Web Structure Mining- A Study on Different Page Ranking Algorithms and their Future Improvements

Amar Ichangimath, Chaitra Joshi, Prof. Vathsala M. K

Department of ISE
P.E.S Institute of Technology
Bangalore, Karnataka, India

Abstract

The rapid advent in internet technology has led the users to get easily confused in large hypertext structure. Fetching the relevant information from this huge web of structured data has become the need nowadays. In order to achieve this goal, we employ the concept of web mining. Specifically, we concentrate on a subsidiary of Web Mining: Web Structure Mining which is defined as the process of analysing the structure of hyperlink using graph theory. There are many algorithms for web structure mining such as PageRank Algorithm, HITS, Weighted PageRank Algorithm, Topic Sensitive PageRank Algorithm (TSPR), Weighted Page Content Rank Algorithm (WPCR) etc. In this paper, we have described the outline of all the algorithms, identify their strengths and limitations and also suggest a few future Improvements.

Keywords: *Web Structure mining, HITS, Efficiency, Hyperlink, Weighted PageRank.*

I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined: Web Content Mining, Web Structure Mining and Web Usage Mining.

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR), Natural Language Processing (NLP) and Machine Learning.

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

.1. Hyperlinks: A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an Intra- document hyperlink, and a hyperlink that connects two different pages is called an inter document hyperlink. Hence, with the help of Link Analysis, we have many algorithms to rank Web Pages.

2. Document Structure: Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting Document Object Model (DOM) structures out of documents

.Web Usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications.

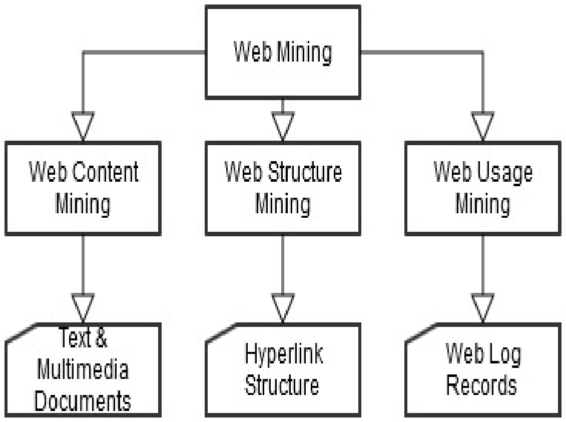

Fig 1. Web Mining and its Branches.

II. LITERATURE SURVEY

Internet search engines are special sites on the Web that are designed to help people find information stored on other sites. There are differences in the ways various search engines work, but they all perform three basic tasks:

- They search the Internet- or select pieces of the Internet -based on important words.
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.

It is very important to note that the link structure of the web serves to bind all of the pages together. Hence, the emphasis is on Web Structure Mining,

Search engines use automated software (known as robots or spiders) to follow links on Websites, harvesting information as they go. When someone submits a query to a search engine, the engine returns a list of sites, ranking them on their relevance to the keywords used in the search. How search engines assess your site and determine the relevance of the words often depends on how the specific search engine works. To understand this further, we need to have a look at the various page ranking algorithms. Without the help of the knowledge of how exactly Search Engines work, we cannot analyse the different algorithms properly.

With respect to Link Analysis in Web Structure Mining, we talk about five important Page ranking algorithms. The various Journal papers referred by talk about these algorithms in detail and give us an insight as to how exactly are we supposed to go about in coming up with a definite analysis for further improvements of these algorithms.

First, we begin with the PageRank algorithm which was developed by Larry Page. This algorithm was the first step in ranking the various web pages which formed the basis for the various other algorithms that were later proposed. Secondly, we have HITS which was developed by Kleinberg which tells us about Hubs and Authorities. After HITS, we look at Weighted PageRank Algorithm which is basically an upgrade to the original PageRank algorithm. After this, we study Weighted Page Content Rank (WPCR) algorithm which makes use of both Web Structure Mining as well as Web Content Mining concepts. Lastly, we have a look at Topic Sensitive PageRank algorithm which is based on the topic sensitivity of user Query.

After we look at these algorithms, we will understand that there is a need to implement efficient algorithms and also improve the user experience at the same time. Let us now look at the various page ranking algorithms now.

III. METHODOLOGY

A. Page Rank Algorithm

PageRank algorithm is a type of link analysis algorithm that was discovered by Larry page, CEO, Google. This algorithm is used by Google internet search engine. In this algorithm numerical weight is assigned to each element of hyperlink set of document such as World Wide Web, with the purpose of measuring the relative importance of that particular set in that hyperlink. PageRank algorithm relies on the concept of Probability Distribution in order to measure the relative importance of a web page in a hyperlink.

Working: This algorithm computes the score of web pages at the time of indexing.

Limitations: Results come at the time of indexing and not at the time of querying. The algorithm is recursive in nature and tends to become less efficient over a period of time.

B. HITS

This algorithm was proposed by Kleinberg in 1997. According to this algorithm first step is to collect the root Set. That root set hits from the search engine. Then the next step is to construct the base set that includes the entire page that points to that root set. The size should be in between 1000-5000. Third step is to construct the focused graph that includes graph structure of the base set. It deletes the intrinsic link, (the link between the same domains). Then, it iteratively computes the hub and authority scores. In HITS concept, he identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to this, a good hub is a page that points to many good authorities; a Good authority is a page that is pointed to by many good hubs". Although HITS provides good search results for a wide range of queries, HITS did not work well in all cases due to the following three reasons:

(1) Mutually reinforced relationships between hosts. Sometimes a set of documents on one host point to a single document on a second host, or sometimes a single document on one host point to a set of document on second host.

(2) Automatically generated links. Web document generated by tools often have links that were inserted by the tool.

(3) Non-relevant nodes. Sometimes pages point to other pages with no relevance to the query topic.

Working: it computes the hubs and authorities of the relevant web pages.

Limitations: Topic Drift and Efficiency Problems.

C. Weighted PageRank Algorithm

This algorithm is an extension of PageRank Algorithm. This algorithm assigns a larger rank values to the more important pages rather than Dividing the rank value of page evenly among its outgoing linked pages, each outgoing link gets a value proportional to its importance.

In this algorithm weight is assigned to both backlink and forward link. Incoming link is defined as number of link points to that particular page and outgoing link is defined as number of links goes out. This algorithm is more efficient than PageRank algorithm because it uses two parameters i.e. Backlink and forward link.

Working: Weight of a web page is calculated on the basis of both ingoing and outgoing links and further the rank is calculated.

Limitations: Relevancy of the web page is completely ignored in this algorithm.

D. Weighted Page Content Rank Algorithm (WPCR)

Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much relevant a page is. Importance refers to the popularity of the web page, that is, how many pages are pointing to it. Relevancy means matching the web page with the relevant query being entered by the user.

Working: It gives different weights to web links based on three important attributes: Relative position of web page, tag where link is present and length of anchor text.

Limitations: Relative position of a web page is not always effective indicating that the logical position of a web page may not always point to the physical page.

E. Topic Sensitive PageRank Algorithm

This algorithm is still being developed in theory. In this algorithm, different scores are computed, multiple important scores for each page under several topics that form composite PageRank score for those pages matching the query. At query time, the similarity of the query is compared to each of these vectors or topics; and subsequently, instead of using a single global ranking vector, the linear combination of the topic- sensitive vectors is weighed using the similarity of the query to the topics. This method yields a very accurate set of results relevant to the context of the particular query. For each web document query sensitive importance score. The results are ranked according to this composite score. It provides a scalable approach for search rankings using Link analysis. For each Web page, compute an importance score per topic. At query time, these importance scores are combined based on the topics of the query and associated context to form a composite PageRank score for those pages matching the query. This score can be used in conjunction with other scoring schemes to produce a final rank for the result pages with respect to the query. This algorithm will improve the order of web pages in the result list so that user may get the relevant pages easily.

Working: it computes the score of a web page based on the topic sensitivity of a web page.

Limitations: Only available to text based web pages as of now.

IV. IMPROVEMENTS

Keeping in mind the different proposed algorithms to rank the web pages; their working and their limitations, we can come up with certain improvements so that they can be implemented more efficiently in the web to enhance the user experience to fetch the required data.

All the algorithms proposed above may provide satisfactory performance in some cases but many times the user may not get the relevant information. The problem we all face when we search a topic in the web using a search engine like Google is that we are presented with millions of search results. First of all it not practically feasible to visit all these millions of web pages to find the required information. Sometimes, we may also not get the relevant information. The major problem is that all these algorithms is that none of them include the “Intelligent Search Factor”. By this, we mean that there is a need for interpreting the inherent meaning of the query and indexing should be based on that.

Hence, we can further concentrate on the last algorithm- ' Topic Sensitive PageRank Algorithm wherein we can concentrate on fetching the relevant content correctly and also fetch that relevant web page efficiently thereby maintaining a perfect balance between the two factors and also keeping in mind, the Intelligent Search factor to improve the user experience.

A large amount of extensive research has to go in analysing all the algorithms for their efficiency because we are supposed to deal with real time data and every day the amount of data that is being uploaded on the web is increasing exponentially. This will lead to lot of future problems while fetching the relevant content. Providing the user with an enhanced experience without compromising on the performance of the algorithms should be considered a priority.

V. CONCLUSIONS

This paper described several proposed web structure mining algorithms like PageRank algorithm, Weighted PageRank algorithm, Weighted Content PageRank algorithm (WPCR), HITS etc. We analysed their working and limitations. Also, we came up with certain improvements to increase the efficiency of the algorithms and also to enhance the user experience. Hence, this paper can be used as a reference to understand the limitations of the various page ranking algorithms and to understand how they can be improved further. Special emphasis is laid on Topic Sensitive PageRank Algorithm since it can be improved further in order to increase the efficiency with which fetch the relevant content that we need rather than just fetch some content which is not at all relevant.

REFERENCES

[1] Preeti Chopra and Md. Ataullah, “A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms”, IJEAT, 2013

[2] Ashish Jain, Rajeev Sharma, Gireesh Dixit and Varsha Tomar, “Page Ranking Algorithms in Web Mining, Limitations of Existing Methods and a New Method for Indexing Web Pages”,International Conference on Communication Systems and Network Technologies, 2013

[3] Mojtaba Rezvani and S. Mehdi Hashemi, "Enhancing Accuracy of Topic Sensitive PageRank", International Conference on Web Intelligence, 2012

[4] Pooja Sharma, Deepak Tyagi and Pawan Bhadana, "Weighted Page Content Rank for Ordering Web Search result", IJEST,2010

[5] Mishra Shesh Narayan et al (2010), "An Effective algorithm for web mining based on Topic sensitive PageRank algorithm", International journal of computer science and software engineering.

[6] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", in proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[7] M. G da Gomes Jr. and Z.Gong,"Web Structure Mining:An Introduction", Proceedings of the IEEE International Conference on Information Acquisition,2006

[8] Mukhopadhyay et al.(2006)," A Syntactic Classification based Web Page Ranking Algorithm", 6th International Workshop on MSPT Proceedings.

[9] http://www.csbdu.in/econtent/Data%20Mining%20&%20Warehousing/Unit%20III.pdf