International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-2, Issue-2, March-2014

Enhancement of the Web Search Engine Results using Page Ranking Algorithm

Nilima V. Pardakhe, Prof. R. R. Keole

Abstract— **As web is the largest collection of information and plenty of pages or documents are newly added and deleted on frequent basis due to the dynamic nature of the web. The information present on the web is of great need, the world is full of questions and the web is serving as the major source of gaining information about specific query made by the user. Search engines generally return a large number of pages in response to user queries. To assist the users to navigate in the result list, ranking methods are applied on the search results. Most of the ranking algorithms proposed in the literature are either link or content oriented, which do not consider user usage trends. In this paper, a page ranking mechanism called Page Ranking based on Visits of Links is being devised for search engines, which works on the basic ranking algorithm of Google i.e. PageRank and takes number of visits of inbound links of Web pages into account. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduces the search space to a large scale.**

Index Terms— **Information Retrieval, PageRank, Search Engine, Web Mining, World Wide Web.**

I. INTRODUCTION

The World Wide Web consists billions of web pages and huge amount of information available within pages. To retrieve required information from World Wide Web, search engines perform number of task based on their respective architecture. These can be complicated and time consuming processes. Every search engine process goes from Crawling, Indexing, Searching, and Sorting/Ranking of information. A Crawler visits and downloads all the webpage of the website and retrieve information needed from them. The information provided by Crawler has to be stored in some order to be accessed by the search engine; the information is indexed in order to decrease the time needed to look into it.

Today, the World Wide Web is the popular and interactive medium to disseminate information. The Web is huge, diverse and dynamic. The Web contains vast amount of information and provides an access to it at any place at any time. The most of the people use the internet for retrieving information. But most of the time, they gets lots of insignificant and irrelevant document even after navigating several links. For retrieving information from the Web, Web mining techniques are used.

A. Web Mining Overview

Web mining is an application of the data mining techniques to automatically discover and extract knowledge from the Web. Web mining consists of the following tasks:
Resource finding**:** the task of retrieving intended Web documents. [7] [10] [11]
Information selection and pre-processing**:** automatically selecting and pre-processing specific information from retrieved Web re-sources.
Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
Analysis: validation and/or interpretation of the mined patterns.
There are three areas of Web mining namely, Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM).

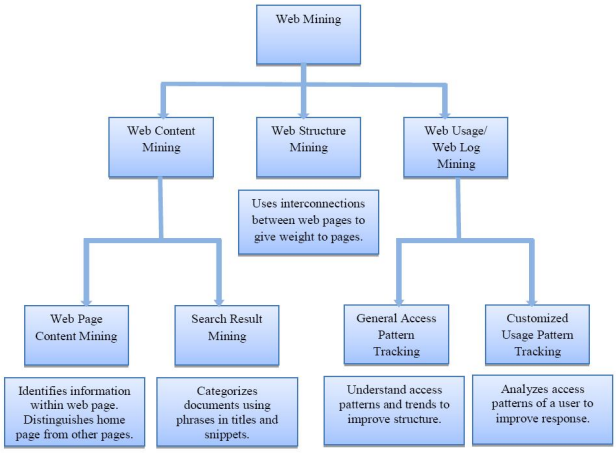

Fig1: Web Mining Taxonomy

A. Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of web documents. The web documents may consists of text, images, audio, video or structured records like tables and lists. Mining can be applied on the web documents as well the results pages produced from a search engine. There are two types of approach in content mining called agent based approach and database based approach. The three types of agents are Intelligent search agents, Information filtering/Categorizing agent, Personalized web agents. Intelligent Search agents automatically searches for information according to a particular query using domain

characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine instructions. Personalized web agents learn user preferences and discovers documents related to those user profiles. In Database approach it consists of well formed database containing schemas and attributes with defined domains. Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data. [10]

- Unstructured Data Mining Techniques: Content mining can be done on unstructured data such as text. Mining of unstructured data give unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques. Basic Content Mining is a type of text mining. Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization.
- Structured Data Mining Techniques: The techniques used for mining structured data are Web Crawler, Wrapper Generation, Page content Mining.
- Semi-Structured Data Mining Techniques: The techniques used for semi structured data mining are Object Exchange Model (OEM), Top Down Extraction, and Web Data Extraction language.
- Multimedia Data Mining Techniques: Some of the Multimedia Data Mining Techniques are SKICAT, color Histogram Matching, Multimedia Miner and Shot Boundary Detection.

B. Web Usage Mining (WUM)

Web Usage Mining is the process of extracting useful information from the secondary data derived from the interactions of the user while surfing on the Web. It extracts data stored in server access logs, referrer logs, agent logs, client-side cookies, user profile and metadata. [7]

The challenges involved in web usage mining could be divided in three phases:

1. Pre-processing. The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. It includes data cleaning, data integration, data transformation and data reduction.

2. Pattern discovery. Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns.

3. Pattern Analysis. This process targets to understand, visualize and give interpretation to these patterns. [13]

C. Web Structure Mining

The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page. It tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and generate the information like similarity and relationship between different Web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level (inter-page). It is important to understand the Web data structure for Information Retrieval. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents.

The objects in the WWW are web pages, and links are in, out and co-citation i.e. two pages that are both linked to the same page. There are some possible tasks of link mining which are applicable in Web structure mining and are described as follows: [2] [13]

1. Link-based Classification: - is the most recent upgrade of a classic data mining task to linked Domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

2. Link-based Cluster Analysis. The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. Link Type. There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

4. Link Strength. Links could be associated with weights.

5. Link Cardinality. The main task here is to predict the number of links between objects. There are some uses of web structure mining like it is:

- Used to rank the user's query
- Deciding what page will be added to the collection
- Page categorization
- Finding related pages
- Finding duplicated web sites
- And also to find out similarity between them

B. Page Ranking Algorithms

An efficient ranking of query words has a major role in efficient searching for query words. There are various challenges associated with the ranking of web pages such that some web pages are made only for navigation purpose and some pages of the web do not possess the quality of self descriptiveness. For ranking of web pages, several algorithms are proposed in the literatures.[7]

Three important algorithms are

- PageRank
- Weighted PageRank
- HITS (Hyper-link Induced Topic Search)

Two graph based page ranking algorithms i.e. google Page Rank proposed by Brin and Page in 1998 and Kleinberg's hypertext induced topic selection (HITS) algorithm proposed by Kleinberg in 1998 are used successfully and traditionally in the area of web structure mining. Both of these algorithms give equal weights to all links for deciding the rank score.

II. LITERATURE REVIEW

The web search engine represents the user interface needed to permit the user to query the information. It is the connection between user and the information repository when user sends query to search engine, then there exist incredible number of web pages related to given query. But only small number of web pages is really needed to the user. Still this number is very large (in millions). Search engine uses ranking algorithm in order to sort the results to be displayed. That way user will have the most important and useful result first. There are various ranking algorithms developed, few of them are PageRank, HITS, SALSA, RANDOMZE HITS, SUBSPACE HITS, SIMRANK etc. In this paper we will focus only on PageRank and a proposed improvement of PageRank.

A. PageRank Algorithm

The PageRank algorithm assigns a PageRank score to more than 25 billion web pages on the WWW. During the processing of a query, Google's search algorithm combines precomputed PageRank scores with text matching scores to obtain an overall ranking score for each web page. Although overall ranking is determined by considering many other factors but Google claims that the heart of its search engine software is PageRank. A simplified version of PageRank is defined as follows.

$$PR(u) = c \sum_{v \in B(u)} PR(v) / N_V$$

Where u represents a web page, B(u) is the set of pages that point to u. PR(u) and PR(v) are rank scores of page u and v, respectively. N_v denotes the number of outgoing links of page v, c is a factor used for normalization. In PageRank, the rank score of a page, p, is evenly divided among its outgoing links. The values assigned to the outgoing links of page p are in turn used to calculate the ranks of the pages to which page p is pointing. Later PageRank was modified observing that not all users follow the direct links on WWW. The modified version is given in following equation

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) / N_V$$

Where d is a dampening factor that is usually set to 0.85. d can be thought of as the probability of users' following the links and could regard (1 − d) as the page rank distribution from non-directly linked pages.

B. Weighted PageRank Algorithm

Wenpu Xing and Ali Ghorbani proposed an extension to standard PageRank called Weighted PageRank (WPR). It assumes that more popular the web pages are, more linkages other web pages tend to have to them or are linked to by them. This algorithm assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its outgoing linked pages. Each outlink page gets a value proportional to its popularity. The popularity is measured by its number of inlinks and outlinks. The popularity from the number of inlinks and outlinks is recorded as $W^{in}(v, u)$ and $W^{out}(v, u)$, respectively. $W^{in}(v, u)$ is the weight of *link* (v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.

$$W^{in}_{(v,u)} = I_u / \sum_{p \in R(v)} I_p$$

Where I_u and I_p represent the number of inlinks of page u and page p, respectively. $R(v)$ denotes the reference page list of page v. $W^{out}(v, u)$ is the weight of link (v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v.

$$W^{out}_{(v,u)} = O_u / \sum_{p \in R(v)} O_p$$

Where O_u and O_p represent the number of outlinks of page u and page p, respectively. Considering the importance of pages, the original PageRank formula is modified.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{out}_{(v,u)} W^{in}_{(v,u)}$$

Wenpu Xing and Ali Ghorbani proposed a Weighted PageRank algorithm which is an extension of the 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. The popularity from the number of in links and out links is recorded as W_{in} and W_{out} respectively. W_{in} (v, u) is the weight of link (v, u) calculated based on the number of in links of page u and the number of in links of all reference pages of page v. [2][3]

C. HITS (Hyper-link Induced Topic Search)

This algorithm was given 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 him, 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][13]

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 a 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.

D. Topic Sensitive PageRank

In Topic Sensitive PageRank, several scores are computed: multiple importance scores for each page under several topics that form a composite PageRank score for those pages matching the query. During the offline crawling process, 16 topic-sensitive PageRank vectors are generated, using as a guideline the top-level category from Open Directory Project (ODP). 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.

III. ANALYSIS OF PROBLEM

All the algorithms such as PageRank (PR), Weighted PageRank (WPR), and Hyperlink-Induced Topic Search (HITS) etc. 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. It is not practically feasible to visit all these millions of web pages to manually find the required information [1].

When we visit few initial links shown in the search results, we may not get the relevant information. Therefore, we feel the requirement of a mechanism so that we can get the relevant information according to the query posted by us. By Relevant search we mean that there is a need for interpreting the inherent meaning of the query and indexing should be based on that.

Again the difficulties are:

- The major source of information is from web. Manually extracting real required information is difficult to read and analyze.
- Many search engines gives a long list of documents and most of them are irrelevant.

So the main problem is ranking of web documents to improve search engine results. The PageRank algorithm can reflect link relationships between web pages on Internet to certain extent, and it can further dig out the importance of Web pages effectively, however, there are still some limitations.

IV. PROPOSED WORK

Due to the increasing amount of data available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. The standard search engines usually result in a large number of pages in response to users' queries, while the user always desires to get the best in a petite time. Therefore there is need to apply the techniques from web mining and machine learning to web data and documents which plays an important role in identifying the relevant web page. Relevancy of web page denotes how well a retrieved web page or set of web pages meets the information need of the user. To assist the users to navigate in the result list, ranking methods are applied on the search result. Page Ranking based on Visits of Links serves the purpose of bringing the most important web pages or information in front of users,

As Page Ranking based on Visits of Links uses link structure of pages and their browsing information, the top returned pages in the result list are supposed to be highly relevant to the user information needs. A link with high probability of visit contributes more towards the rank of its out linked pages. The rank value of any page by PageRank method will be same either it is seen by user or not, because it is totally dependent upon link structure of Web graph. While the ordering of pages using Visits of Links is more target-oriented. In Page Ranking based on Visits of Links , a user can not intentionally increase the rank of a page by visiting the page multiple times because the rank of the page depends on the probability of visits (not on the count of visits) on back linked pages. The main issue to address is the periodic crawling of web servers so as to collect the accurate and up to date visit count of pages. Specialized crawlers need to be designed for fetching the required information of pages.

V. CONCLUSION

As years passed World Wide Web became overloaded with information and it became hard to retrieve data according to the need. The goal of search engines is to provide relevant information to the users to cater to their needs. Therefore, finding the content of the Web and retrieving the users' interests and needs have become increasingly important. The different algorithms used for link analysis like PageRank (PR), Weighted PageRank (WPR), Hyperlink-Induced Topic Search (HITS) algorithms are discussed. Page Ranking based on Visits of Links calculates rank value of a web page based on the user visits on incoming links of that page. The ordering of pages in this way increases the relevancy of pages and thereof provides the user with quality search results. As a result, user may find the desired content in the top few pages, thus search space can be reduced to a large scale.

REFERENCES

[1] Ashish Jain, Rajeev Sharma, Gireesh Dixit, Varsha Tomar ," Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", 2013 IEEE International Conference on Communication Systems and Network Technologies.

[2] Seifedine Kadry , Ali Kalakech ," On the Improvement of Weighted Page Content Rank", *Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013.*

[3] Rashmi Rani, Vinod Jain ," Weighted PageRank using the Rank Improvement" International Journal of Scientific and Research Publications, Volume 3, Issue 7, July 2013.

[4] Preeti Chopra, Md. Ataullah ,"A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms", International Journal of

Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.

[5] B.Aysha Banu, Dr.M.Chitra.," A Novel Ensemble Vision Based Deep Web Data Extraction
Technique for WebMining Applications", 2012 IEEE Intemational Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

[6] P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar," Content Based Ranking for Search Engines",Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol I, Hong Kong.

[7] Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2670-2676.

[8] Mohamed-K HUSSEIN, Mohamed-H MOUSA ," An Effective Web Mining Algorithm using Link Analysis", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (3) , 2010, 190-197.

[9] Shesh Narayan Mishra, Alka Jaiswal, Asha Ambhaikar ," Web Mining Using Topic Sensitive Weighted PageRank", International Journal of Scientific & Engineering Research Volume 3, Issue 2, February-2012 , ISSN 2229-5518.

[10] Faustina Johnson , Santosh Kumar Gupta," Web Content Mining Techniques: A Survey", *International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012*.

[11] Shesh Narayan Mishra ,Alka Jaiswal,Asha Ambhaikar ," An Effective Algorithm for Web Mining Based on Topic Sensitive Link Analysis ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012 ISSN: 2277 128X.

[12] V. Lakshmi Praba , T. Vasantha," EVALUATION OF WEB SEARCHING METHOD USING A NOVEL WPRR ALGORITHM FOR TWO DIFFERENT CASE STUDIES "Ictact Journal on Soft Computing, April 2012, Volume: 02, Issue: 03.

[13] Miguel Gomes da Costa, Júnior Zhiguo Gong," Web Structure Mining: An Introduction", Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China.

[14] Neelam Tyagi, Simple Sharma," Comparative study of various Page Ranking Algorithms in Web Structure Mining (WSM)" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-1, June 2012.

[15] Ms.N.Preethi , Dr.T.Devi," New Integrated Case And Relation Based (CARE) Page Rank Algorithm" 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 – 06, 2013, Coimbatore, INDIA.