Indian Journal of Science and Technology, *Vol 7(10), 1712–1719, October 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Link-click-concept based Ranking Algorithm for Ranking Search Results

S. Geetha Rani[1*] and M. Sorana Mageswari[2]

[1]PSG College of Arts and Science, Coimbatore, Tamil Nadu, India; geetharani527@gmail.com
[2]PG and Research Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, India

Abstract

Personalized search is an essential research area that has main goal to determine the uncertainty of query terms. In order to enhance the relevance of search results, personalized search engines form user profiles which capture the users' personal preferences and by using those preferences find out the actual goal of the input query. By using User profile we can rank the documents in a search engine according to the query which is submitted by user. A better user profiling strategy is an important and primary component in search engine personalization. In this work, we propose a scheme that supports mining a user's conceptual preferences from users' click through data resulted from web search. This discovered preference is helpful to adapt a search engine's ranking function. In the system, an absolute set of conceptual preferences is derived for a user such as the concepts extracted from the search results and the click through data. After that, a Concept-based User Profile (CUP) in other words a concept ontology tree is generated. Our system proposes a novel approach such as Link-Click-Concept based Ranking Algorithm. The proposed system considers the concept for the user profile construction and has high efficiency than the existing system.

Keywords: Concept-based User Profile and Search Engine, Personalized Search, Ranking Algorithm, User profile

1. Introduction

At the present time the amount of information on the web is increasing rapidly, which has become increasingly harder for the web search engines to get the information that satisfies the user's individual interests. Commonly, the queries submitted by the user have lower length. This may lead to uncertainty for the search engine to recover the results for a particular user query. As a result, large numbers of retrieved results may not equivalent to the users' interests. Foremost commercial search engines offers query suggestions which is used for users to make more valuable queries and it is used to improve user's search experience. Whenever a user submitted a query, the meaningful related terms for a given query are offered to help the user recognize terms that they really want. Undoubtedly it will improve the efficiency of retrieval. Yahoo's "Also Try" and Google's "Searches related to" features offers correlated queries for narrowing search. Unluckily, these above systems give the same suggestions to the same query without taking into account users' specific interests.

Personalized search is an efficient way to develop quality of the search by make to ordering search results for people with various necessities. Numerous latest research efforts have paying attention on this area. Most of them could be classified into two common approaches: Re-ranking query results which are retrieved from search engines locally using personal information; or sending personal information and queries jointly to the search engine. Because users are generally unwilling to giving their preferences clearly owing to the extra manual effort is needed, current examine has paying attention on the automatic learning of user preferences from user's search histories or browsed documents and the development of personalized systems based on the learned user preferences. An enhanced user profiling strategy is an important and primary component in search engine personalization.

*Author for correspondence

The traditional click through-based user profiling schemes can be classified into two approaches. They are: A. Document-based approaches and B. Concept-based approaches.

These two approaches assume that user clicks is well used for identify the users' real interests. However their inference techniques and the outcomes from those inference techniques are different. Users' document preferences are approximated by Document-based profiling methods (i.e., users are interested in some documents more than others). Concept-based profiling methods has main goal as deriving the topics or concepts wherein users has highly interested.

Several user profiling approaches are only taking into account of the documents in which the users are interested in (i.e., users' positive preferences) but pay no attention to documents which are out of favour by the user (i.e., users' negative preferences). In actuality, not only positive preferences are sufficient to detain the fine interests of a user but also require negative preference as well.

Our main contribution of this work as follows:

- We propose novel methods that gain knowledge of concept preferences of user which is extracted directly from user click throughs in addition tousing of concept ontology. Such a way the concept preferences for user are extracted. This addresses the click sparsity problem which is not able to create sufficient concept preferences for training in search personalization.
- We proposed concept based user profile approaches can able to use the advantages of both document-based and concept-based approaches. The primary reason is that our system supports using click throughs to find out user document preferences via converting the document preferences into concept preferences through the concept ontology extraction method.
- We are introducing the natural language processing scheme such as Word Net which is used to derive the semantic information of concepts in the concept ontology.
- In addition, we are extracting the click through data from the multiple search engines such as Google, MSN, and Yahoo. We show that the user preference that is extracted from the concept ontology which is improves quality of the ranking and it is much better than existing techniques based on document features only.

1.1 Document-based Methods

Document-based user profiling techniques has important goal as to detain the users' clicking and browsing behaviours. In the beginning, users' document preferences are extracted from the click through data then it learns the behaviour model of user generally referred as users' behaviour model. In addition to this model, it is demonstrated as a set of weighted features. Concept-based user profiling approaches derives only conceptual needs of user that means the users' interests. User browsed documents and histories of search are characterized and mapped automatically. User profiles are built according to the users' preferences on the extracted topical categories. Mainly document-based approaches has centre of attention in the users' clicks and browsing behaviours. These behaviours are evaluated and stored in the users' click through data. Click through data are one of the significant understood feedback system on search engines in users opinion. With the purpose of capture users' interest several authors have been proposed click through data in [1–5] and these data were employed by several personalized systems.

1.2 Concept-based Methods

Generally concept-based techniques develop users' interest by discovering the contents of users search histories and browsed documents automatically. Several user profiling approaches that are based on users' search history and the Open Directory Project (ODP)[7] were proposed by Liu et al.[6]. In general, the user profile is corresponding to the set of categories. Additionally a set of keywords with weights is connected for each set of categories. These categories were accumulated in the user profiles and consequently it provide as a context for disambiguating users' query. When the users' interest in particular categories is exposed by profile, the search can be pointed down by providing some suggested results based on users' preferred categories.

Various user profiling approaches are only taking into account of the documents in which the users are interested in (i.e., users' positive preferences) but pay no attention to documents which are out of favour by the user (i.e., users' negative preferences). In actuality, not only positive preferences are sufficient to detain the fine interests of a user but also require negative preferences as well. Personalization system takes account of a negative preferences in[3–5] personalization approaches. However, these are all document-based approaches and that cannot give the users' topical interests.

Derived from the hyper link structure, Page Rank algorithm was proposed by Brin and Page[8] at Stanford University. For the most of web pages these ranking algorithms can be used repeatedly. During the processing of a query, search algorithm merges precalculated. As a result, the process of ranking can be completed by Page Rank. This score along with the text matching scores is used to gain an overall ranking score for each web page. Page Rank algorithm function is related to the link structure of the web pages. The concept of Page Rank algorithm is if a page surrounds an essential links on the way to it, then the links of this page near the other page are also to be assumed as imperative pages. The Rank score conclusion can be restricted on the back link of the Page Rank. When the addition of the ranks in the back links are high, then the page holds a high rank as well.

Preference mining and machine learning to model users' clicking and browsing behaviour are employed by a method, which was proposed by Joachims[3].Users' clicking and browsing behaviour are modelled by Machine learning and Preference mining. These models are employed by using a method, which was proposed by Joachims[3]. During query processing, the relations are lost and given keywords are treated as individual keywords, thus creating the major problem of isolated keyword matching. Though the ranking of the retrieved web pages has not accounted for relations, such that it is purely based on link analysis like i PageRank[8,9] and some on page relevance factors[10].

A combination of spying technique and novel voting procedure is employed for determining users' document preferences from the click through data by an algorithm, proposed by Ng et al.[4]. In order to learn the user behaviour model as a set of weight features, RSVM algorithm is also employed by them. More recently, explicit feedback (i.e., click through data, individual user behaviour etc.) from search engine users is noisy was suggested by Agichtein et al.[1] . In the following sections we proposed user profile strategies and ranking algorithm for inbound and outbound links and the relevancy of pages can be returned.

2. Link-click based Ranking Approach

In this research, propose a new algorithm to calculate the rank of pages for the submitted query. In this system, click count made by users of the page which is obtained by click through data and inbound outbound link of pages are used for calculating rank of pages.

Step 1: Calculate the number of click made on page by the user
Step 2: Calculate the number of inbound and outbound links in clicked page
Step 3: Combine step1 and Step2 to get the result.

Initially the total number of click can be counted and multiplied by its weight, the weight for each click is denoted by ω_1. Click count is denoted by C_x

$$\text{Total number of click} = \omega_1 * C_x \quad (1)$$

Similarly, Inbound and outbound link can be calculated. The weight of inbound link is ω_{in} and weight of outbound link is denoted as ω_{out}. The inbound and outbound link can be denoted as C_{in} and C_{out} respectively.

$$\text{Total number of inbound and outbound link} = (\omega_{in} * C_{in}) + (\omega_{out} * C_{out}) \quad (2)$$

By combining above two Equations the result is as follows

$$(\omega_1 * C_x) + ((\omega_{in} * C_{in}) + (\omega_{out} * C_{out})) \quad (3)$$

Ranking of pages can be made done by checking the relevancy between the User profiles based page ranking and the user profile based click made on the page.

3. Proposed Methodology

We propose the novel approach called link-click-concept based user profile strategy which is combines the link based, click based and concept based user profile approaches. The concepts are extracted by using the ontology. Thus the concept ontology is constructed. To discover further the benefits of concept ontology, here we propose a innovative *Concept-based User Profiling* (*CUP*) technique which is used to capture users' topical preferences. This concept ontology is built by using the Concept Extraction. In the Concept Extraction step, while a query is given, significant concepts and their relations are mined online from web-snippets to construct concept *ontology*.

In general, concept ontology (or simply ontology) can be regarded as the formal representation of a set of concepts within a domain. In general web searching, ontology is defined as the formal demonstration of a set of concepts within the search results as well as relationship between the concepts. In the ontology, two types of relationships are there. They are similarity and Parent-Child Relationship, which is working to represent

the relations of the extracted concepts for constructing the Concept based user profile. Correspondingly, the semantic information of concepts is extracted by using the natural language processing scheme. We used Word Net as the NLP technique. By using this Word Net tool we can derive the semantic information for all concepts which are extracted from the concept ontology. In addition, we are extracting the click through data from the multiple search engines such as Google, MSN, and Yahoo. Finally, our proposed system is well effective than the existing system by incorporating these approaches. Based on this, the efficiency and performance of the system is improved.

3.1 Link-Click-Concept based Ranking Approach

In this section, we propose four user profiling strategies which are both concept-based approach and utilize users' positive and negative preferences.

3.1.1 Link-click-concept based Method ($P_{(Link-Click-Concept)}$)

The concepts extracted for a query q using the concept ontology. It captures both positive as well as negative preferences. Therefore, we propose the following formulas to capture a user's degree of interest ω_{c_i} on the extracted concepts c_c using the ontology, when a Web-snippet s_j is clicked by the user. (click(s_j))

$$\text{click}(s_j) + \text{Link}(s_j) \Rightarrow \forall c_{i,\text{in, out}} \in S_j,$$

$$\omega(C + C_c)_{ci} = \omega_{c_i} + \omega(C_c)_{ci+1}, \tag{4}$$

$$\text{click}(s_j) + \text{Link}(s_j) \Rightarrow \forall c_{i,\text{in,out}} \in S_j$$

$$\omega(C + C_c)_{Cj} = \omega_{c_j} + \omega(C_c)_{Ci} + \text{sim}_R(c_i, c_j) \text{ if } \text{sim}_R(c_i, c_j) > 0 \tag{5}$$

where s_j is a web – snippet, in and out describes the inbound and outbound links. $\omega(C + C_c)_{ci}$ is a users degree of interest on the concept c_i, and c_j is the neighbourhood concept of c_i.

When a Web-snippet s_j has been clicked by a user, the weight $\omega(C_c)_{ci}$ of concepts c_i appearing in s_j is incremented by 1. For other concepts c_j that are related to c_i on the concept relationship graph, they are incremented according to the similarity score given in the above equation. The similarity between the concepts is calculated with the help of natural language processing (NLP) techniques such as Word Net. This similarity is referred as semantic similarity. This novel technique is used to upgrade the performance of the system.

3.1.2 Link-Click-Concept+Joachims-CMethod ($P_{(Link-Click-Concept+Joachims-C)}$)

In this section we incorporated the click-based method, Joachims-C method as well as Concept based approach. Click-based method captures only positive preferences, while Joachims-C method captures negative preferences. In addition our concept based method is derives the concept from the ontology.

Since both the user profiles, click-based P_{Click} and Joachims-C method $P_{Joachims-C}$ are represented as weighted concept vectors, a hybrid method Link-Click-Concept+Joachims-C Method is proposed. It is combines $P_{Link-Click-Concept}$ and $P_{Joachims-C}$. These two profiles are combined using the following formula:

$$\omega(C + J + C_c)_{Ci+\text{in, out}} = \begin{cases} \omega(C)_{Ci+\text{in,out}} + \omega(j)_{Ci+\text{in,out}} + \omega(C_c)_{Ci+\text{in,out}} \\ \quad if\ \omega(J)_{Ci+\text{in,out}} < 0 \\ \omega(C)_{Ci+\text{in,out}} + \omega(C_c)_{Ci+\text{in,out}} \\ \quad otherwise \end{cases} \tag{6}$$

Where

$$\omega(C + J + C_c)_{Ci+\text{in, out}} \in P_{Link-Click-Concept+Joachims-C}$$

$$\omega(C)_{Ci} \in P_{Link\ and\ Click}$$

$$\omega(J)_{Ci+\text{in, out}} \in P_{Joachims-C}$$

$$\omega(C_c)_{Ci+\text{in, out}} \in P_{Link\ and\ Concept}$$

If a concept c_i and Link(in and out) has a negative weight in $P_{Joachims-c}$ (i.e. $\omega(J)_{Ci+\text{in, out}} < 0$) the negative weight will be added to $\omega(C)_{Ci+\text{in, out}}$ in $P_{Link\ and\ Click}$ and $\omega(C_c)_{Ci+\text{in, out}}$ in $P_{Link\ and\ Concept}$ forming the weighted concept vector for the hybrid profile $P_{Link-Click-Concept+Joachims-c}$.

3.1.3 Link-Click-Concept+mJoachims-CMethod ($P_{(Link\ and\ Click+mjoachims-C)}$)

Similar to Link-Click-Concept+Joachims-C method, a hybrid method Link-Click-Concept+mJoachims-C Method is proposed. It combines $P_{Link-Click-Concept}$ and

$P_{mjoachims-C}$. These two profiles are combined using the following formula:

$$\omega(C+mJ+C_c)_{Ci+in,out} = \begin{cases} \omega(C)_{Ci+in,out}+\omega(mj)_{Ci+in,out}+\omega(C_c)_{Ci+in,ou} \\ \quad if\ \omega(mJ)_{Ci+in,out}<0 \\ \omega(C)_{Ci+in,out}+\omega(C_c)_{Ci+in,out} \\ \quad otherwise \end{cases} \quad (7)$$

Where

$$\omega(C+mJ+C_c)_{Ci+in,out} \in P_{Link-Click-Concept+mJoachims-C}$$

$$\omega(C)_{Ci} \in P_{Link\ and\ Click}$$

$$\omega(mJ)_{Ci+in,out} \in P_{mJoachims-C}$$

$$\omega(C_c)_{Ci+in,out} \in P_{Link\ and\ Concept}$$

If a concept c_i and Link(in and out) has a negative weight in $P_{mJoachims-C}$ (i.e.. $\omega(mJ)_{Ci+in,out}<0$) the negative weight will be added to $\omega(C)_{Ci+in,out}$ in $P_{Link\ and\ Click}$ and $\omega(C_c)_{Ci+in,out}$ in $P_{Link\ and\ Concept}$ forming the weighted concept vector for the hybrid profile $P_{Link-Click-Concept+mJoachims-C}$.

3.1.4 Link-Click-Concept+SpyNB-CMethod ($P_{(Link-Click-Concept+SpyNB-C)}$)

Similar to Link-Click-Concept+Joachims-C and Link-Click-Concept+mJoachims-C methods, the following formula is used to create a hybrid profile Link-Click-Concept+SpyNB-C Method $P_{Link-Click-Concept+SpyNB-C}$ that combines $P_{Link-Click-Concept}$ and $P_{spyNB-C}$:

$$\omega(C+sNB+C_c)_{Ci+in,out} = \begin{cases} \omega(C)_{Ci+in,out}+\omega(sNB)_{Ci+in,out}+\omega(C_c)_{Ci+in,out} \\ \quad if\ \omega(sNB)_{Ci+in,out}<0 \\ \omega(C)_{Ci+in,out}+\omega(C_c)_{Ci+in,out} \\ \quad otherwise \end{cases} \quad (8)$$

Where

$$\omega(C+sNB+C_c)_{Ci+in,out} \in P_{Link-Click-Concept+spyNB-C}$$

$$\omega(C)_{Ci} \in P_{Link\ and\ Click}$$

$$\omega(sNB)_{Ci+in,out} \in P_{spyNB-C}$$

$$\omega(C_c)_{Ci+in,out} \in P_{Link\ and\ Concept}$$

If a concept c_i and Link(in and out) has a negative weight in $P_{spyNB-C}$ (i.e.. $\omega(sNB)_{Ci+in,out}<0$) the negative weight will be added to $\omega(C)_{Ci+in,out}$ in $P_{Link\ and\ Click}$ and $\omega(C_c)_{Ci+in,out}$ in $P_{Link\ and\ Concept}$ forming the weighted concept vector for the hybrid profile $P_{Link-Click-Concept+spyNB-C}$.

3.2 Link-click-concept based Algorithm

In this research we propose a new algorithm to calculate the rank of pages for the submitted query. In this approach for calculating rank of pages we use click count made by users of the page and inbound outbound link of pages.

Step 1: Calculate the number of click made on page by the user
Step 2: Calculate the number of inbound and outbound links in clicked page
Step 3: Calculate the concept extracted from the ontology
Step 4: Combine Step 1, 2 & 3 to get the result.

Initially the total number of click can be counted and multiplied by its weight, the weight for each click is denoted by ω_1. Click count is denoted by C_x

$$\text{Total number of click} = \omega_1 * C_x \quad (9)$$

Similarly, Inbound and outbound link can be calculated. The weight of inbound link is ω_{in} and weight of outbound link is denoted as ω_{out}. The inbound and outbound link can be denoted as C_{in} and C_{out} respectively.

$$\text{Total number of inbound and outbound link} = (\omega_{in} * C_{in}) + (\omega_{out} * C_{out}) \quad (10)$$

The concepts are extracted by using ontology with the help of user click through data. It is defined as,

$$\text{Total number of click} = \omega_c * C_c \quad (11)$$

By combining above three equations the result is as follows

$$(\omega_1 * C_x) + (\omega_{in} * C_{in}) + (\omega_{out} * C_{out}) + (\omega_c * C_c) \quad (12)$$

Ranking of pages can be made done by checking the relevancy between the User profiles based on our proposed ranking algorithm and the user profile based click made on the page.

4. Experimental Results

In this research, we have analyzed the three concept based user profiling strategies. These approaches are ranked with proposed Link-Click-Concept based ranking algorithm and existing Link and click based ranking algorithm. This algorithm mainly deals with the concept of when the submitted query does not give expected result then the links returned by the given query gives out the best result. Experimental result showed better result by using this proposed algorithm when compared to Link and click based ranking algorithm. The relevancy measure can be applied by using Precision and recall method as follows:

4.1 Precision

Precision, also called positive predictive value, is the fraction of retrieved instances that are relevant to the search. In general it is defined as:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False Positive}} \tag{13}$$

Precision takes all retrieved pages into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system.

$$\text{Precision} = \frac{|\{relevantpages\} \cap \{retrievedpages\}|}{|\{retrievedpages\}|} \tag{14}$$

In Figure 1, the precision values are compared between the LinkandClick+Joachims-Cmethod, Link and Click+mJoachims-Cmethod and Link and Click+SpyNB-Cmethod according to the number of queries. In this graph, x axis will be the number of queries and y axis will be precision rate. When the numbers of queries are increased then the precision rate decreased. This existing system has the lower precision rate compared to the proposed Link-Click-Concept based ranking method.

In Figure 2, the precision values are compared between the Link-Click-Concept+Joachims-Cmethod, Link-Click-Concept+mJoachims-CmethodandLink-Click-Concept+SpyNB-Cmethod according to the number of queries. In this graph, x axis will be the number of queries and y axis will be precision rate. When the numbers of queries are increased then the precision rate is decreased. The proposed system has the highest precision rate compared to the existing Link-Click based ranking method.

4.2 Recall

Recall value is calculated is based on the retrieval of information at true positive prediction, false negative. In general it is defined as,

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \tag{15}$$

Recall is also known as sensitivity. Recall in information retrieval is the fraction of the pages that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{|\{relevantpages\} \cap \{retrievedpages\}|}{|\{relevantpages\}|} \tag{16}$$

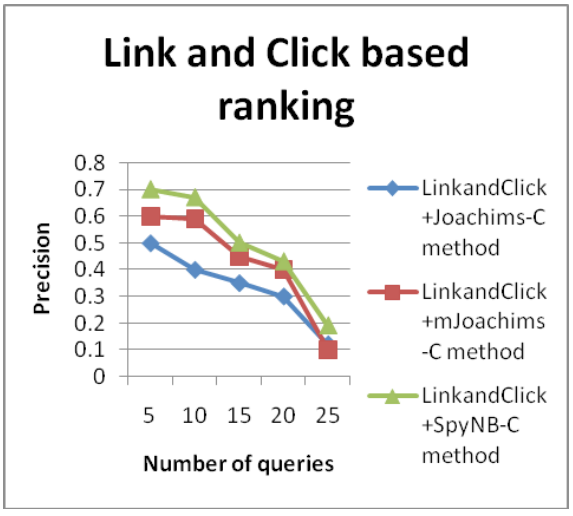

Figure 1. Precision comparison for link-click.

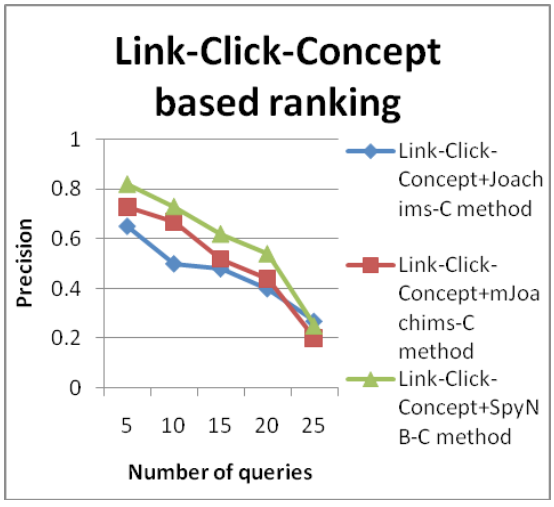

Figure 2. Precision comparison for link-click-concept.

In Figure 3, the recall values are compared between the LinkandClick+Joachims-Cmethod, LinkandClick+mJoachims-Cmethod and LinkandClick+SpyNB-Cmethod according to the number of queries. In this graph, x axis will be the number of queries and y axis will be recall rate. When the numbers of queries are increased then the recall rate is decreased. This existing system has the lower recall rate compared to the proposed Link-Click-Concept based ranking method.

In Figure 4, the recall values are compared between the Link-Click-Concept+Joachims-Cmethod, Link-Click-Concept+mJoachims-Cmethod and Link-Click-Concept+SpyNB-Cmethod according to the number of queries. In this graph, x axis will be the number of queries and y axis will be recall rate. When the numbers of queries are increased then the recall rate is decreased. The proposed system has the highest recall rate compared to the existing Link-Click based ranking method.

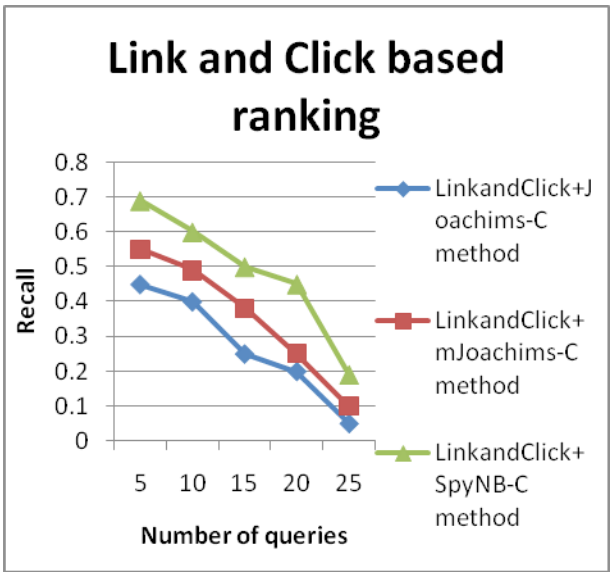

Figure 3. Recall comparison for link-click.

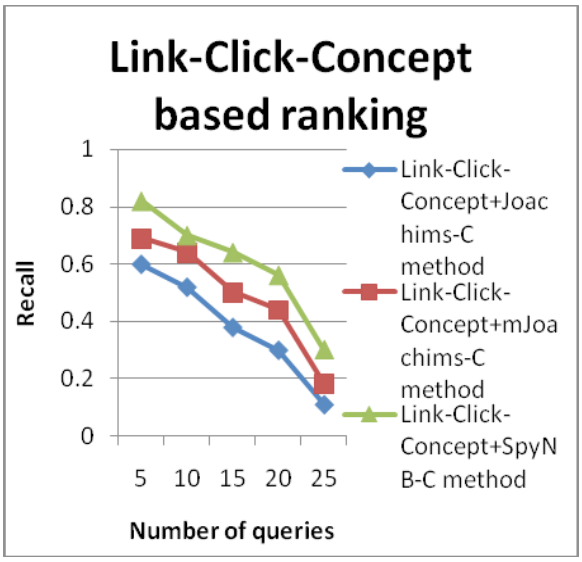

Figure 4. Recall comparison for link-click-concept.

5. Conclusion

Search engines performance can be improved by an accurate user profiles. This can be done by identifying the information which is exactly needed for individual users. In this research, we proposed and evaluated ranking algorithm. Our proposed CUP profiling strategy captures and organizes users' topical preferences in concept ontology. The ontology provides us to understand rich user concept preferences as well as those concepts are derived directly from user click through data. Based on Link and click & concept based approach, we proposed four user profile strategies. These strategies are ranked by two approaches. One is our proposed algorithm which is called Link-Click-Concept based algorithm and the existing one is Link and Click based algorithm. These four user profile approaches are evaluated and ranked by using these two techniques. Our results confirm that the CUP framework is able to accurately capture the users' topical preferences and outperforms existing methods. Finally, our proposed algorithm worked well and returned maximum value when compared with the existing Link-Click-Concept based approach.

In future, temporal dynamics of user's behaviour can be considered to rank search results to improve its accuracy.

6. References

1. Agichtein E, Brill E, Dumais S. Improving Web Search Ranking by Incorporating User Behaviour Information. Proc ACM SIGIR; 2006.
2. Agichtein E, Brill E, Dumais S, Ragno R. Learning User Interaction Models for Predicting Web Search Result Preferences. Proc. ACM SIGIR; 2006.
3. Joachims T. Optimizing Search Engines Using Clickthrough Data. Proc. ACM SIGKDD; 2002.
4. Ng W, Deng L, Lee DL. Mining User Preference Using Spy Voting for Search Engine Personalization. ACM Trans Internet Technology. 2007; 7(4).

5. Tan Q, Chai X, Ng W, Lee D. Applying Co-training to Clickthrough Data for Search Engine Adaptation. Proc Database Systems for Advanced Applications (DASFAA) Conf; 2004.
6. Liu F, Yu C, Meng W. Personalized Web Search by Mapping User Queries to Categories. Proc Int'l Conf Information and Knowledge Management (CIKM); 2002.
7. Open Directory Project; 2009. Available from: www.dmoz.org
8. Brin S, Page L. The anatomy of a large scale hypertextual web search engine. Computer Network and ISDN Systems. 1998; 30(1–7):107–17.
9. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project; 1998.
10. Search engine ranking variables and algorithms. Available from: www.SEMJ.org