ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Increasing performance in Information Retrieval Using Extended Agent Based Weighted Page Ranking Algorithm

S.Ganesan
M.Phil Research Scholar,
Department of Computer Science and Engineering,
Alagappa University, Karaikudi, Tamilnadu
ganesangk15@gmail.com

Abstract— **Web mining is most popular application of Data mining means search important and useful information from web sites.Find information in the form of document, connected number of document with hyperlink, statistical graph, image, video clip etc.When a user refers a query to the search engine, it generally returns a large number of pages in response to user's query. Information Retrieval (IR) is a significant area in web mining where the users procure their needed information from the web. Web content mining(WCM) goads this problem with the help of agent by retrieving explicit information from different web sites for its access and knowledge discovery. Page Rank focuses on ranking a page based on the number of inlinks and outlinks to a page. Most of the search engines are ranking their search results in response to users queries to make their search navigation easier. This paper also explores increasing performance in Information Retrieval using Extended Agent Based Weighted Page Ranking Algorithm for web content mining to retrieve more relevant information. So Extended AWPR Algorithm retrieve the most important and relevant content information or web pages using their web content and links in front of end users.**

Keywords—Webmining, Agent based weighted page rank Algorithm, Page Rank

I.INTRODUCTION

The World Wide Web (WWW) is rapidly growing on all aspects and is a massive, explosive, diverse, dynamic and mostly unstructured data repository. As on today WWW is the huge information repository for knowledge reference. There are a lot of challenges in the Web: Web is large, Web pages are semi structured, and Web information tends to be diversity in meaning, degree of quality of the information extracted and the conclusion of the knowledge from the extracted information [1]. So it is important to understand and analyze the underlying data structure of the Web for efficient Information Retrieval. Many of the existing Information Retrieval Systems still relies on various approach of ranking algorithms, like Content-based ranking algorithms that use the words in each document to determine its ranking; Link-based ranking algorithms assign scores to web pages based on the number and quality of hyperlinks between pages. Links that point to a particular page or endorse a page can help to improve link based rankings; Usage-based ranking algorithms score documents by how often they are viewed by Internet users.

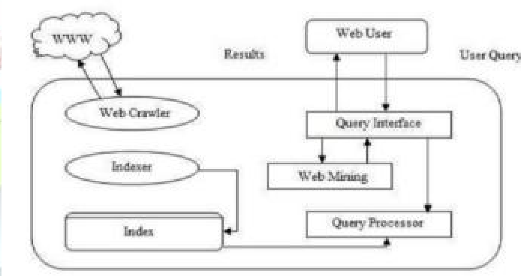

Fig.1 Architecture of Search Engine

Web Mining is categories into Web Structure Mining (WSM), Web Usage Mining (WUM), Web Content Mining (WCM) types [2].Web Structure Mining gathering the structure of the web consider it as a graph. Web Structure Mining can be used to ranking pages present in the web, to improve the effectiveness of search engines.

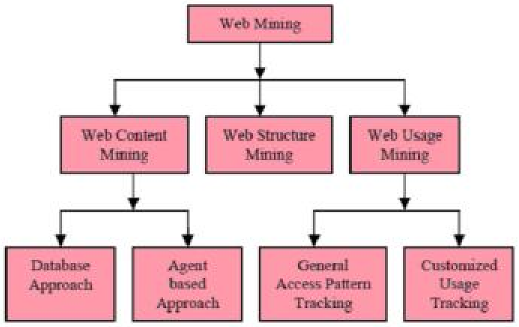

Fig.2 Web mining Classification Structure

Web structure mining tries to discover useful knowledge from the structure of hyperlinks which helps to investigate the node and connection structure of web sites. According the type of web structure data can be divided into two kinds 1)Web contains extracting the documents from hyperlinks 2) Page structure is analysis of the tree-like structure. Based on the hyperlinks topology , web structure mining(WCM) will formed the web pages and invoked the information, such as the similarity and mining is concerned with the retrieval of information from WWW into more structured form and indexing the information to retrieve it quickly. Web classification Structure is shown in Fig 1.

A.Web Content Mining (WCM)

WCM is the process of extracting useful information from the contents of Webpages. Content data corresponds to the collection of facts a Web page was designed to communicate to the users.Web content mining is related to data mining because many data mining techniques can be applied in web content mining.

B.Web Usage Mining (WUM)

WUM is the application of data mining techniques to discover custom patterns from Web datain order to understand and better serve needs of Web based applications. It consists of three stages, namely preprocessing, pattern discovery, and pattern analysis.

C.Web Structure Mining (WSM)

WSM is to generate structural summary about the website and web page. The first kind of web structure mining is extracting patterns from hyperlinks in the web. The other kind of the web structure mining is mining the document structure. This type of mining can be performed at the document level (intra-page) or at the hyperlink level (interpage). It is important to understand the Web data structure for Information Retrieval.

II.RANKING ALGORITHMS

A.Page Rank

Page Rank is a numeric value that represents how important a page is on the web. PageRank is the Best algorithm of evaluate a page's "value." When all other elements such as Title tags and keywords are taken into account, Search Engines uses Page Rank to adjust results so that more "value" pages move up in the results page of a user's search result display. Search Engine numbers that when a page links to another page, it is efficiently casting a vote for the other page. Google calculates a page's importance from the votes cast for it with the help agent. Every vote is taken into account when a page's PageRank is calculated [3][4]. It things because it is one of the elements that determine a page’s ranking in the search results. It isn't the only factor that Search Engines uses to rank pages, but it is an important one. The order of ranking in Search Engines works like this:

1) Find all pages matching the keywords of the search.
2) Adjust the results by PageRank scores.

B.HITS Algorithm

HITS algorithm identifies two different forms of Web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many good hub pages on the same subject. In this a page may be a good hub and a good authority at the same time. This circular relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Selection). HITS algorithm is ranking the web page by using inlinks and outlinks of the web pages. In this a web page is named as authority if the web page is pointed by many hyper links and a web page is named as hub if the page point to various hyperlinks.

C.Weighted Page Rank

Weighted Page Rank [5] Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original pagerank algorithm. WPR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages. This algorithm gives high value of rank to the more popular pages and does not equally divide the rank of a page among its out-link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links. Simulation results show that WPR algorithm finds larger number of related pages compared to standard page rank algorithm.

Original Weighted PageRank formula is

$$\mathbf{PR(U)=(1-d)+d\sum PR(V)W^{in}(U,V)W^{out}(U,V)}$$

V€B(u)

D.Agent based Weighted Pagerank

Agent based Weighted Page Rank Algorithm (AWPR) is a proposed page ranking algorithm which is used to give assorted order to the web pages with an help of agent and returned by a search engine in response to a user query. AWPR is a numerical value based on which the web pages are given an order. This algorithm is used for web structure mining as well as web content mining techniques. Web structure mining is used to compute the importance of the page and web content mining is used to check the page is how much related. Importance here means the popularity of the page.

III.PROPOSED WORK

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. Among a view to resolve the problems found in both algorithms, a new Enhanced algorithm called Extended Agent based Weighted PageRank has been proposed which is used for Web structure mining as well as Web content mining and Query relevance Strings with the help of agents. This algorithm is designed at helpful the order of the pages in the result list so that the user may get the relevant and important pages easily in the list.

A.Modified Search Engine Architecture

Search engines [6] are the key to finding specific information on the vast span of the World Wide Web.There are at least three elements which contain important detection of the database, the user search, arrangement and ranking of results. At the same time we focus density of contents based on query words. With the proposed Extended AWPR, the search engine architecture is modified so as to add the components for calculating importance and relevancy of pages and contents based on Queries.

B. Extended AWPR Algorithm

Algorithm: Extended AWPR

Input: Page P, In-link and Out-link Weights of
all back-links of P, Query Q, d (damping factor).
Output: Rank score

Step 1: Relevance calculation:

a) Find all meaningful word strings of Q (say N)
b) Find whether the N strings are occurring in P or not?
c) Z= Sum of frequencies of all N strings.
d) S= Set of the maximum possible strings occurring inP.
e) X= Sum of frequencies of strings in S.
f) Content Weight(CW)= X/Z
g) C= No. of query terms in P
h) D= No. of all query terms of Q while ignoring stopwords.

i) Probability Weight(PW)= C/D
j)Find query relevant words in all contents QE

Step 2: Rank calculation:

a) Find all backlinks of P (say set B).

b) PR(P)= (1-d) + d [™PR(V) dWin(P,V)
v€Bu
Wout(P,V)] (CW+PW)

c) Output PR(P) i.e. the Rank score

The formula to calculate the Enhanced Weighted Page Rank of a page U is

PR(U)= (1-d)+ d∑PR(V) Win(P,V)Wout(P,V)*(CW+PW+QW)
v€Bu

Here PR(U)=PageRank of page U,
B(U)= Set of all pages referring to page U.
D= Damping factor which can be set between 0 and 1
Win(U,V)= in-weight of link (U,V)
Wout(U,V)= out-weight of link (U,V),
Cw=Content weight of page U
Pw=Probability weight of page U

Qw=Query Relevance content weight of the Page U

IV.EVALUATION

The Calculation part of the algorithm are explained under in detail.

A.Weight calculation

The Win(v,u) and Wout(v,u) are the preprocessed weights. The weights are just given as input to the algorithm. Win(v,u) is the weight of the 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 and is given in eq (1.1).

$$Win(U,V) = 1.1\ \frac{IU}{\sum Ip_{P\text{€}R(V)}}$$

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 The Wout(v,u) is the weight of link(v, u) calculated based on the number of out-links of page u and the number ofOut-links of all reference pages of page v given in eq (1.2).

$$Wout(U,V) = 1.2\ \frac{Ou}{\sum Op_{P\text{€}R(V)}}$$

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

B.Relevance Calculation

Relevance calculator agent which calculates the relevance of a page on the fly in terms of two factors: one represents the probability of the query in the page and other gives the maximum matching of the query to the page. Probability Weight: It is the probability of the query terms in the web page. This aspect is the ratio of the query terms present in the document and the total number of terms in the fired query.

The formula is given in eq (2.1)

Probability weight (PWi) =Yi/N (2.1)

where Yi= Number of query terms in ith document., N=sum of terms in query

Content Weight: It is the weight of content of the web page with respect to query terms. This aspect is the ratio of the sum of frequencies of highest possible query strings in order and sum of frequencies of all query strings in order. The maximum possible strings are selected in such a way that all such strings represent a different logical combination of words.
The formula is given in eq (5.2).

Content weight (CWi) = Xi/M (5.2)

Where Xi= Total number of cardinalities of highest possible query strings in order
M= Total number of cardinalities of all possible significant query strings in order

V.CONCLUSION

Web Mining is the use of the data mining techniques to automatically discover and extract information from web documents/services. Due to ever increasing information present on the web, the users have to spend lot of time to retrieve relevant information to them. The PageRank and Weighted PageRank algorithms are used by many search engines but the users may not get the required relevant documents easily on the top few pages. This Extended Agent based weighted page rank algorithm is intended at improving the order of the pages in the exact result list with the help of an agent so that the user may get the relevant and important pages easily in the list.

References

[1] A. Scime, Web mining: Applications and Techniques, London: Idea Group Publishing, 2005, ch. 5.

[2] R. Kosala and H. Blockeel."Web mining research": A survey.ACM SIGKDD Explorations, 2(1):1–15, 2000.

[3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] POOJA SHARMA B.S, DEEPAK TYAGI St. Weighted Page Content Rank for Ordering Web Search Result. International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7301- 7310

[5] Wenpu Xing and Ali Ghorbani.Weighted PageRank Algorithm, Proceedings of the Second Annual Conference on (CNSR'04) IEEE

[6] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Science Department, Stanford University, Stanford, CA 94305.