A New Ranking Algorithm for Ranking Search Results of Search Engine based on Personalized User Profile

S.Geetha rani, MCA, M. Phil
Assistant professor
PSG College of Arts and Science
Coimbatore, Tamil nadu, India

ABSTRACT

The user profile is a main component in personalization applications. An accurate user profile can greatly improve a search engine's performance by identifying the information needs of individual users. The desired information can be obtained by submitting the respective query. Different query gives different information. The information which is more relevant for the given query can be analyzed and evaluated. The user profile is used to rank the documents in a search engine for a submitted query. Many user profiling strategies based on positive preferences preferences (i.e., Objects that users are interested in). Later some user profiling strategies were based on both positive as well as negative preferences (i.e., objects that users dislike).In Existing research only the count of a click can be evaluated. In this research, click count as well as Link-Click based Ranking Algorithm is being proposed. In this Algorithm the count of click of each query concept can be evaluated as well as the link also evaluated in the submitted query for three user profiling strategies. The relevance between the query and information obtained can be analyzed, evaluated and ranked. The goal of the proposed ranking approach is providing the user with more satisfied results to get relevant information based on Link and Click approach rather than Click count.

General Terms

Search Engine, Data mining

Keywords

User profile strategies, Search engine, Link and Click–based Ranking Algorithm.

1. INTRODUCTION

Search Engines are developed to help the users to search for what they are looking for. The information what the users are looking are found in the web data repository. Search Engines work with the help of predefined automated software programs (Spiders or bots and are used for crawling,) that is, going through the entire website (all the pages) and then records down the content in the form of databases , which is known as indexes.

In most commercial search engine, the user gets the same result returned for the same query regardless of their real interest. Since queries given by the user are short and ambiguous, search engines are unable to express the user's precise needs.

In existing click through-based user profiling strategies can be categorized into two approaches. They are

I. Document-based approaches

II. Concept-based approaches.

These both approaches assume that user clicks can be used to infer users' real interests, but their inference methods and the outcomes of such inference are different. Users' document preferences are estimated by Document-based profiling methods (i.e., Users are interested in some documents more than others)

Concept-based profiling methods goal is to derive topics or concepts in which users are highly interested.

These two approaches will be reviewed in Section 2.

Most existing user profiling strategies only consider the documents in which the users are interested in (i.e., Users' positive preferences) but ignore documents which are disliked by the user (i.e., Users' negative preferences). In reality, not only positive preferences are enough to capture the fine interests of a user but also need negative preference as well.

The proposed work in this research is:

- Concept-based user profiling method is used to consider both users' positive and negative preferences in building user profiles.
- Three user profiling methods that exploit a user's positive and negative preferences to produce a profile for the user by using a Link and click based Ranking is proposed.

We proposed Link and click based ranking algorithm to rank the pages for the submitted query .Ranking through this algorithm is based upon a probability measure of Precision and Recall. This relevancy can be checked for the queries submitted by the user. This checks the relevancy of a page to the query.

2. RELATED WORK

User profiling strategies can be broadly classified into two main approaches: They are

- Document-based approach
- Concept-based approach.

2.1 Document-Based Methods

Document-based user profiling methods goal is to capture the users' clicking and browsing behaviors. Initially click through data extracts Users' document preferences then it learns the behavior model of user commonly called as users' behavior model. And this model is represented as a set of weighted

features. Concept-based user profiling methods capture only conceptual needs of the user. Browsed documents and search histories of users are categorized and mapped automatically. User profiles are created by depends upon users' preferences on the extracted topical categories. Most document-based methods focus users' clicks and browsing behaviors. These behaviors are analyzed and recorded in the users' click through data. Click through data are one of the important implicit feedback mechanism on search engines in users' point of view. In order to capture the users' interest we have been proposed click through data in [1] [2] [3] [4] [5] and these data were employed by several personalized systems.

2.2 Concept-Based Methods

Most concepts-based method derives users' interest by exploring the contents of users search histories and browsed documents automatically.

Some user profiling method which is based on users' search history and the Open Directory Project (ODP) [7] were proposed by Liu et al. [6].

The user profile is represented as like set of categories. And a set of keywords with weights is associated for each set of categories. These categories were stored in the user profiles and thus it serves as a context for disambiguating users' query. Once the users' interest in certain categories is shown in profile, the search can be narrowed down by giving some suggested results based on users' preferred categories.

Most existing user profiling strategies only consider the documents in which the users are interested in (i.e., Users' positive preferences) but ignore documents which are disliked by the user (i.e., Users' negative preferences). In reality, not only positive preferences are enough to capture the fine interests of a user but also need negative preference as well.

Personalization process includes negative preferences in [3], [4], [5] personalization strategies. But these are all document-based methods, and it cannot return users' general topical interests.

Based on the hyperlink structure, PageRank algorithm was developed by Brin and Page [8] at Stanford University. A famous search engine, Google is being used by the PageRank algorithm. In ranking millions and billions of web pages these ranking algorithms can be used frequently. During the processing of a query, search algorithm combines precomputed.

Thus the process of ranking can be done by PageRank, which scores with the text matching scores to obtain an overall ranking score for each web page.

Page Rank algorithm function depends upon the link structure of the web pages. The concept of PageRank algorithm is if a page surrounds an important link towards it, then the links on this page near the other page are also to be believed as imperative pages. The Rank score decision can be limited on the back link of the PageRank. When the addition to the ranks in the back links is high , then the page holds a high rank as well.

Preference mining and machine learning to model users' clicking and browsing behavior are employed by a method, which was proposed by Joachims [3].

Users' clicking and browsing behavior 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 [9] [10] and some on page relevance factors [11] .

A combination of spying technique and novel voting procedure is employed for determining users' document preferences from the clickthrough data by an algorithm , proposed by Ng et al. [4]. In order to learn the user behavior model as a set of weight features, RSVM algorithm is also employed by them.

More recently, explicit feedback (i.e., clickthrough data, individual user behavior 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.

.3. USER PROFILING STRATEGIES

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

3.1LinkandClick-Based Method (P_{Click})and (P_{Link})

The concepts extracted for a query q using the concept extraction. It captures only positive preferences.

Therefore, we propose the following formulas to capture a user's degree of interest ω_{c_i} On the extracted concepts c_i,when a Web-snippet s_j is clicked by the user.(click(s_j))

click (s_j) + Link(s_j) => $\forall c_i$, $link(in, out)$ ϵs_j , ω_{c_i}= ω_{c_i}+1, ---(1)

click (s_j) + Link(s_j) => $\forall c_i$, $link(in, out)\epsilon s_j$,

$\omega_j = \omega_{c_j} + sim_R(c_i , c_j)$ if $sim_R(c_i , c_j) > 0$ ----(2)

where s_j is a web − snippet,in and out describes the inbound and outbound links.

ω_{c_i} is a users degree of interest on the concept c_i, and c_j is the neighborhood concept of c_i.

When a Web-snippet s_j has been clicked by a user, the weight ω_{c_i} 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(2).

3.2LinkandClick+Joachims-CMethod ($P_{Link\ and\ Click+Joachims-C}$)

In this research we integrated Link and click-based method and Joachims-C method.

Click-based method captures only positive preferences, while Joachims-C method captures negative preferences. In the proposed work we found that Joachims-C is best in predicting users' negative preferences.

Since both the user profiles, click-based P_{Click} and Joachims-C method $P_{Joac\ hims-C}$ are represented as weighted concept vectors,these two vectors can be combined using the following formula:

$\omega(\mathrm{C+J})_{\mathrm{Ci+link(in,out)}} = \omega(C)_{\mathrm{Ci+link(in,out)}} + \omega(J)_{\mathrm{Ci'+link(in,out)}}$ if $\omega(J)_{\mathrm{Ci+link(in,out)}} < 0$,

$\omega(C+J)_{\mathrm{Ci+link(in,out)}} = \omega(C)_{\mathrm{Ci'+link(in,out)}}$ otherwise

Where $\omega(C+J)_{\mathrm{Ci+link(in,out)}} \in P_{Link\ andClick\ +Joac\ hims-C}$,$\omega(C)_{\mathrm{Ci+link(in,out)}} \in P_{Link\ andClick}$, $\omega(J)_{\mathrm{Ci+link(in,out)}} \in P_{Link\ andJoac\ hims-C}$

If a concept c_i and Link of inbound and outbound link(in and out) has a negative weight in $P_{Joac\ hims-C}$ (i.e.. $\omega(J)_{\mathrm{Ci+link(in,out)}} < 0$) then negative weight will be added to $\omega(C)_{\mathrm{Ci+link(in,out}}$ in P_{Click} (i.e.,$\omega(C)_{\mathrm{Ci+link(in,out)}} + \omega(J)_{\mathrm{Ci+link(in,out)}}$) and forms the weighted concept vector for the hybrid profile $P_{Link\ and\ Click\ +Joac\ hims-C}$.

3.3LinkandClick+mJoachims-CMethod ($P_{LinkandClick+mJoachims-C}$)

Similar to LinkandClick+Joachims-C method, a hybrid method LinkandClick+mJoachims-C Method**,** which combines $P_{LinkandClick}$ and $P_{mJoac\ hims-C}$ is proposed. These two profiles are combined using the following formula:

$\omega(\mathrm{C+mJ})_{\mathrm{Ci+link(in,out)}} = \omega(C)_{\mathrm{Ci+link(in,out)}} + \omega(mJ)_{\mathrm{Ci'+link(in,out)}}$ if $\omega(J)_{\mathrm{Ci+link(in,out)}} < 0$,

$\omega(C+mJ)_{\mathrm{Ci+link(in,out)}} = \omega(C)_{\mathrm{Ci'+link(in,out)}}$ otherwise

Where $\omega(C+mJ)_{\mathrm{Ci+link(in,out)}} \in P_{LinkandClick\ +mJoac\ hims-C}$, $\omega(C)_{\mathrm{Ci,+link(in,out)}} \in P_{LinkandClick}$, $\omega(mJ)_{\mathrm{Ci+link(in,out)}} \in P_{mJoac\ hims-C}$

If a concept c_i and Link of inbound and outbound link (in and out) has a negative weight in $P_{mJoac\ hims-C}$ (i.e.. $\omega(mJ)_{\mathrm{Ci,+link(in,out)}} < 0$) the negative weight will be added to $\omega(C)_{\mathrm{Ci+link(in,out)}}$ in $P_{LinkandClick}$ (i.e., $\omega(C)_{\mathrm{Ci+link(in,out)}} + \omega(mJ)_{\mathrm{Ci+link(in,out)}}$) forming the weighted concept vector for the hybrid profile $P_{LinkandClick\ +mJoac\ hims-C}$.

3.4LinkandClick+SpyNB-CMethod ($P_{LinkandClick+SpyNB-C}$)

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

$\omega(\mathrm{C+sNB})_{\mathrm{Ci+link(in,out)}} = \omega(C)_{\mathrm{Ci+link(in,out)}} + \omega(sNB)_{\mathrm{Ci'+link(in,out)}}$

if $\omega(sNB)_{\mathrm{Ci+link(in,out)}} < 0$,

$\omega(C+sNB)_{\mathrm{Ci+link(in,out)}} = \omega(C)_{\mathrm{Ci'+link(in,out)}}$ otherwise

Where $\omega(\mathrm{C+sNB})_{\mathrm{Ci+link(in,out)}} \in P_{LinkandClick\ +SpyNB-C}$, , $\omega(C)_{\mathrm{Ci+link(in,out)}} \in P_{LinkandClick}$ and$\omega(sNB)_{\mathrm{Ci+link(in,out)}} \in P_{SpyNB-C}$:

If a concept c_i and Link of inbound and outbound link (in and out) has a negative weight in $P_{SpyNB-C}$ (i.e., (i.e., $\omega(sNB)_{\mathrm{Ci+link(in,out)}} < 0$) the negative weight will be added to $\omega(\mathrm{C})_{\mathrm{Ci+link(in,out)}}$ in $P_{LinkandClick}$ (i.e., $\omega(C)_{\mathrm{Ci+link(in,out)}} + \omega(sNB)_{\mathrm{Ci+link(in,out)}}$) forming the weighted concept vector for the hybrid profile $P_{LinkandClick\ +SpyNB-C}$.

3.5Link and Click 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 a click count made by users of the page and inbound outbound link of pages.

Step1: Calculate the number of clicks made on the page by the user

Step2: Calculate the number of inbound and outbound links in clicked page

Step3: Combine step1 and Step2 to get the result.

Initially the total number of clicks 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

Total number of clicks = $\omega_1 * C_x$ --→(3)

Similarly, Inbound and outbound link can be calculated .The weight of inbound link is ω_{in}

And the weight of outbound link is denoted as ω_{out}. The inbound and outbound link can be denoted as C_{in} and C_{out} respectively.

Total number of inbound and outbound links = $(\omega_{in} * C_{in}) + (\omega_{out} * C_{out})$ -→(4)

By combining Equation (3) and (4) the result is as follows

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

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

This relevancy dependence can be found by Precision and recall method.

4 EXPERIMENTAL RESULTS

In this research, the three concepts based user profiling strategies were evaluated and analyzed. These strategies are ranked with Link and click based ranking algorithm as well as with a click-count ranking approach. This algorithm mainly deals with the concept of when the submitted query does not give the expected result then the links returned by the given query gives out the best result .Experimental results showed a better result by using this proposed algorithm against Click-Count. The relevant measure can be applied by using Precision and recall method as follows.

Precision:

Precision is also called positive predictive value is the fraction of retrieved instances that are relevant to the search.

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

$$\text{Precision} = \frac{|\{relevant\ \ pages\} \cap \{retrieved\ \ pages\}|}{|\{retrieved\ \ pages\}|}$$

Recall:

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.

$$\text{Recall} = \frac{|\{relevant\ \ pages\} \cap \{retrieved\ \ pages\}|}{|\{relevant\ \ pages\}|}$$

The graph can be plotted to show the relevancy below:

The following graph shows the result for both Link and Click based Ranking and Click-Count based method

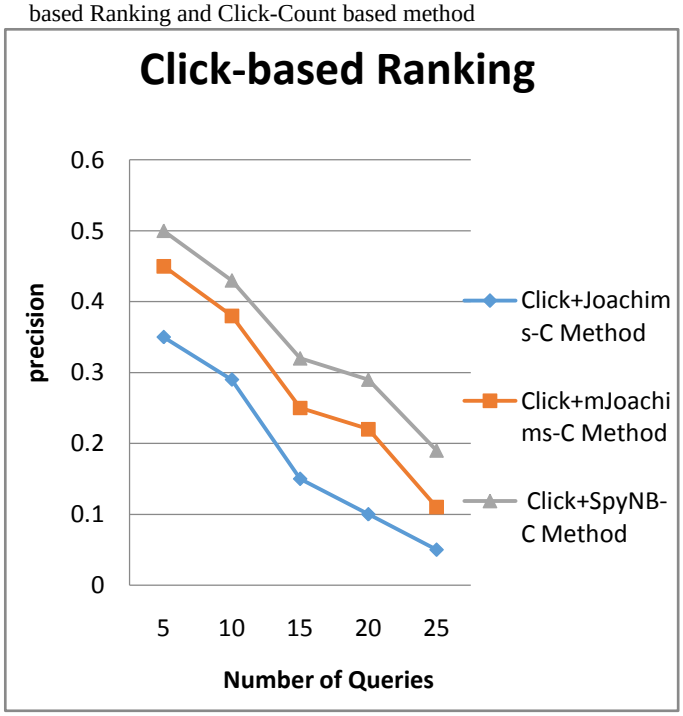

Figure 1:Number of Query Vs Precision

The above graph shows that the number of queries is compared with precision.

These three strategies returned minimum level value when comparing with the proposed algorithm.

When comparing (Figure1) click based approach to (Figure2) Link and Click based, The Precision value gets increasing for the number of queries . In Click based approach, The precision value obtained for Strategies Click+Joachims-C, Click+mJoachims-C Method and Click+SpyNB-C Method for a query range 25 is 0.35, 0.43, 0.5 respectively. In Link and Click based approach, The precision value obtained for Strategies LinkandClick+Joachims-C, LinkandClick+mJoachims-C Method and LinkandClick+SpyNB-C Method for a query range 25 is0.5,0.6,0.69 in the following graph. Thus the proposed method shows high precision value than existing click-based method.

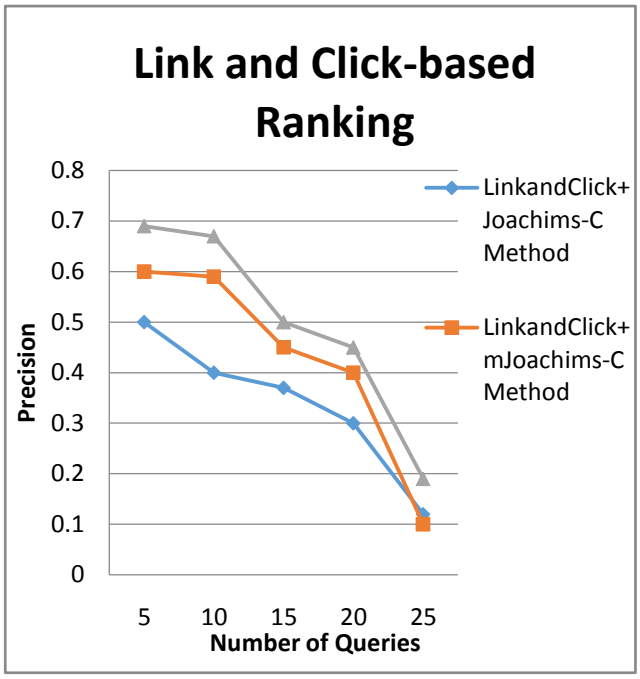

Figure 2:Number of Query Vs Precision

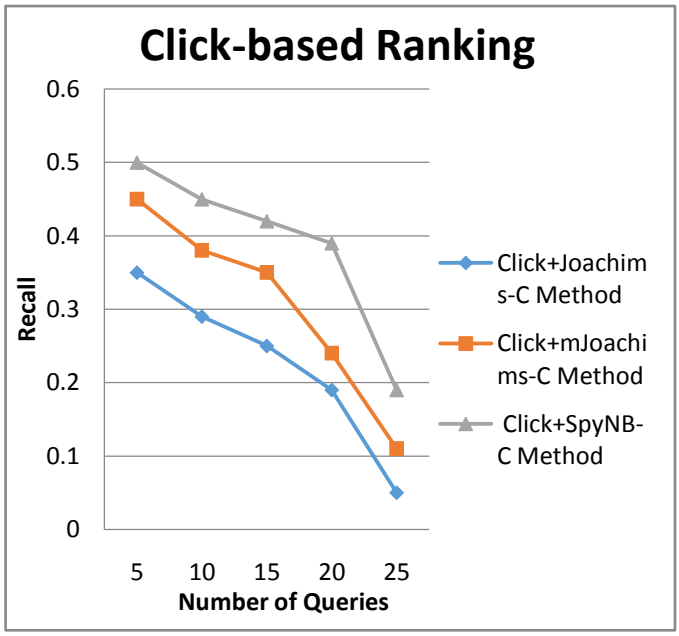

Figure 3:Number of Query Vs Recall

The above graph shows that the numbers of queries are compared with Recall.

The value returned by the above strategies are minimum when compare with the proposed Algorithm.

When comparing (Figure))click based approach to ((Figure4) Link and Click based, The Recall value gets increasing for the number of queries. In Click based approach, The Recall value obtained for Strategies Click+Joachims-C, Click+mJoachims-C Method and Click+SpyNB-C Method for a query range 25 is 0.35,0.45,0.5 respectively. In Link and Click based approach, The Recall value obtained for Strategies LinkandClick+Joachims-C, LinkandClick+mJoachims-C Method and LinkandClick+SpyNB-C Method for a query range 25 is0.45,0.55,0.7 respectively. Thus the proposed method shows high Recall value than existing click-based method.

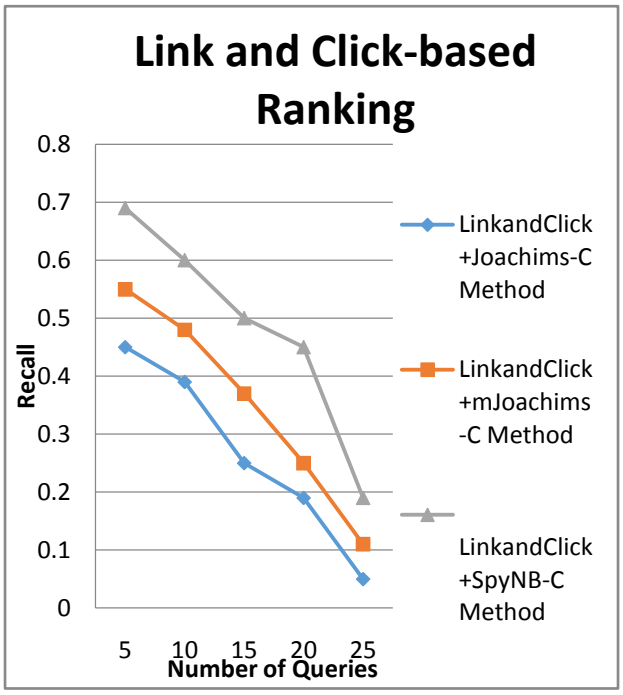

Figure 4: Number of Query Vs Recall

The above graph shows the result returned by the proposed algorithm by plotting the number of queries against Recall value.

The value obtained by using the proposed algorithm gives out maximum value when compare with the Click-Count based method.

5. CONCLUSION AND FUTURE WORK

Search engine performance can be improved by an accurate user profile. This can be done by identifying the information which is exactly needed for individual users. In this research, the ranking algorithm is evaluated and proposed. Based on a Link and click approach, three user profile strategies are used. These strategies are ranked by two approaches. One is proposed algorithm which is called Link and Click based algorithm and the other one is Click-Count based approach. The relevant measure has taken by using Precision and Recall method. The three user profile strategies are evaluated and ranked by using these two approaches. Finally, the proposed algorithm worked well and returned maximum value when compared with the Click-count approach.

In future, the work can be done by considering temporal dynamics of user's behavior to rank search results, it will improve the accuracy of search results.

6. REFERENCE

[1] E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," Proc. ACM SIGIR, 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning User Interaction Models for Predicting Web Search Result Preferences," Proc. ACM SIGIR, 2006.

[3] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. ACM SIGKDD, 2002.

[4] W. Ng, L. Deng, and D.L. Lee, "Mining User Preference Using Spy Voting for Search Engine Personalization," ACM Trans. Internet Technology, vol. 7, no. 4, article 19, 2007.

[5] Q. Tan, X. Chai, W. Ng, and D. Lee, "Applying Co-training to Clickthrough Data for Search Engine Adaptation," Proc. Database Systems for Advanced Applications (DASFAA) Conf., 2004.

[6] F. Liu, C. Yu, and W. Meng, "Personalized Web Search by Mapping User Queries to Categories," Proc. Int'l Conf. Information and Knowledge Management (CIKM), 2002.

[7] Open Directory Project, http://www.dmoz.org/, 2009.

[8] S. Brin, and Page L., "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network .

[9] ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Sergey Brin and Lawrence Page.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford Digital Library Technologies Project, 1998.

[11] Search Engine Ranking Variables and Algorithms. Sean A. Golliher – Publisher, SEMJ.org

AUTHOR PROFILE

S.Geetha rani completed her MCA in 1996 at Bharathiyar University and completed her M.phil in Data mining in 2006 at Bharathiyar University. Currently she is working as a Assistant Professor in PSG College of Arts and Science. She has successfully submitted Two National and International Journals.She guided more than 15 M.phil students .She has a 12 years of teaching experience in her career. Her Research area is in Data mining.