A GEOGRAPHICAL LOCATION INFLUENCED PAGE RANKING TECHNIQUE FOR INFORMATION RETRIEVAL IN SEARCH ENGINE

Sanjib Kumar Sahu[1], Vinod Kumar J.[2], D. P. Mahapatra[3] and R. C. Balabantaray[4]
[1]Department of Computer Science, Utkal University, Bhubaneswar, Odisha, India
[2]USICT, Guru Gobind Singh Indraprastha University, Delhi, India
[3]Department of Computer Science and Engineering NIT, Rourkela, Odisha, India
[4]Department of Computer Science and Engineering IIIT, Bhubaneswar, Odisha, India
E-Mail: sahu_sanjib@rediffmail.com

ABSTRACT

Internet contains huge amount of information, getting the desired page at the top of search results is always a challenging task, as the expectation varies from one users to other users. Each user performs the search expecting certain categories of pages like research, definition, downloads etc., at the top of the search result but users doesn't always provided their desired category during the search. In this paper we present a novel technique where the category of pages that need to be given more priority while calculating the Page Rank, can be judged using the geographical location from where the search is initiated. The Category preference for that geographic location is gradually developed based on the various searches performed by different user from that location. This Category preference is added as one of the factor while calculating the Page Rank.

Keywords: page ranking, geographical navigation, information retrieval, query engine.

1. INTRODUCTION

World Wide Web contains huge amount of information. Each user spend considerable amount of time in searching for his desired information. Search Engine aids user in finding the required information. Primary component of the Search Engines is its Crawler or Spider which crawl the complete World Wide Web and fetches various pages. The fetched pages are indexed using various information present on that page and stored in organized way so these information are provided to the user as and when the search is performed. The Quality of the search engine mainly depends on the prioritization of the searched pages from the result. If the user gets his relevant pages on the top of the search result then we could assume the search engine quality to be high. For finding the relevant pages for the user among the set of matched pages various Ranking techniques are applied. Hence ranking techniques primarily decides the quality of the search engines.

The Ranking techniques are classified in the following ways:

1.1 Web content and web structure mining

This is a technique where the matched documents of search results are ranked based on the content of the pages. Web structure mining is a technique where the ranking is performed based on the link structure of each pages. The complete web is viewed as a graph structure with webpage as node and each link of that webpage as an edge connecting two webpages.

1.2 Web usage mining

This is a technique where the usage pattern of the web such as time spends over the page, navigation patterns etc., are stored on the server as a logs over a period of time. These server logs are analysed and the rank of the page is determined.

Most successful web page ranking Algorithm is Google's Page Rank Algorithm [1], which is based on the Web Structure Mining technique. This ranking technique uses the link structure of the web pages along with various add-on parameters to find the rank of the pages. In most of the commonly available ranking algorithms, category of search result such as research downloads, gaming, information etc., desired by user is not considered. It is assumed that the User expected category can be highly influenced by the geographical location from where the search is performed, example, if search initiated from the Geographical location where more academic institution exist, then the preference would be given to pages containing more research content. With this view, in this paper we propose a model where the system initially works using a Web Structure Mining and logs of each user using the search engine from that region is stored. These logs can be used to identify the most preferred category of pages used by the user belonging to that geographical location. This category preference factor is added as one of the factor along with other factors to decide the page ranking.

The rest of the paper is organized as follows. The Section II covers the review work based on the existing page ranking algorithm; Section III covers the architecture of the proposed model along with its components and algorithms. Section IV covers the result and analysis. Finally concluded with conclusion and future scope in section V.

2. LITERATURE REVIEW OF EXISTING RANKING ALGORITHMS AND MOTIVATION

Quality of search engine depends on the factors such as time required performing the search, relevancy among the results, complexity of interface etc. The relevancy of the search results is given more priority compared to the other parameters. Among the list of several search engines, Google is one of the highly used search engine. Google tops the competition by providing more relevant information at the top of the search results; the core behind the Google's highly successfully page ranking technique is the Google's Page Rank Algorithm [1]

2.1 Page rank algorithm (PR)

Brin *et al.* [1] developed a ranking algorithm for prioritizing their search result and named as Page Rank Algorithm. This algorithm states that if there exist some important page in the web then, all the other pages linked via this page also should be considered as important. Page Rank Algorithm works with the concept of back links, the page containing more back links would have more Ranking. When the search is performed PageRank already calculated for each page is considered along with many other factor and the sorting of result is done.

Page Rank of each page is calculated using the following equation:

$$PR(u) = (1-d) + d \sum_{v \epsilon B(u)} \frac{PR(v)}{N_v} \quad (1)$$

Where B(u) denotes set of Back link pages pointing to the Current page, PR(v) denotes the Page Rank of one of the Back link pages, N_v denotes the total outgoing links from the page p, d is considered as a damping factor normally set to 0.85 which is considered as a probability of users who would use the page directly.

2.2 Weighted page rank algorithm (WPR)

Xing *et al.* [3] developed Weighted Page Rank (WPR) algorithm by assigning weights to each page. In Page Rank algorithm all pages have assumed to have same page rank. In Weighted Page Rank algorithm more weightage is given to page which are more popular or important, the algorithm doesn't distribute the rank evenly to all pages. The Weightage of incoming and outgoing links are denoted as $W^{in}(v,u)$, $W^{out}(v,u)$, The equation for Weighted Page Rank Algorithm is modified as

$$\mathrm{WPR(u)} = (1-\mathrm{d}) + \mathrm{d} \sum_{\mathrm{v \epsilon B(u)}} (\mathrm{WPR(u)} \,.\, \mathrm{W^{in}(v,u)} . \mathrm{W^{out}(v,u)}) \quad (2)$$

2.3 Page ranking based on timestamp and link (WTPR)

Shiguag *et al.* [4] used the last modification time which is present in the HTTP response of the each node along with the weight of in-link and out-link like WPR to calculate the Page Rank. WTPR makes the new page to appear at the top of the page search result

2.4 Customised page ranking

Anjali *et al.* [5] proposed a customized page ranking algorithm where the user preference is recorded and the result are displayed as their preference. The algorithm requires a user to register into the system and decide their preference. Domain authority, Page Authority, Social signals and the search history are also considered while calculating the Page Rank.

2.5 Feedback based page ranking technique

Gupta *et al.* [6] suggested a novel user preference and Feedback based Page Ranking technique. This technique combines the Page similarity, link structure information with user preference based on internet domains, and implicit user feedback (based on number of clicks and time spent on web page) for page for calculating of Page Rank.

Page Rank is calculated using the following equation,

Page Rank=0.2*PR + 0.2*PH_Score + 0.3*PC_Score + 0.3*DoMn (3)

Where,

PR is the Weighted Page Rank Score [2]

PH_Score is the Page history score which denotes the Average time Spent by user on the page using the Server logs.

PC_Score is the Page Content Score, which is calculated as follows:

PC_Score=0.2 * URLText + 0.2*TitleText + 0.3*LinkText + 0.3*BodyText (4)

Where,

URLText= No of Query keywords that appear in WebPage URL / Total number of terms in URLText

TitleText =No of Query keywords that appear in Title tag of web page / Total number of terms in Title tag

LinkText= No of Query keywords that appear in Link tag of web page / Total number of terms in Link Text

BodyText= No of Query keywords that appear in Body tag of web page / Total number of terms in Body

DoMn is the Domain score, it is acquired from the user, its value is set as 1 if website matches the user specified domain else its set to 0 i.e, DoMn is not considered as a factor.

2.6 Motivation of research

In the Feedback based page ranking algorithm[6], the DoMn factor is based on the user choice of selecting the domain, the user can select the domain from where the page need to be selected like .com, .in, etc. But as of today's scenario, .com doesn't only contain the commercial web application; hence the relevancy of domain DoMn factor doesn't contribute much in the page ranking. If the DoMn factor if modified with the Category preference of user then ranking of relevant result at the top would be high and if the Category preference can be automated then there is no stress on personalization.

3. PROPOSED MODEL AND ANALYSIS

3.1 Architecture

The Architecture of the proposed model contains the following components as shown in the Figure-1.

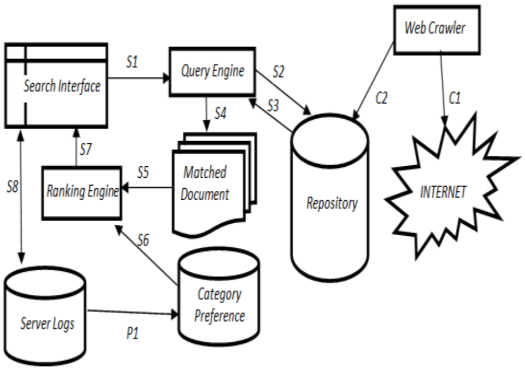

Figure-1. Architecture diagram of the proposed model.

3.2 Components

3.2.1 Web crawler

Web Crawler is the core component of the search engine, WebCrawler fetches each and every pages from the internet hence this component should be powered with sufficient processing power for indexing the fetched content and more bandwidth for fetching more pages at a single instance of time. It also maintains a URL Seen list containing the list of URL that is already processed to avoid crawling the same URL again.

3.2.2 Repository or store

Repository store is a Data store where are the fetched and indexed page are stored along with their indexed information.

3.2.3 Search interface

It's a web interface through which the User Initiates the Search. Interface contains a text box, where the user enters his desired keywords and clicks the search button to perform the search. The Search interface runs a script which updates the logs such as pages selected by user, location from where the search is performed etc.

3.2.4 Query engine

Query engines analyse the keyword or phrase presented by the user and uses indexed store to find the relevant match. The set of matched documents are brought to the memory for processing and given to the Page Ranking Engine, The Query Engine should be provided with more processing power to decrease the search time.

3.2.5 Page ranking engine

Page Ranking Engine, uses the proposed Geographical Page Ranking (GPR) algorithm to sort the page matched by the Query Engine.

3.2.6 Server logs

Server logs contain logs of various searches performed and its statistics. It gets updated via a script which is run on the Search Interface of the client browser.

3.2.7 Category preference maintainer

Category preference maintainer is a store that frequently updates the Preference list of category for every geographical location based on the Server logs.

3.3 Working of proposed model

C1- The Search engine need to be initialized before the user performs the search, web crawler which is the core component of the search engine fetches each and every page in the Internet.

C2- Keywords are extracted from the fetched pages, category of the webpage is identified and it is stored in the repository store indexed by its keyword. The Weighted Page Rank (WPR) is calculated for the fetched page based on the link structure and its updated in the repository. This process recurs in the server looking for new pages or for updating the already fetched pages.

S1- User, who wants to perform the search, opens the search interface and provides the phrase or keyword for search. These keywords and phrases are sent to the query engine.

S2- The query engine extracts the keyword that needs to be searched from the phrases and performs the search on the repository store.

S3- The repository returns the set of matched document to the Query engine.

S4- Query engine identifies the location from where the search is performed using the IP address present in the request header. Matched documents along with the location details are given to the Ranking engine for sorting the result as per the proposed Geographical Page Ranking Algorithm.

S5, S6- Ranking engine sorts the results as per the proposed algorithm and sends the result to the Search interface.

S7- The background script running at the Search interface monitors the user's choice of selection among the results, time spent over the page, location etc., And sends it to the server and logs are maintained.

P1- The server logs containing the user's choice of selection and the location from where the search is performed etc. are identified and the Preference for each location is updated.

3.4 Explanation of Geographical Page Rank Technique

Geographical page rank technique (GPRT) presented in this paper, is calculated by considering the following factors 40% of Page Rank Score, 30% of Page Content Similarity Score, 30% Category Preference Score. It's calculated as follows:

$$GPR\ (u) = 0.4 * PageRank + 0.3 * PageContentSimilarityScore + 0.3 * CategoryPreferenceScore \qquad (5)$$

Page rank: Page rank is calculated using the Weighted Page Rank Algorithm [2] using the equation 2.

Page content similarity score: Various content present in the pages such as URL text, Title text, Link text, Body Text are analysed with the user provided keyword and the match score is generated. Equation 4 of Feedback based Page Ranking technique [6] could be used for generating this score.

Category preference score: The main consideration of our model is this category preference score. The score is calculated using the following algorithms

3.5 Algorithm 1: Category Preference Score

```
Input-
    Location (Location from where the search is initiated)
    Category (category of the current page for which the preference score need to be calculated)
Output-
    PreferenceScore (Score normalized between 0-1)
Step 1: Find CategoryCount from CategoryPreferenceStore where location=Location and category=Category
Step 2: If CategoryCount is NULL or 0, PreferenceScore=0, return
Step 3: Find the MaxCount=Max (CategoryCount) where the location=Location
        //Normalize the current category score between 0-1
Step 4: PreferenceScore = CategoryCount / MaxCount
Step 5: Return PreferenceScore
```

3.6 Algorithm 2: Category preference maintainer

Category Preference need to be updated constantly with the help of server logs. The following algorithms need to be followed for m

```
Input- Location (Location from where the search was performed)
       Category (Category of the link from the search result which the user found it as relevant or spend more amount of time)

Output - VOID

Step 1: Find Category Count from Category Preference Store where location= Location and category= Category
Step 2: If CategoryCount is NOT NULL, goto Step 4
Step 3: Create a new CategoryCount for the Location and initialize it to 0
Step 4: Increment the CategoryCount by 1
```

4. RESULT AND ANALYSIS

To analyse and prove the efficiency of the proposed technique we here take the reference study and proof of various other models and we can infer the performance of our proposed algorithm. The real study can be conducted only after the implementation of the proposed technique. In User Preference and Feedback Based Page Ranking Technique [6] the user category preference is considered and the study has been shown the improved performance with reference to the Google Page Rank Algorithm [1]. In User Preference and Feedback Based Page Ranking Technique [6] user preference had contributed much personalized search and the result shown much improved ranking in the searched result but the limitation is that the user wouldn't be willing to provide his preference choice all the time. The result is analysed based on the "Precision Value" which is calculated from the following way,

$$Precision = \frac{\begin{array}{c} Sum\ of\ scores\ of\ relevant\ pages \\ returned\ by\ ranking\ method \end{array}}{\begin{array}{c} Total\ number\ of\ pages\ returned \\ by\ ranking\ method \end{array}}$$

The preliminary study results derived on the paper User Preference and Feedback Based Page Ranking Technique [6] is normalized to work at 75% efficiency for the proposed algorithm. The proposed algorithm initially works similar to that of the Google Page Rank Algorithm as the category preference won't be available initially. When the search is performed from the geographical area over a period of time the search result precision increases and move towards the Precision Feedback Algorithm result but the advantage is that it works without the preference choice from the user.

Query No.	No of top n pages selected for evaluation	Precision feedback algorithms precision	Google PR algorithm precision	Proposed algorithm Precision (considered at 75% efficiency of user feedback algorithm)
1	15	0.7	0.43	0.53
2	15	0.66	0.36	0.50
3	15	0.63	0.46	0.47
4	15	0.76	0.4	0.57
5	10	0.8	0.4	0.60
6	10	0.68	0.45	0.51
7	10	0.75	0.6	0.56
8	10	0.8	0.35	0.60
9	5	0.9	0.3	0.68
10	5	0.9	0.24	0.68
11	5	0.9	0.24	0.68
12	5	0.8	0.3	0.60

Figure-2. Analysis Report with Comparative Study.

The result shows that the proposed algorithms precision is better than Google's page Rank algorithm and the efficiency slowly increased towards the efficiency of the Precision feedback Algorithm's prevision but without the user personalization.

4. CONCLUSIONS

In this paper we presented a New Page ranking technique in which geographical location from where the search is performed is considered as one of the factor for deciding the page rank. Currently existed techniques requires user to provide input choice option for their customised search whereas by employing our model the user gets a feel of personalised search without employing much effort. Web Structure Mining, Web Content mining, Web usage mining are employed together in our model, the proposed page ranking algorithm if implemented it would be better than existing page ranking algorithms by providing more relevant page at the top.

REFERENCES

[1] L. Page, S. Brin, R. Motwani and T. Winograd. 1998. The pageRank citation ranking: bringing order to the web. Technical Report, Stanford InfoLab.

[2] S.Brin and L.Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems. 30(1-7): 107-117.

[3] Wenpu Xing and Ali Ghorbani. 2004. Weighted PageRank Algorithm. International Journal of Engineering and Innovative Technology (IJEIT). 1(2): 305-314.

[4] Shiguang Ju, Zheng Wang and Xia Lv. 2008. Improvement of Page Ranking Algorithm Based on Timestamp and Link. 2008 International Symposiums on Information Processing.

[5] Anjali, Ankita Sadhwani, Nidhi Saxena. 2015. A New Approach to Ranking Algorithm-Custom Personalized Searching. 20152nd International Conference on Computing for Sustainable Global Development (INDIACom), Delhi.

[6] Ashlesha Gupta, Ashutosh Dixit, Pooja Devi. 2015. A Novel User Preference and Feedback Based Page Ranking Technique. 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).

[7] Delhi Normalizing the Data,http://stats.stackexchange.com/questions/70801/how-to-normalize-data-to-0-1-range.