Personalized Page Rank Algorithm Based on User Profile and Meta Keyword

Nirja Pathak[1] Kruti Patel[2]
[1,2]Department of Information & Technology
[1,2]KITRC, Gujarat, India

Abstract— During the last 3 decades, world wide web has been filled with so many resources from so many topics. So many organizations and personals avail their data and knowledge using websites and various social media platforms. Information regarding any topic is available by searching with keywords on various search engines. Mostly when user search for any product, service or any other information; search engines provides the result based on keywords and rank the result based on matching score of various keywords and other keyword related methods. In most of cases user gets the same result by searching the same query in a particular search engine. There is a scope to re rank the result for every user based on his own profile and user's own criteria. Every common result may not be same useful for each user in general way. During our research work we have tried to identify the user behaviour and used the same user behaviour parameter in his next search result. We have expanded the query based on his profile and re-ranked the result. After evaluation we have got significant improvement in data retrieval compared to existing system.

Key words: Page Ranking, Query Expansion, HITS Algorithm, Symantec Query, Web Mining

I. Introduction

In current era large amount of websites and blogs are published for various services and products. Social media websites has already boost up the industry of information sharing. It's very easy for anyone to publish his own business, service or product information using so many facilities of web. WWW has been ocean of information in last 3 decades. Searching of information from WWW has been a crucial service for any user. For all this task importance of Web Mining has been crucial. Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents[4]. In general case web mining consists mainly four separate functionalities:

A. Resource finding:

Resource finding includes the extraction of web pages and find the related resources from the webpages those are available on web.

B. Information Selection and Pre-Processing:

After resource selection next step is to select appropriate information from those resource and pre-processing contains the task of cleaning the data before presenting to the end user or using the same resources for any use.

C. Generalization:

Generalization includes the finding out general patterns of data and information. Mostly data mining and machine learning techniques are used for the same purpose.

D. Analysis:

This step used by experts or machines to find out meaningful result from the patterns or generalized data.

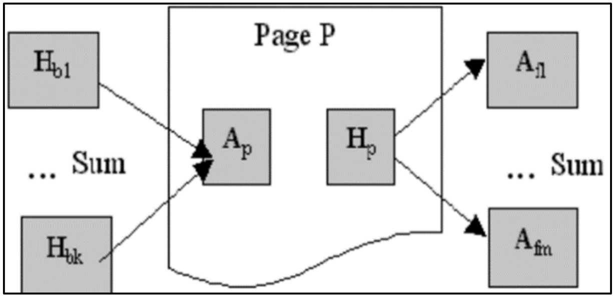
Based on usage of Web Mining we can categorize the same in major three categories that is

1) Web Content Mining: Mining of the content available on website and use it for analysis purpose.
2) Web Structure Mining: this type of category includes the analysis of Web Structure like how each pages are connected to another, relation between pages etc. 3
3) Web Usage mining includes the data of users and usage of webpages; how, who, where data of website is being used is the main result of this type of category.

For web mining and data retrieval Page Ranking is an important factor. Page Ranking is basically a numeric value for the webpage that ranks the webpage in particular category. Page ranking like Google Index, Yandex are used for displaying the sequence of pages to the user as a part of their search result. Each search engine has its own method for indexing the webpages and based on their method they calculate the PageIndex of any particular webpage. Search Engines frequently changes their method of page indexing based on their experiments.

For Page Ranking mostly two methods are used:

1) HITS Algorithm: HITS stands for Hyperlink Induced Topic Search. It was developed by Jon Kleinberg [6]. HITS algorithm analyse the various inbound and outbound links for any webpage.



2) Page Rank (PR) : Page Rank algorithm is used by well-known searching Google. Google ranks the pages based on this algorithm, this algorithm was named after Larry Page, one of the founder of Google. Page Rank algorithm mostly focus on importance of the webpages in terms of query. To calculate Page Rank of given page A we can use formula like:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

where,
PR(A) – Page Rank of page A
PR(Ti) – Page Rank of pages Ti which link to page A
C(Ti) - number of outbound links on page Ti
d - damping factor which can be set between 0 and 1

For better clarification we have generated a table below with the summary of both algorithm (HITS and PR)

Parameter	HITS	Page Rank
Mining technique used	WSM & WCM	WSM
Working	Computes hub and authority scores of n highly relevant pages on the fly	Computes scores at indexing time. Results are sorted according to importance of pages.
I/P Parameter	Backlinks Forward Links & contents	Backlinks
Limitation	Efficiency Problem	Query Independent

Table 1: Comparison of HITS and Page Rank

II. Related Work

As so many researchers has worked toward Page Ranking Algorithms. We have reviewed many of research paper related to their work. In research work "Agent Based Weighted Page Ranking Algorithm for Web Content Information Retrieval" by "Nagappan, V.K ,Dr. P. Elango"[1] introduced a new method for page ranking by modifying existing classical methods. Their work paper focuses on WCM and provides a new Weighted Page Content Rank Algorithm with the help of an agent. As we have expanded the same research work in next section it is explained in depth. In another research work named "Efficient Key Hash Indexing Scheme with Page Rank for Category Based Search Engine Big Data[2]" by "Ragavan N" published in IEEE in year 2017, researchers proposed a new method of Indexing Search Engine Big Data called Key Hash Indexing scheme followed by the implementation of Page Rank. In their method they have focused on indexing in BigData Storage. After evaluation of their research they have tested their system over various real time data and founded their system faster than existing. We also reviewed a research work by "Lissa Rodrigues, Shree Jaswal" named "Hybrid Model for Improvised Page Ranking Algorithm [3]". In their model they have used Hybrid Model for Page Ranking. They have used NE04J, webgraph and databases. In another research work "Toward Efficient Hub-Less Real Time Personalized PageRank [4]" by "MATIN PIROUZ AND JUSTIN ZHAN" also considered BigData storage and Page Retrieval system for BigData. They have used pruning graph technique to obtain better result. As Social Media is booming now a days, we also reviewed a research that is aimed for social media network. For that we have reviewed "Realtime Personalized PageRank Query for Social Network Search[5]" by "Jiang Wu" For their research work they have combined several optimization methods to make it practical for personalized social network searching and obtained a good result.

III. Existing Work

As mentioned in sector II, in research paper "Agent Based Weighted Page Ranking Algorithm for Web Content Information Retrieval"[1] they have used a new approach named AWPR algorithm. In their research they have founded that most of the search engines are ranking their search results in response to users queries to make their search navigation easier but there is a scope of improving the user experience by modifying the classical approach with applying weighted page rankings. They also mentioned "Although Page Rank and Weighted Page Rank algorithms are used by many search engines but the users may not get the necessary relevant documents easily on the displayed pages"[1]. So that they have given a solution is designed at enlightening the order of the pages in the result list so that the user may get the relevant and important pages easily in the list.

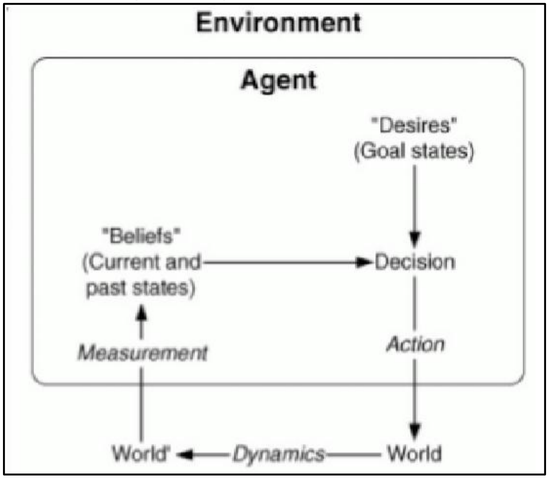


Fig. 2: Model for AWPR.

They have calculated weight for links for both type of connections in and out. Formulas for weight calculation used are:

$$Win(U,V) = 1.1 \frac{Iu}{\sum_{P \in R(V)} Ip}$$

Where IU = number of in-links of page u , Ip= number of in-links of page p, R(v)=Reference page text of page v

$$Wout(U,V) = 1.2 \frac{Ou}{\sum_{P \in R(V)} Op}$$

Where Ou=number of out-link of page u, Op= number of out-link of page p

They have improved the listing of Pages and their ranking system. After applying their algorithm user gets the pages based on their requirement properly. But they have missed the user behaviour feature for decding the page rank; that is important for personalized page ranking.

IV. Proposed Methodology

As dicussed in section III existing system a new aproch for Personalized Page Ranking has been developed where Agent based Approch and Query Expansion approach is applied. We have expanded the same work as a part of our research. We have added user behaviour feature that includes the User Click information and User Profile Information. We have generated an user profile based on their search and clicks. This profile is automatically updated at every search that is performed by user. We have applied our system on WikiPedia Dataset. After evaluation we have founded the proposed system gives better pages those are suitable to users at top passion, so that overall process of searching for user takes less time compare to existing system.

Steps for the Proposed System

1) Step 1: Get User IP for Profile
2) Step 2: Get Query
3) Step 3: Clean Query if Required
4) Step 4: Get the Result as per Base Algorithm
5) Step 5 : Re Rank the Result Based on Meta Keyword Weight
6) Step 6 :Display Search result & Calculate Relevancy
7) Step 7: Get Click Details
8) Step 8 : Update User Profile

For Relevancy we will use:

$$\text{Relevancy}\ (t, p) = \sum Wkt*Wkp / \sqrt{\sum Wkt^2 * Wkp^2}$$

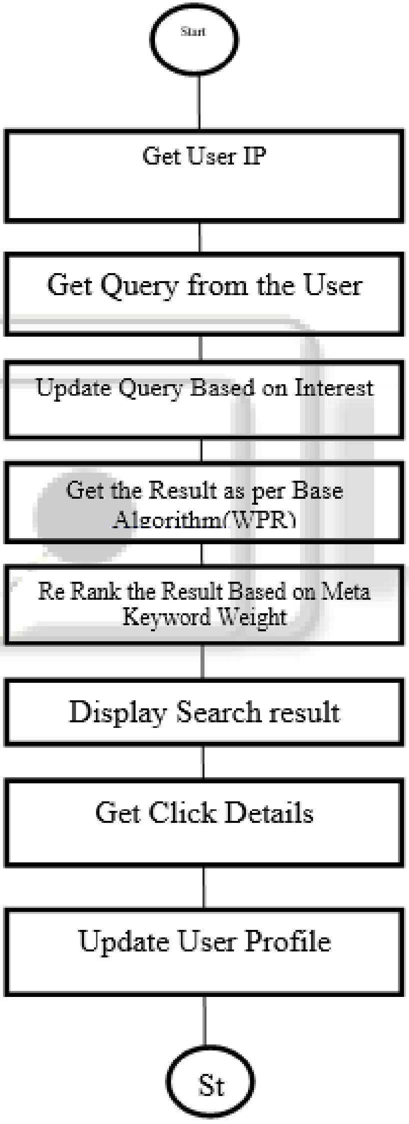

Start
Get User IP
Get Query from the User
Update Query Based on Interest
Get the Result as per Base Algorithm(WPR)
Re Rank the Result Based on Meta Keyword Weight
Display Search result
Get Click Details
Update User Profile
St

Fig. 3: Proposed Flow Diagram

V. Conclusion & Future Work

As agent based personalized page ranking system performed well for getting relevance pages. But they had still some points to cover with user profile generation and apply the same feature in searching and page ranking. We have applied IP based user profile creation and user behaviour is recorded to update user's profile based on clicks. We have applied the method and on dataset we have achieved better relevance score of web pages based on search term. Still there is scope to add more parameters to build a user profile and apply the same parameters for getting better page ranking.

References

[1] Nagappan V.K and P. Elango, "Agent based weighted page ranking algorithm for Web content information retrieval," *2015 International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, 2015, pp. 31-36. doi: 10.1109/ICCCT2.2015.7292715

[2] N. Ragavan, "Efficient key hash indexing scheme with page rank for category based search engine big data," *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Srivilliputhur, 2017, pp. 1-6. doi: 10.1109/ITCOSP.2017.8303118

[3] L. Rodrigues and S. Jaswal, "Hybrid model for improvised page ranking algorithm," *2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kumaracoil, 2015, pp. 466-469. doi: 10.1109/ICCICCT.2015.7475324

[4] M. Pirouz and J. Zhan, "Toward Efficient Hub-Less Real Time Personalized PageRank," in *IEEE Access*, vol. 5, pp. 26364-26375, 2017. doi: 10.1109/ACCESS.2017.2773038

[5] J. Wu, "Realtime personalized PageRank query for social network search," *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, 2017, pp. 647-651. doi: 10.1109/IAEAC.2017.8054096

[6] W. Zheng, S. Mo, P. Duan and X. Jin, "An improved pagerank algorithm based on fuzzy C-means clustering and information entropy," *2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE)*, Beijing, 2017, pp. 615-618. doi: 10.1109/CCSSE.2017.8088006

[7] D. Silvestre, J. Hespanha and C. Silvestre, "A PageRank Algorithm based on Asynchronous Gauss-Seidel Iterations," *2018 Annual American Control Conference (ACC)*, Milwaukee, WI, 2018, pp. 484-489. doi: 10.23919/ACC.2018.8431212.

[8] T. Sen and D. K. Chaudhary, "Contrastive study of Simple PageRank, HITS and Weighted PageRank algorithms: Review," *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Noida, 2017, pp. 721-727. doi: 10.1109/CONFLUENCE.2017.7943245

[9] L. Z. Xiang, "Research and Improvement of PageRank Sort Algorithm Based on Retrieval Results," *2014 7th International Conference on Intelligent Computation Technology and Automation*, Changsha, 2014, pp. 468-471. doi:10.1109/ICICTA.2014.119

[10] S. G. Pawar and P. Natani, "Effective utilization of page ranking and HITS in significant information retrieval," *International Conference for Convergence for Technology-2014*, Pune, 2014, pp. 1-6. doi:10.1109/I2CT.2014.7092030