A Hybrid Page Rank Algorithm using Content and Link Based Algorithms

Madhurdeep Kaur1, Asst. Prof. Charanjit Singh2

1Research Scholar (Department of Computer Science), Rimt-Iet, Mandi Gobindgarh, India.
2Asst. Professor (CSE Dept), Rimt-Iet, Mandi Gobindgarh, India.

Abstract: - World Wide Web (WWW) is a huge repository of data. It has become one of the most important media to store, share and distribute information. The fast growth of the internet and the World Wide Web, results in network congestion and server overloading. Web Mining is a challenging activity that aims to discover new, relevant and reliable information and knowledge by investigating the web structure, its contents and its usage. Web search engines are the main application of Web Mining. Various page ranking algorithms are used by them, in order to rank their search results. In this paper, we propose a Hybrid Page Rank (HPR) algorithm which is based on both content and link structure of the web pages. The main aim of this research is to discover an efficient approach so that users can get the relevant results from the web search engines.

Keywords:- WWW; Data mining; Web mining; Search engine; Page ranking.

I.INTRODUCTION

Today internet has become the main source of retrieving information. The WWW is one of the services that run on the internet. It is an enormous, contrary diverse, dynamic and mostly formless data warehouse. As the volume of information on WWW is increasing day by day so it becomes difficult for the website owners to provide the relevant information to the internet users.

In such situations, various web search engines are used by users for finding suitable answers for their queries. Every web search engine has its own respective architecture and based on that, it performs number of tasks.

There are three basic components of a web search engine known as Crawler, Indexer and Ranking Mechanism. The search engine downloads, index and store up hundreds of millions of web pages. Every search engine uses web mining techniques and ranking mechanism for retrieving the information effectively.

The ranking mechanism is either based on the content of the web pages or link structure of the web pages. The main purpose of these ranking algorithms is to sort the results to be displayed so that user will have the most important and useful results first.

The work proposed in this paper aims to optimize the results of a search engine by returning the more relevant pages on top of the search result list. To perform the required task, both content and link structure of web pages have to be considered.

The paper is divided into different sections: Web Mining is introduced in section II. In section III, this paper describes the related work done for ranking web pages. Section IV includes the proposed framework that computes the rank of the web pages on the basis of content and link structure of the web pages. Finally we conclude the paper in section V.

II. WEB MINING

Web mining is defined as the application of data mining techniques to extract the useful information from the web data [22]. Web mining is used to discover the content of the web, the user’s behavior in the past and the web pages that the users want to view in the future [23]. The web data that can be collected and used in Web Mining analysis includes content data (text, images, records, etc); structure data (hyperlinks, tags, etc) and usage data (http logs, app server logs, etc) [24]. Web Mining can be divided into three categories [1, 2] namely web content mining, web structure mining and web usage mining.

Web Content Mining (WCM) is responsible for exploring the useful information from the contents of web documents. It may consist of text, images, audio, video, or structured records such as lists and tables.
**Web Structure Mining (WSM)** is the process used to extract structure information from the web using graph theory. WSM can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level.

**Web Usage Mining (WUM)** is the application of data mining techniques to discover the interesting usage patterns from web data generated by client-server transactions on one or more web localities.

The three categories of Web Mining described above have their own appliance areas including business intelligence, site improvement and modification, web personalization, ranking of pages etc.

### III. RELATED WORK

Most of the search engines use Page Ranking Algorithms for arranging the web documents in order of their relevance, importance and content score. Many algorithms have been proposed in the literature such as Page Rank, Weighted Page Rank, Sim Rank, HITS [3, 4, 5, 7, 8, 9, 15].

These are used to rank the query results of web pages in an effective and efficient fashion. Some algorithms rely only on the link structure of the document i.e their popularity score (web structure mining), some look for the content of the documents (web content mining), while other use a combination of both i.e they use links as well as the content of the document to assign a rank value to the concerned document.

- **Page Rank Algorithm** was developed by Sergey Brin and Larry Page [4, 5]. It is used by the Google [6] web search engine to rank websites in their search engine results. It is used to measure the importance of website pages by counting the number and quality of links to a page.

- In [9], the algorithm was proposed called Weighted Page Rank by Wenpu Xing and Ali Ghorbani, which is an extension of Page Rank algorithm. It is based on the popularity of the web pages and assigns the rank values accordingly rather than dividing it evenly.

- **Sim Rank** [15] is a new page rank algorithm which is based on similarity measure from the vector space model.

- **HITS (Hyperlinked Induced Topic Search)** [8] was developed by Jon Kleinberg [7] which gives two forms of web pages called Hubs and Authorities.

In existing algorithms, the page rank or score is calculated either on the basis of link structure of web pages or the content of the web pages but they have not considered both. We are proposing a Hybrid Page Rank algorithm which is based on both, the link structure of the web pages as well as the content of the web pages.

### IV. PROPOSED FRAMEWORK

**A. Objectives:**

- To study various link based and content based algorithms.
- To design an efficient algorithm by combining both content based (Sim Rank algorithm) and link based (Page Rank algorithm) web mining techniques.
- To enhance the technique by considering Sim Rank algorithm, this will calculate the score of the web page by comparing the webpage content to user query and Page Rank algorithm will calculate it on the basis of links of a web page.
- To analyze the scores of both algorithms and combine them to find the final score or rank of the page.

**B. Outline of the Algorithm:**

**Page Rank based on Links of the web pages:** Page Rank algorithm is a link based algorithm. The rank or score of the page is calculated on the basis of in-bound and out-bound links of the pages. Inbound links or commonly known as “back links” are the links pointing to our web page which decides the rank of the page. Outbound links are the links pointing to other pages from our web page. In our approach we are using a local web repository. A link structure is defined, showing how pages are linked to each other. This link structure will be used to calculate the page
rank. Eq.1 is used to calculate the page rank on the basis of links.

\[
PR(u) = (1 - d) + d \sum_{v \in S(u)} PR(v)/TL(v)
\]  

(1)

Where \( S(u) \) is the set of pages that points to \( u \), \( PR(u) \) and \( PR(v) \) denotes the rank scores of page \( u \) and \( v \) respectively. \( TL(v) \) denotes total number of outgoing links of page \( v \), \( d \) is a damping factor that is usually set to 0.85. \( d \) can be thought of as the probability of users following the links and (1-d) as the page rank distribution from non-directly linked page.

Page Rank based on Content of the web pages: Web Mining is basically extracting the information from the web. Retrieving the content of a web page is a process of web content mining. While searching for the pages according to user query, only using link based algorithm, cannot be considered an appropriate approach.

User is searching the web for some relevant content that can never be found just listing pages on the basis of link structure. Therefore content based scoring algorithm is important. In our approach we use SimRank algorithm [15] which is a content based algorithm to assign more relevant score to the pages using similarity measure. Pages are scored on the basis of user query and best scoring pages are listed further to create final list of pages. To calculate the rank on the basis of content of web pages, Eq.2 is used.

\[
CPR(u) = \text{occur}(q,u) \times d \times \sum_{u \in D} \text{wpc}(u) \times \text{fieldscore}(f,u)
\]

(2)

where CPR(u) denotes content based page rank of page u. Occur(q, u) denotes query based score that is how many terms of the query are found in the page u. Field score (f, u) defines a score on the basis of the field (f) of page (u) that contains the terms of the query. Note that the content of a crawled page contains two parts or fields: title and body. The weight values assigned to them are different and computed using Eq.3.

\[
wpc(u) = tf (iu) \times idf (i)
\]

(3)

Where tf(iu) denotes term frequency that is number of time term i appears in page u. And idf(i) is the inverse page frequency that is total number of pages divided by pages in which term i appear.

V. RESEARCH METHODOLOGY

Input: Web document \( D = \{D1, D2, \ldots, DN\} \)

Output: Relevant documents.

Step 1: We will create a repository (database) of web pages.

Step 2: After creating the database a link structure will be created that will explain how pages are linked to each other. On the basis of links, page rank will be calculated for each page at the beginning.

Step 3: User will add a query and database will be searched for the pages related to user query.

Step 4: Pages will be searched for user query. Web Pages will be selected on the basis of their similarity content and those are similar to user search will be selected for user. Web Page similarity will be calculated using modified Sim-Rank technique i.e. content based rank.

Step 5: Once we have web pages those are matched with user query, their page ranks will be compared. Pages with high page and content rank will be placed on top of the search result list. To build our final search list we will consider both web pages content and links. HPR denotes hybrid page rank.

\[
\text{HPR}_{vol}(u) = x\% \text{ of } PR(u) \times y\% \text{ of } CPR(u)
\]

(4)

By using Eq.4 a Hybrid Page Rank is computed by combining the percentage of page rank and content based page rank to get a final score and to generate a final list of pages.

VI. CONCLUSION

On the basis of this study we conclude that both link and content based algorithms are important to calculate a final score or page rank of a web page. The usual search engines results in large number of pages in response to user’s queries, while the user always want to get the best in a petite time. In order to rank massive web pages accurately and effectively, we propose a Hybrid Page Rank Algorithm which computes the score on the basis of content as well as link structure of the web pages.

VII. ACKNOWLEDGEMENT

I am extremely grateful and remain indebted to all the people who have given their intellectual support throughout the course of this work. And a special acknowledgement to the authors of various research papers and books which help me a lot.

REFERENCES


Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004 IEEE.


