selvan [1], Dr. A. Chandrasekar[2],

Department of CSE ,Faculty of Computing, Sathyabama University, Chennai., Tamil Nadu
Department of CSE, St. Joseph's College of Engineering, Chennai, Tamil Nadu.

Abstract—

With the rich and growing wealth of information on the internet the process of finding a specific piece of information may often become frustrating and time-consuming for users. In the previous research to learn user interest using ontological profiles, the web pages visited and its dwell time have been considered. But considering these factors alone is not enough to extract user's interest accurately. In this paper, a hybrid personalized model of search engine based on learning ontological user profiles implicitly is presented. The aim of this paper is to optimize the search results to get more relevant information rather than simple keyword matching by user's recent browsing history. The user browsing behaviour is studied initially to get his area of interest and the search results for different users are based on his area of interest. In addition, these web pages are stored in user profile under positive and negative documents. Thus, a hybrid reranking algorithm that is based on the combination of different significant information resources collected from the reference ontology, user profile and original search engine's ranking has been proposed.

Keywords: search engine, personalization, user profile, ontology, information retrieval.

I. INTRODUCTION

Significance of search engine:

Search engines are the window through which the information available on the internet reaches the people corresponding to their needs.Search Engines essentially act as filters for the wealth of information available on the Internet. They allow users to quickly and easily find information that is of genuine interest or value to them, without the need to wade through numerous irrelevant web pages. This makes the job of search engine optimization more crucial and important.

Without any support each system engraves and frames a user's profile that illustrates known information about the individual, including demographic data, interests, preferences, goals and previous history. This information could be collected unambiguously by asking user to rate retrieved documents [9], or inevitably from click-history data [2], semantic web browsers [3] or log files [4].

One common line of research work is using ontology to engrave and frames ontological user profiles in order to fetch personalized search results. Ontological user profiles can be presented as an instance of reference ontology [5,6]. Sieg et. al [5], recommended an ontological user profile that is built by applying a spreading incitement mechanism to learn and maintain user interests. This contour then is used to re-rank the search results based on the users' current curiosity. In the way that [6], stated a web search system based on ontological user profiles. These profiles were built making use of the Open directory project (ODP) 1 as reference ontology.

The paper showed that this advent increases the performance of the system. A prominent line of works has abused contextual information to identify user's curiosity and preferences [4, 5, 6 and 7]. In addition, user context is identified based on current user query [4] or browsing behaviour [6] alongside with current user interests at the time of conducting a search. Xiang et al. [7] tried to remit three main inquiries: (i) how to take advantage of different context information; and (ii) how to aggregate this information to provide a more effective search engine;(iii)how to retrieve the user's recent browsing result. Likewise, [7] takes into the account not just the user's current context but also ignored results for these queries in the same search session. In Overall these systems, personalized search results are based on inquiry expansion [10, 12] or document re-ranking [11, 4, 6, 8, 2, and 13]. In [10] for instance, a hybrid personalization system is proposed to collect user context information using ontology and then expand the user query consequently.

However, with regard to document re-ranking approaches, some studies proposed re-ranking approaches based on a combination of different information sources. In [11] and [4] the re-ranking mechanism is based on using the cosine methods corresponds to compute the similarity between retrieved search results and documents from the ODP reference ontology. The re-ranking mechanism could also depend on a combination of the similarity [2] or the distance [13] between search results and documents in the user profile and the original Google ranking. In [13] both the original Google ranking, which is based on Page Rank scores, and as well as documents in the Google directory were considered in the reranking process.

In this paper, by considering the factors that are explained in previous research one cannot extract user's interest accurately. Therefore a modified algorithm considering the recently browsed web page by maintaining a user's recent browsing behaviour addition to the above factors has been proposed. Finally, the web page's interest weight is calculated considering the rank value of a web page that has been given by the user as an additional factor.

The framework of this paper is briefed below. In section II, the main architecture discussed by describing the two main phases of the architecture. In section III, hybrid re-ranking algorithm base on ontological user profile is discussed. In section IV, experimental evaluation is discussed briefly along with a graph. In section V, the paper ends with the conclusions.

II. MAIN ARCHITECTURE

The architecture mainly consists of users Profile(figure 1).The user profile is about building of the database by engraving and learning the user's behavior that includes collection of the users perusing behavior, preparing reference ontology, attaining and formulating the ontological user profile. The search result is given based on personalization process. The algorithm tailors recent search results to a particular user based on that user's interests and preferences. This process gives the results with regard to user's ontological profile in order to provide personalized search result.

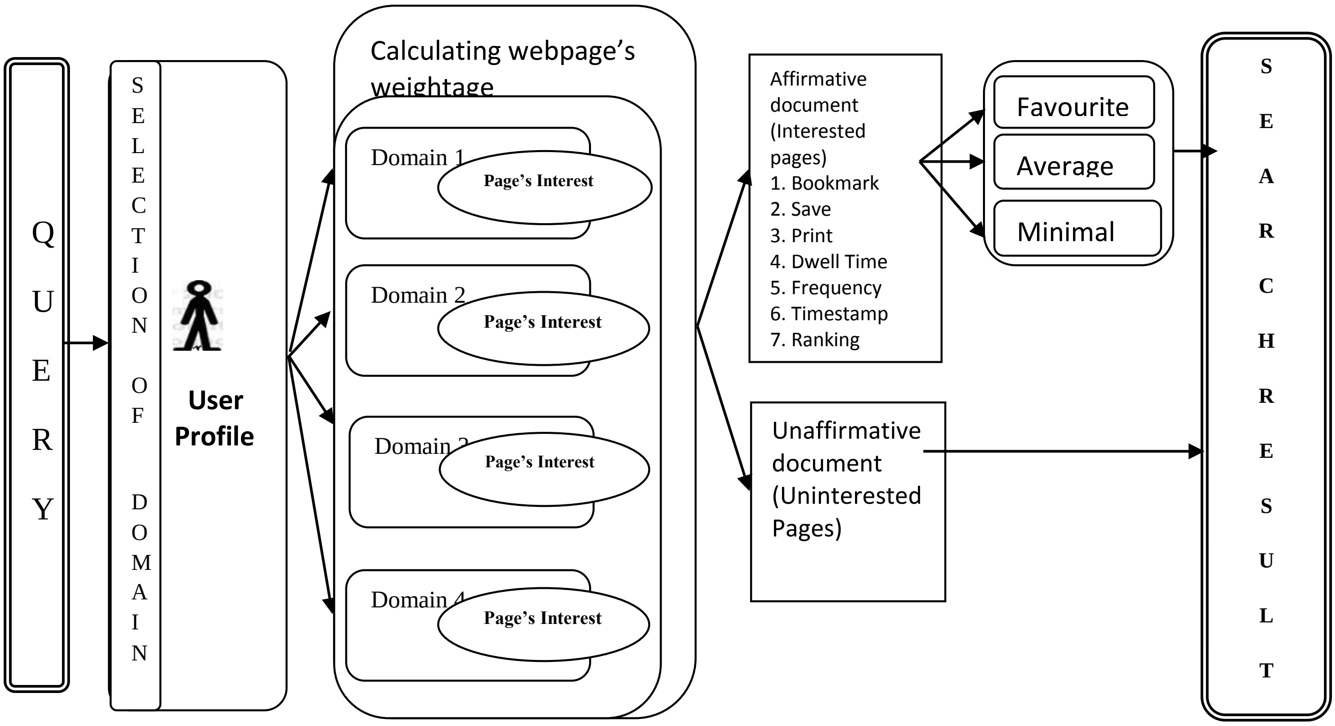

Figure 1. Main Architecture

A. Framing and attaining ontological user profile

Users perusing behavior: By concerning the user's browsing behavior essential information about the user should be collected .In the present trend, the user's information is collected unambiguously and inevitably. Despite, as collecting information unambiguously add more affliction on users, in this paper we lean on collecting user browsing information inevitably based on users browsing behaviour. For each browsing session, we observe different types of information namely: the visited web pages and the time spent in reading these web pages. This information is then enrolled and cached in a users log file. After each session, the contents of each visited web page are extracted and refined using various data extraction techniques including tokenization, sentence splitting and conflation [14].

Reference ontology: The existing methods have utilized ontology to show more striking Personalization systems [5, 15, 10]. In this paper, ontology plays a key role in modelling the user profile. Reference ontology provides a clear illustration of contents of a particular domain of application [15]. Distinctly flat representations, reference ontology provides a richer representation of information in that well-formed and structural relationships are defined unambiguously. In this paper, the user profile is engraved and user interests are generated from the Open Directory Project (ODP) reference ontology. ODP uses a hierarchical ontology scheme for organizing site listings. Listings on a similar topic are grouped into categories which can then include smaller categories.

The main motivations to create ODP were the frustration that many people experienced in getting their sites listed on Yahoo! Directory.

According to the ODP website, there are more than 4.5 million websites categorized in more than 590,000 categories.[1] Each concept in the ODP contains a number of websites and documents. The data in these documents and websites for each concept are elicited and plugged into one document [1]. The vector space (TF-IDF) classifier is then used to give each term in each document a weight from 0 to 1 (see formula 1).

After computing the term weights for each term, the cosine similarity algorithm [16] is used to map visited web pages to appropriate concepts in the ODP ontology.

Attaining and formulating an ontological user profile: The existing approaches create an ontological user profile as an exact example of the ODP reference ontology and its documents [3]. In this section, a new method is proposed for attaining and formulating an ontological user profile. Abducting and learning, amusing and dismaying concepts based on the data obtained by observing user behaviour, the proposed method endows the ontological user profile by populating it with the web pages that the user has visited. Additionally, we divide any visited web page to either an interesting (affirmative) web page or uninteresting (unaffirmative) web page. For each web page we extract its information and store it in the ontological user profile under the corresponding concept. Therefore, each concept in the ontological user profile would contain three documents. The first document is called the ***Ontology document*** which contains information fetched from reference ontology. The second document is the ***affirmative document*** and it holds all the information that was extracted from visited web pages recognized to be interesting to the user. The last document is the ***unaffirmative document*** and contains all the information that was extracted from the visited web pages identified as uninteresting to a user. For classification of the visited web pages as affirmative and affirmative, a novel mechanism is adapted to calculate the interest weight for each visited web page. The interest weight relies on different attributes namely: frequency of visit to that page, the time a user spent in reading a web page, and the context in which a web page occurred. The proposed method employs three different context statuses. The first status is the ***Browsed*** status which occurs when a user navigates a web page. ***Search result*** status is the second type of context and it takes place when a user submits a query and then clicks and browses the retrieved results. The final context status is the ***Favourite*** status that occurs when a user bookmarks, prints or saves a web page. All of these context statuses have different mass and values that reflect their importance. Formula 2 shows how the interest mass for each web page is calculated.

The TW is the total average of all the considered interest weights for all the visited web pages in each session and it is computed using the following formula.

$$Interest\ Weight = \sum_{url \in c}^{n} \left(\frac{\text{Duration}}{100}\right) * \text{Status mass} + \text{Rank}$$

In the above formula, we assign a status mass of 100 to all browsed web pages, while we assign a higher mass (150) to all web pages that are recently retrieved based on a user's submitted query to a search engine. We assign a higher mass because clicking on a particular retrieved search result provides a stronger implicit indication of web pages that might be interesting to users. Finally, we assign the highest status mass of 200 to all visited web pages that were bookmarked, saved or printed. This is because bookmarking, saving or printing a webpage is considered to be a strong implicit sign of interest. Additionally, if the page is considered as the affirmative, then the value assigned is 3 to 10.If the rating comes below 3 then that page is considered to be unaffirmative page.

Once the interest weight for each visited web page is calculated, all the interest weights for all visited web pages are stored in a stack history. At this point, some studies [2, 18] rely on a pre-defined and fixed threshold to identify interesting and uninteresting web pages. Yet, such an approach lacks flexibility and adaptability as assigning a fixed and pre-defined threshold for all users is not effective; each user may have diverse browsing behaviour. Therefore, it is essential to distinguish between different user types including heavy, medium and light Internet users. To trounce this limitation, a new parameter Threshold Weight (TW) which is calculated based on each user's browsing behaviour.

$$TW = \frac{sum\ of\ interest\ webpage}{total\ number\ of\ webpages}$$

All the web pages are categorized into affirmative or unaffirmative corresponding to the calculated TW. The web pages whose interest weight is above the TW are considered to be interesting, while the web pages whose interest weight below than the TW are treated as uninteresting ones. The contents of both groups (affirmative and unaffirmative) web pages are extracted and are associated to the corresponding concept in the ontological user profile.

This project talks about number of time the web page is visited by the users as an additional factor for sorting the result for the given query based on the users perusing history.

Attaining User's recent browsing History:

This paper is to optimize the search results to get more relevant information rather than simple keyword matching by user's recent browsing history. The user browsing behaviour is studied initially to get his area of interest and the search results for different users are based on his area of interest. In addition, these web pages are stored in user profile under positive and negative documents. The web page's interest weight is calculated considering the rank value of a web page that has been given by the user as an additional factor.

III. HYBRID SEARCH ENGINE OPTIMIZATION ALGORITHM

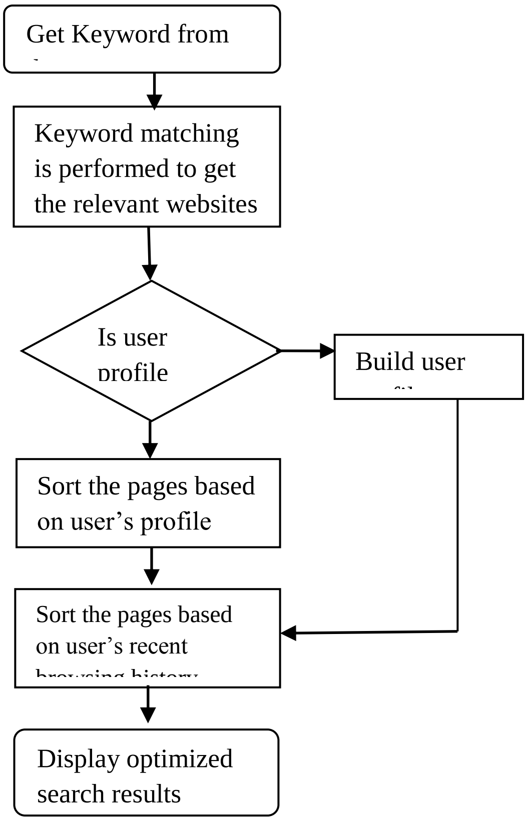
Get Keyword from
Keyword matching is performed to get the relevant websites
Is user profile
Build user
Sort the pages based on user's profile
Sort the pages based on user's recent
Display optimized search results

Figure 2. Flow Chart

When the user registers, the user will not have his personal profile. Thus the initial few times the user enters the keyword in the search engine; the results are sorted with the described factors but excluding the personalized user profile from the list. In these few times, the user profile is simultaneously built.

The algorithm initially takes the keyword from the users as input. Keyword matching is done as the first step. From the built user profile the algorithm concludes which category is the user's area of interest. Following, the websites which have been selected from the keyword matching algorithm under this category is again sorted based on user's recent browsing history that have been visited. Then the sorting is repeated keeping the interest weight of the web page as factor. Thus the search engine gives an optimized result saving the time of the user in searching for the content rich website.

IV. EXPERIMENTAL RESULTS

The proposed system is mainly useful for collecting the user's behavior that is based on the user's recently browsing behavior. For our experiments, we collect the datasets from the 3 user's perusing history. Depending up on the keyword search or query given by the user, we are suggesting the most expected pages as a result. Addition to this we are calculating the interest page weight by keeping some values for both interested and uninterested pages. By calculating the interest weight of a page we give more efficient result through our proposed system.

In this experiment, our goal is to appraise the concert of our proposed re-ranking algorithm. An effective re-ranking algorithm should place the most relevant results based up on the user's recent search and the browsing history in the top of retrieved results. For this experiment, we compare the searching behaviour of 3 users based on the keyword match. For instance, if a user searches with a keyword "Apple" to know about the fruit apple before two months, but recently his search about "Apple" is for Apple software .In this case when again the user seeks for Apple, the result will be given based on his recent search, i.e. for Apple software.

The performance of our proposed system is better than the existing system, because the existing system did not talk about the user's recent behaviour and the interest page weight. When comparing the efficiency of the existing system, our proposed systems Average rank is higher.

$$Average\,rank(p,q) = \frac{Recent\ Query\ Result}{Interest\ weight}$$

Where q is the recent query given by the user and the p is the position of the web page result based on the rank calculated.

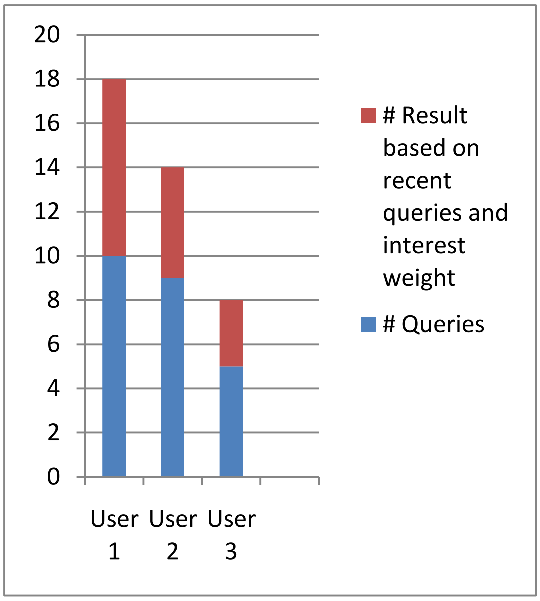

Fig 3: Relevancy Chart

V. CONCLUSION:

In this paper, a hybrid personalized model of search engine based on learning ontological user profiles based on user's recent browsing history is implicitly presented. This helps to optimize the search results to get more relevant information rather than simple keyword matching. The user's recent browsing behaviour is studied initially to get his area of interest and the search results for different users are based on his area of interest. In addition, these web pages are stored in user profile under positive and negative documents. The web page's interest weight is calculated considering the rank value of a web page that has been given by the user as an additional factor. Thus, a hybrid reranking algorithm that is based on the combination of different significant information resources collected from the reference ontology, user profile, mainly the user's recent browsing history and original search engine's ranking is proposed to give more relevant search results.

REFERENCES

[1] Ahmad Hawalah, Maria Fasli, A Hybrid re-ranking algorithm based on Ontological User Profile, 2011. 978-1-4577-1301-9/11/$26.00 © 2011 IEEE

[2] Li, L., Yang, Z., Wang, B. and Kitsuregawa, M. 2007. Dynamic Adaptation Strategies for Long-Term and Short-Term User Profile to Personalize Search. APWeb/WAIM, pp. 228-240.

[3] Sumalatha, M.R., Vaidehi, V., Kannan, A. and Anandhi, S.2007. Information Retrieval using Semantic Web Browser-Personalized and Categorical Web Search. ICSCN '07, pp. 238-243.

[4] Mohammed, N.U., Doung, T.H. and Jo, G.K. 2010 Contextual_Information Search Based on Ontological User Profile. ICCCI 2010, pp.490-500.

[5] Sieg, A., Mobasher, B. and Burke, R. 2007. Representing context in web search with ontological user profiles. In Proceedings of the Sixth International and Interdisciplinary

[6] Conference on Modeling and Using Context. Roskilde, Denmark.

[7] Challam, V., Gauche, S., and Chandramouli, A. 2007. Contextual Search Using Ontology-Based User Profiles, Proceedings of RIAO 2007, Pittsburgh, USA.

[8] Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E. and Li, H. 2010.Context-Aware Ranking in Web Search. SIGIR 2010.

[9] Pan, J., Zhang, B., Wang, S., Wu, G. and Wei, D. 2007. Ontology Based User Profiling in Personalized Information Service Agent. CIT 2007. Pp.1089-1093.

[10] Paramythis, A., König, F., Schwendtner, C. and and Velsen, L.V. 2010. Using Thematic Ontologies for User- and Group-Based Adaptive Personalization in Web Searching. Lecture

[11] Notes in Computer Science, 2010, Volume 5811/2010.

[12] Mittal, N., Nayak, R., Govil, M.C. and Jain, K.C. 2010.Evaluation of a hybrid approach of personalized web information retrieval using the FIRE data set. In A2CWiC 2010.India.

[13] Daoud, M., Tamine, L. and Boughanem, M. 2010. A Personalized Graph-Based Document Ranking Model Using a Semantic User Profile. In UMAP 2010.

[14] Bhogal, J., Macfarlane, A. and Smith, P. 2006. A review of ontology based query expansion. Inf. Process. Management.

[15] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter. Using odp metadata to personalize search. 2005. In Proceedings of SIGIR '05, pp. 178–185.

[16] Kim H., Chan P. 2003. Learning Implicit User Interest Hierarchy for Context in Personalization. In Proceedings of the 2003 International Conference on Intelligent user interfaces 2003, Miami, Florida.

[17] Porter, M. 1980. An algorithm for suffix stripping. Program 14,3, pp.130-137.

[18] Trajkova, J., Gauch, S., 2004. Improving ontology-based user profiles. In Proceedings of RIAO.

[19] Baeza, R. and Ribeiro, B. 1999. Modern Information Retrieval.Addison-Wesley.

[20] Grcar, M., Mladenic, D. and Grobelnik, M. 2005. User profiling for interest-focused browsing history. In SIKDD 2005 at Multiconference IS 2005, Ljubljana, Slovenia.