A NOVEL APPROACH TO PAGE RANKING MECHANISM BASED ON USER INTEREST

Ankur Mittal[1], Shilpa Sethi[2]

[1]PG Scholar, [2]Assistant Professor, Department of Computer Engineering,
YMCA University of Science & technology, Faridabad, Haryana, (India)

ABSTRACT

Information searching on the web is more popular than the traditional method but is not an easy task on the web because the World Wide Web (WWW) is collection of huge no of document which available on internet. Thus a tool is required called search engine which helps in find the required information for user. Search engines are provide the list of document which sort according to page rank algorithm in which lots of not relevant to the user's information needs. Hense the personalized search is required a better page rank mechanism to provide the more relevant document to user. In this paper we design a page ranking mechanism which based on the various factor time-span, interest score, activity performed, and user interest score which helps in improve the ranking.

Keywords: Information Retrievals, Search Engine, Interest Score, Personalized Model, User Profile, Page Rank

I. INTRODUCTİON

In the area of internet technology development is very fast more no of person and information is added in the every second that way World Wide Web (WWW) size in manner of collection of information is increasing rapidly and a huge amount of information is stored on the internet but this growth of the web is create more difficulties to retrieve the information which relevant to user's needs. Here one more problem is present in the previous search these are focused on the keyword of query given by the user not consider context of query. If user searches the “Apple” it mostly provides document related to “Apple Fruit and other fruits” not mention the apple phone or “Apple Mac book” if user interested in electronics product.

The solution to this problem is personalized search which is recently more active research field. Web personalized is requried a better page ranking mechanism to provided the more relevant information to the user to satisfy user needs. Various page ranking mechanism are peresnt but they are used the link oriented mechanism or content oriented mechanismof web mining. Here we proposed the page ranking mechanism which based on link and usage mining by considering their previous browsing history, user interest and preference etc. In case of our previous example searching for “Apple” and user previously browse the “Apple phone” related documents so our system based on our page ranking mechanism try to provide the phone related document not like the traditional page search engines provide most viewed document which not having relevancy to user.

This paper is arranged as it follows. Section II provides the related work. Section III provides our proposed search engine architecture and working of system. And we conclude in the Section IV.

II. RELATED WORK

Many algorithm in the area of personalization have been developed in past which are based on the user interest and user browsing history. In [1] Surgey Brin and Larry Page give the ranking algorithm named as Page Rank (PR) which used by the search engine Google. Google used the Page rank algorithm to rank the web pages. Page rank algorithm is classified in the web structure mining technique where algorithm is based on link structure of web pages. Page rank algorithm tells that both incoming and outgoing link is important and in this algorithm page rank is calculated by added all backlinks rank and final page rank score is calculated. In [6] Page Ranking based on Link Visit (PRLV) user browsing information is additional used to the original Google Page rank algorithm. PRLV is based on the web structure and web usage mining. Here user search behaviour is also considered. In this algorithm outgoing links which is more visited having the more rank score then less visited pages. User profiles can be created explicitly and implicitly. In [7] personalized search system is based on the web usage mining and user profile is created explicitly. Here user profile is created using the user browsing history and it attributes such as no of page visited, no of page clicked, time spent on web page and action performed on page. Previous history is used to re–rank the search results and using these factor more users relevant results are obtained. In [5] proposed the personalized search engine model that is based on the web usage mining. Here user profile is created when user registers first time on the system. User is asked to enter their interest area explicitly which keeps on adding their interest area depending on user browsing patterns. Feature words are extracted from the web page visited by the users which are further used to create the short term and long term user profile.

A critical on the available literature indicate the following short coming which needs to be address

- Although [1] the no of outgoing and incoming links to the page indicates its importance within the web, but it should not be considered as the only parameter to rank a page. The user browsing history may play a major role in finding the page relevance.
- Although the work in [7], considered the browsing patterns of the user while ranking the results but incorporating the short term and long term interest can better determine the page relevance to user.

III. PROPOSED PAGE RANKING MECHANISM

The proposed page ranking mechanism model as shown in the figure1. In the proposed model when the user process the query from the search engine interface then a signal is send to the profile generation module to monitoring the activity. In while query is send to the query processor and query processor retrieved the list of document from the database related to query keywords and performing the ranking on the list and send to search engine interface which display the result to user.

They are described as follows:

- Search Engine Interface
- Profile Generation Module
- Query Processor
- Database
- Page Classifier

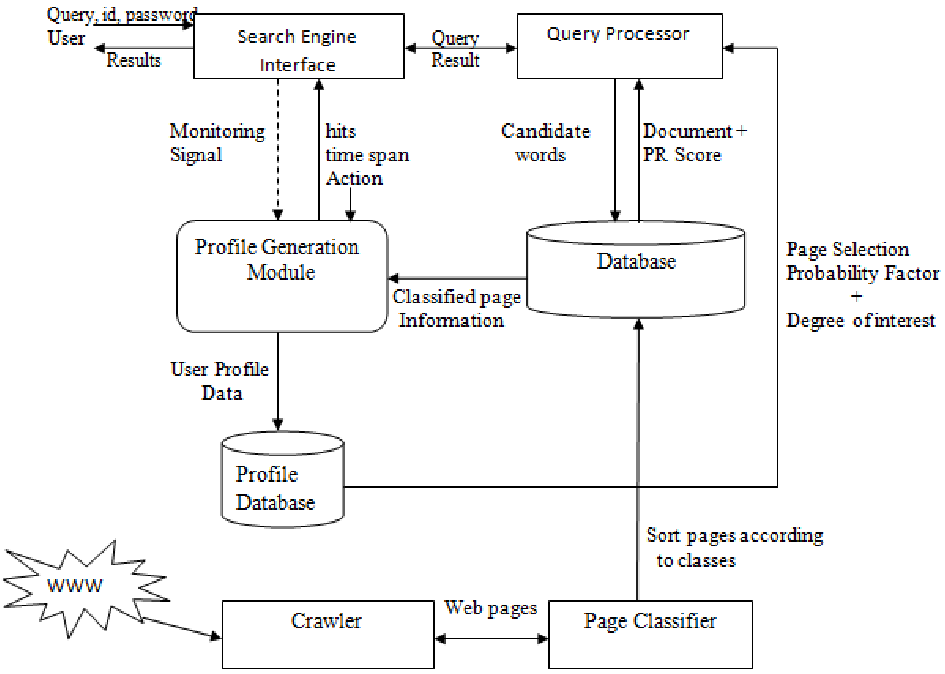

Fig 1 Proposed Architecture

3.1 Search Engine Interface

The search engine interface is a web page which having the text box in which user enters the query and search engine display the results. It registers the new user and authenticates the exiting user with user id and password. Search engine interface send the query submitted by user to the query processor to find the appropriate results related to query and pass signal "monitoring signal" to the user profile generation module. After getting the sorted results from query processor, presents the result to the user.

3.2 Profile Generation Module

The user profile is created by the special module is called "Profile Generation module". It creates the profile for every user which registers on the system. When user enters the query at search engine it gets the signal from search engine interface and become active. It monitors and save the information such as

hits(p_i ,u_i) = no of hits on Page P_i by user u_i

Action (p_i ,u_i) = action performed on page p_i by user u_i

ts($u_{i,}p_i$) = time-span on page p_i by user u_i

Interest score Is(u_i,c_i) = degree of interest of user u_i in class c_i

Fig 2 Profile Information

It stores the user profile information in the separate profile database. Three different table are used to store user information namely; Hits_info table, Time_span Table, user interest_class table

1) The Hits_info table store the page as the column and row as the different user and each entry is no of hits made by user u_i on the page p_i as shown in table 1.

Table 1 Hits_info Table

Class	C_1					C_2					C_k				
Page user	P_1	P_2	P_3	----	P_n	P_1	P_2	P_3	---	P_m	P_1	P_2	P_3	---	P_q
U_1	10	15	6	-	-	7	41	20	-	-	2	14	25	-	-
U_2	3	7	11	-	-	13	5	14	-	-	41	17	2	-	-
U_3	4	22	2	-	-	26	25	1	-	-	47	5	4	-	-
⋮	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U_n	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

2) The Time_span table store the pages as the column and row as the different user and each entry of table contain time span on page p_i by user u_i as shown in table 2

Table 2 Time_Span Table

Class	C_1					C_2					C_k				
Page user	P_1	P_2	P_3	----	P_n	P_1	P_2	P_3	---	P_m	P_1	P_2	P_3	---	P_q
U_1	30	25	8	-	-	11	16	18	-	-	32	21	23	-	-
U_2	20	12	28	-	-	10	25	11	-	-	13	27	21	-	-
U_3	42	22	24	-	-	14	35	10	-	-	7	35	14	-	-
⋮	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
U_n	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

3) The interest score table store the class of page as column and row as different user and each entry is contain the degree of user interest in the particular classes as shown in the table 3

Table 3 Interest_Score Table

class **user**	**A**	**B**	**C**	**D**	**E**
U1	65%	5%	15%	10%	5%
U2	20%	20%	25%	15%	20%
U3	10%	10%	50%	15%	5%
U4	5%	15%	40%	25%	15%

Example :- if user U1 is search about the India and it found in page P1 and P2. P1 in class A and P2 in class D the P1 have 65% interest_score and P2 have 10% interest_score.

Interest Score is calculation of degree of interest of user in different classes. The interest score is calculated with corresponding the classes of web page which define in database. User interest is can divided in short term and long term interest. Due Short term interest change in degree of interest is minor and long term having major role and it easily calculated. Calculation of interest score is shown in eq. 1

$$\text{Interest score } Is(u_i,c_i) = \frac{P(u_i, c_i)}{P_t} \quad \text{-------- } 1$$

Where :

- Interest score (u_i,c_i) is interest of user u_i in class c_i
- $P(u_i,p_i)$ is total page browsed by the u_i related to the web page class c_i
- P_t is the total page browsed by the user u_i

Here we also computed the Page Selection Probability Factor (PSPF). Which based on the user browsing pattern and PSPF is further used by the query processor in calculating the final page rank. PSPF tells the user browsing behaviour by which finding page relevance to user is may become more accurate. PSPF is computed by the hits, time-span, and action factor as we consider personalizing.

$$PSPF(P_i)=hits_{wt}(P_i)+ts_{wt}(P_i)+Action_{wt}(P_i) \qquad \text{------2}$$

Where:

- $hits_{wt}(P_i)$ is the ratio of number of HITs done by user u_i on the page p_i w.r.t. total no. Of HITs on all the pages made by user u_i
- $ts_{wt}(P_i)$ is ratio of time-span spent by the user u_i on the web page w.r.t. highest time-span spent on any page.
- $Action_{wt}(P_i)$ is activity performed by the user on web page like the print, saving, bookmark etc.

1) Hits weight is used in personalization because as many clicks or visits made by user means it more interested in this web page. Hits weight is count the number of click or visit made by user u on page p. In eq. 3 calculate the hits weight.

$$hits_{wt}(P_i)= \frac{hits(u_i,p_i)}{hits(u_i,*)} \qquad \text{------------ 3}$$

Where:

- $hits(u_i,p_i)$ is number of hits on Page P_i by user u_i
- $hits(u_i,*)$ is total no of hits on all Pages of belongs to all classes by user u_i

2) Time-span is a important factor in web personalization if any user spent more time on the web page it means it more like the page content means more relevance to user. Time-span weight is calculated of page p_i shown in eq. 4

$$ts_{wt}(p_i)= \frac{ts\ (u_{i,}p_i)}{\text{Highest time-span } (u_{i,}p_i)} \qquad \text{------- 4}$$

Where:

- $ts(u_{i,}p_i)$ is time-span spent on page p_i by user u_i
- Highest time-span spent is the time spent on any page by user u_i

3) Action is mainly classify in four type [3] which are print, save, bookmark and send. These four action listed in the below. We give the highest weight to print after that save, bookmark and last send. Action weight is calculated by sum of all action weight given in eq. 5

$$Action_{wt} = \frac{Print_{wt}+Save_{wt}+Bookmark_{wt}+Send_{wt}}{4} \qquad \text{---------- 5}$$

Where

- $Print_{wt}= \log\{click(u_i,p_i)+1\}*n*4$
- $Save_{wt}=\log\{click(u_i,p_i)+1\}*n3$
- $Bookmark_{wt}=\log\{ click(u_i,p_i)+1\}*n*2$

- Send_{wt}=log{ click(u_i,p_i)+1}*n*1
- click(u_i,p_i) is no times page viewed by user u
- n is no of time action performed

3.3 Query Processor

Query processor extract the candidate words form the query by apply stop word removal. After that query processor executes query on the database and fetch the page based on the term relevance.

After that query processor calculate the term relevane_score of all in the list pages. Term _relevance is calculated using the query candidate words and page keywords by checking the similarity between page and query and which matching score is called term_relevance. This term relevance score is used in the eq 6 to calculate final rank of all pages.

Query processor gets the PSPF and degree of interest from the profile database and applies the formula given in eq-6 on the list of document fetched form the database. Finally shorted list of document is provides to search engine interface. Formula of modified page rank:

$$PR(P_i) = \text{Term Relevance\_score} + \text{Degree of Interest}\{\text{interest score}\} + PSPF \quad \text{---------} \; 6$$

3.4 Database

Database is collection of summary of list of documents and their URL of page. Database is having the various attributes related to pages in the table as follows:

- URL of pages
- Class Ids
- Keywords of page
- Page Rank

3.5 Page Classifier

Page Classifier gets the keywords and summery of downloaded web page form the crawler and puts in the repository called database. Page classifier is classifying the pages in the different classes by matching the keyword of page to class. One page is may be belong to one or more classes.

Example: Page classifier classifying page in class like the:

- Entertainment
- Sports
- Real Estates
- Stock Market
- Education and etc.

Here a page related to sports class may also belong to entertainment because sports are also entertainment.

Hence some pages are comes under the one or more class.

IV. CONCLUSION

In this paper we have presented a new page rank mechanism which based on the user browsing history and degree of interest of user and a personalized search engine is build based on this new proposed page rank mechanism. Its provide the more relevant information to user needs.our page rank mechanism work better then previous availabe page rank algoritm but its have some limitation it create some extra burden of dividing page in classes but its not major. It give the better result to user and user needs of inormation is likely be more satisfy then our previous page rank algorithm. Comparison in between the different page rank algorithm is given in the below table 4

Model Item	PR[1]	PRLV[2]	PSE[3]	Proposed
Page Rank	Original	Modified	Modified	Modified
User Profile	Not created	Not created	Explicit	Explicit
Search Engine	Depended (Google)	Google	Google	Own
Ranking Factor	Ingoing and Outgoing links	Browsing patterns	Action, Click, Time	Browsing history attributes, degree of interest
Limitation	Not include user browsing behaviour	Link visit is not major factor some other like action, time not consider	Not consider degree of interest	Extra effort to divide page in classes and its some case not feasible

REFERENCES

[1] S. Brin and L. Page., “The Anatomy of a Large-Scale Hyper textual Web Search Engine”, Proceedings of 7th International World Wide Web Conference, pages 107–117, 1998

[2] Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, “Adaptive Web Search Based on User Profile Constructed without Any Effort from Users”, Proceedings of the 13th international conference on World Wide Web,2004

[3] Ivan Marecialis and Emanuela De vita, “SEARCHY: An agent to personalize search result”, Third International Conference on Internet and Web Application and Services, 2008

[4] Hany M. Harb, Ahmed R. Khalifa, Hossam M. Ishkewy, “Personal Search Engine Based on User Interests and Modified Page Rank”, International Conference on “Computer Engineering & Systems, (ICCES)” 2009

[5] Liu Zhongbao, “Research of a Personalized Search Engine Based on User Interest Mining”, International Conference on “Intelligent Computing and Integrated Systems (ICISS)”, 2010

[6] Gyanendra Kumar, Neelam Duahn, and Sharma A. K., “Page Ranking Based on Number of Visits of Web Pages”, International Conference on Computer & Communication Technology (ICCCT)-2011, 978-1-4577-1385-9.

[7] Shilpa Sethi, Ashutosh Dixit, “Design of Personalised Search System Based on User Interest and Query Structuring”, International Conference on “Computing for Sustainable Global Development” 2015