ISSN:2249-5789

Improving the Efficiency of Weighted Page Content Rank Algorithm using Clustering Method

Shruti Aggarwal
Assistant Professor,
Dept. of CSE, S.G.G.S.W.U.,
Fatehgarh Sahib (Punjab), India
shruti_cse@sggswu.org

Gurpreet Kaur
Research Scholar,
Dept. of CSE, S.G.G.S.W.U.,
Fatehgarh Sahib (Punjab), India
ggill208@gmail.com

ABSTRACT

Web mining is defined as the application of data mining techniques on the World Wide Web to find hidden information. Most of the search engines are ranking their search results in response to users' queries to make their search navigation easier. It includes Link Analysis algorithms i.e., Page Rank, Weighted Page Rank and Weighted Page Content Rank. PageRank is a commonly used algorithm in Web Structure Mining. Weighted Page Rank also takes the importance of the inlinks and outlinks of the pages but the rank score to all links is not equally distributed. i.e. unequal distribution is performed. Weighted Page Content Rank based on web content mining and structure mining that shows the relevancy of the pages to a given query is better determined, as compared to the existing PageRank and Weighted PageRank algorithms. The implementation of Weighted Page Content Rank algorithm has been carried out and the result shows that it takes too much time to process the Web Pages. As the number of Web Pages increases, there is also increase in the processing time. So, in this paper, the Fuzzy Logic is implemented on WPCR algorithm to decrease the processing time of Web pages.

Keywords:- Web Mining, Page Rank, Weighted Page Rank, Weighted Page Content Rank, Fuzzy K Mean, Fuzzy C Mean.

I INTRODUCTION

The *World Wide Web* is the collection of information resources on the Internet that are using the Hypertext Transfer Protocol. It is a repository of many interlinked hypertext documents, accessed via the Internet. Web may contain text, images, video and other multimedia data. In order to analyze such data, some techniques called web mining techniques are used by various web applications and tools. Web mining describes the use of data mining techniques to automatically discover Web documents and services, to extract information from the Web resources and to uncover general patterns on the Web. Over the years, Web mining research has been extended to cover the use of data mining and similar techniques to discover resources, patterns, and knowledge from the Web-related data (such as Web usage data or Web server logs).It is used to understand customer behavior, evaluate the effectiveness of a particular Web and help quantify the success of a marketing campaign. It is a rapidly growing research area.

II Web Mining

In 1996 it's Etzioni who first coined the term web mining. Etzioni starts by making a hypothesis that information on web is sufficiently structured and outliers the subtasks of web mining.[1].It refers to overall process of discovering potentially useful and previously unknown information from web document and

services web mining could be viewed as an extension of standard data mining to web data.

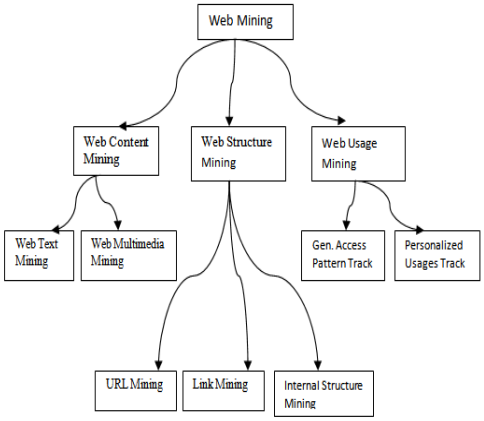

Fig.1. Web Mining Categories.

Web Mining Process: - Web Mining can be decomposed into the following subtasks:-

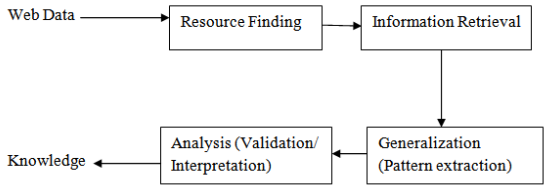

Fig.2. Web Mining Subtasks [1].

a) **Resource Finding**: the function of retrieving relevant web documents.
b) **Information Selection and Pre-processing**: the automatic selection and preprocessing of specific information from retrieved web resources.
c) **Generalization**: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine Learning are used in generalization.
d) **Analysis**: the validation and interpretation of the mined patterns. It plays an important role in pattern mining.

A. Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content mining is related but is different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in web content mining.

B. Web Usage Mining

Web usage mining is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve needs of Web based applications. It consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Web servers, proxies, and client applications can quite easily capture data about Web usage.

C. Web Structure Mining

The goal of web structure mining 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. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure.

III Link Analysis Algorithm

Web mining technique provides the additional information through hyperlinks where different documents are associated. The web can be examined as a directed labeled graph whose node is the credentials or pages and edges are the hyperlinks between them. This directed graph configuration is known as web graph. There are several algorithms proposed based on link analysis. The important algorithms Page Rank, Weighted Page Rank, Weighted Page Content Rank.

A. Page Rank

Page Rank is a numeric value that represents how important a page is on the web. Page Rank is the Google's method of measuring a page's "importance." When all other factors such as Title tag and keywords are taken into account, search engine uses Page Rank to adjust results so that more "important" pages move up in the results page of a user's search result display. Search Engine Figs that when a page links to another page, it is effectively casting a vote for the other page. It calculates a page's importance from the votes cast for it. How important each vote is taken into account when a page's Page Rank is calculated. It matters because it is one of the factors that determine a page's ranking in the search results. It isn't the only factor that Google 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 Page Rank scores.

Page Rank takes the back links into account and propagates the ranking through links. A page has a higher rank, if the sum of the ranks of its backlinks is high. The original Page Rank algorithm is given in following equation

PR(P)=(1-d)+d(PR(T1)/C(T1)+…..PR(Tn) / C(Tn))…. (1)

Where,
PR (P) = Page Rank of page P
PR (Ti) = Page Rank of page Ti which link to page
C (Ti) = Number of outbound links on page T
D = Damping factor which can be set between 0 and 1.

B. Weighted Page Rank

Weighted PageRank algorithm (WPR).This algorithm is an extension of Page Rank algorithm. WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. WPR performs better than the conventional Page Rank algorithm in terms of returning larger numbers of relevant pages to a given query. According to author the more popular web pages are the more linkages that other WebPages tend to have to them or are linked to by them. The proposed extended Page Rank algorithm–a Weighted Page Rank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as Win (v, u) and Wout (v, u), respectively. WPR supplies the most important web pages or information in front of users.

Original Weighted PageRank formula is

$$PR(U)=(1-d)+d\sum_{v\in B(u)} PR(V)\, W^{in}(U,V)W^{out}(U,V) \quad \ldots (2)$$

C. Weighted Page Content Rank

Weighted Page Content Rank Algorithm (WPCR) [2] 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 here means the

popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of in links and out links of the page. Relevancy means matching of the page with the fired query. If a page is maximally matched to the query, that becomes more relevant.

D. HITS Algorithm

The HITS algorithm is proposed by Kleinberg in 1988. 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. An Illustration of hub and authority are shown in figure 3.

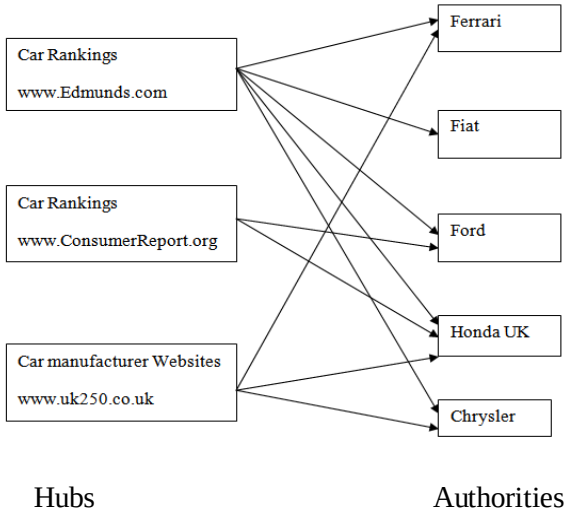

Fig.3. Illustration of hubs and authorities [3].

HITS Algorithm

1. Initialize all weights to 1
2. Repeat until the weights converge
3. for every p∈H
4. Hp=$\sum_{q \in Ip} A_q$
5. For every authority p∈ A
6. Ap=$\sum_{q \in Bp} H_q$
7. Normalize

E. Comparison of Page Ranking Algorithms

ISSN:2249-5789

Algorithm	**Page Rank**	**HITS**	**WPR/WPCR**
Main Technology	WSM	WSM,WCM	WSM
Input Parameter	Backlinks	Content, Back & Forward links	Backlinks & Forward links
Relevancy	Less(rank the pages on the indexing time)	More(use the hyperlinks)	Less as ranking is based on the calculation of weight of the web page at the time of indexing
Methodology	This algorithm computes the score for pages at the time of indexing of pages	It computes the hubs and authorities of the relevant pages. It relevant as well as important page as the result	Weight of web page calculated on the basis of input & outgoing links & on the basis of weight the importance of page is decided
Quality of Results	Medium	Less than PR	Higher than PR
Importance	High. Backlinks are considered	Moderate. Hub & authorities scores are utilized	High. The pages are sorted according to the importance
Limitation	Results come at the time of indexing & not at the query time	Topic drift & efficient problem	Relevancy is ignored

Table 1: Comparison of Page Ranking Algorithms [4].

IV Ranking and Clustering

Ranking retrieval systems have also been closely associated with clustering. Early efforts to improve the efficiency of ranking systems for use in large data sets proposed the use of clustering techniques to avoid dealing with ranking the entire collection. It was also suggested that clustering could improve the performance of retrieval by pregrouping like documents.

A. Fuzzy K-Means Clustering Algorithm

The fuzzy *k*-means clustering algorithm partitions data points into *k* clusters S_l (l = 1,2, …, k) and clusters S_l are associated with representatives (cluster center) **C***l*. The relationship between a data point and cluster representative is fuzzy. That is, a membership *uij* ϵ [0, 1] is used to represent the degree of belongingness of data point **X***i* and cluster center **C***j*. Denote the set of data points as $S = \{\mathbf{X}i\}$. The FKM algorithm is based on minimizing the following distortion:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{i,j}^{m} d_{ij} \quad \ldots\ldots (3)$$

with respect to the cluster representatives **C***j* and memberships *uij*, where N is the number of data points; m is the fuzzifier parameter; k is the number of clusters; and dij is the squared Euclidean distance between data point **X***i* and cluster representative **C***j*. It is noted that *uij* should satisfy the following constraint:

$$\sum_{j=1}^{k} u_{i,j} = 1, \text{ for } i = 1 \text{ to } N. \quad \ldots (4)$$

The major process of FKM is mapping a given set of representative vectors into an improved one through partitioning data points. It begins with a set of initial cluster centers and repeats this mapping process until a stopping criterion is satisfied. It is supposed that no two clusters have the same cluster representative. In the case that two cluster centers coincide, a cluster center should be perturbed to avoid coincidence in the iterative process. If $dij < \eta$, then $uij = 1$ and $ui,l = 0$ for $l \neq j$, where η is a very small positive number.

The fuzzy *k*-means clustering algorithm is now presented as follows.

(1) Input a set of initial cluster centers $SC0 = \{\mathbf{C}j(0)\}$ and the value of ε. Set $p = 1$.

(2) Given the set of cluster centers *SCp*, compute *dij* for $i = 1$ to N and $j = 1$ to k. Update membership's *uij* using the following equation:

$$u_{i,j} = \left((d_{ij})^{1/m-1} \sum_{l=1}^{k} \left(\frac{1}{d_{il}} \right)^{1/m-1} \right)^{-1} . \quad \ldots (5)$$

If $dij < \eta$, set $uij = 1$, where η is a very small positive number.

(3) Compute the center for each cluster using Eq. (6) to obtain a new set of cluster representatives *SCp*+1.

$$\mathbf{C}_j(p) = \frac{\sum_{i=1}^{N} u_{ij}^m \mathbf{X}_i}{\sum_{i=1}^{N} u_{ij}^m} \quad \ldots.. (6)$$

(4) If $\|Cj(p) - Cj(p-1)\| < \varepsilon$ for $j = 1$ to k, then stop, where $\varepsilon > 0$ is a very small positive number. Otherwise set $p + 1 \rightarrow p$ and go to step 2.

The major computational complexity of FKM is from steps 2 and 3. However, the computational complexity of step 3 is much less than that of step 2. Therefore the computational complexity, in terms of the number of distance calculations, of FKM is $O(Nkt)$, where t is the number of iterations.

Limitations of K-means

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes
- Problems with outliers
- Empty clusters

Advantages of Fuzzy C-Mean over K-mean

1) Gives best result for overlapped data set and comparatively better then k-means algorithm.
2) Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.

B. Fuzzy C-means Clustering Algorithm

The concept of Fuzzy Logic (FL)[5] was conceived by Lotfi Zadeh, a professor at the University of California at Berkley, and presented not as a control methodology, but as a way of processing data by allowing partial set membership rather than crisp set membership or non-membership. This approach to set theory was not applied to control systems until the 70's due to insufficient small-computer capability prior to that time. Professor Zadeh reasoned that people do not require precise, numerical information input, and yet they are capable of highly adaptive control. If feedback controllers could be programmed to accept noisy, imprecise input, they would be much more effective and perhaps easier to implement. Unfortunately, U.S. manufacturers have not been so quick to embrace this technology while the Europeans and Japanese have been aggressively building real products around it.

1.) Web Mining can improve the search results by exploiting the new semantic structures in the Web, such as extracting and utilizing semantics.

2.) Web Mining technique, can be able increasingly treat the web content, web structure, and web usage.

3.) The Fuzzy Logic is able to solving queries, because it not necessarily matching the query exactly.

4.) The average of the time for the different tests by using fuzzy logic it takes less time.

5.) The increase percentage on databases during the experiment and this result is considered very good.

V Proposed Method

1. Initially enter a keyword to search.

2. Apply the Weighted Page Content Rank Algorithm and calculate the rank.

WPCR is a numerical value to represent the rank of a web page.
Input: Page P, Inlink and Outlink Weights of all backlinks 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?
Z= Sum of frequencies of all N strings.
c) S= Set of the maximum possible strings occurring in P.
d) X= Sum of frequencies of strings in S.
e) Content Weight (CW) = X/Z
f) C= No. of query terms in P
g) D= No. of all query terms of Q while ignoring stop words.
h) Probability Weight (PW) = C/D

Step 2: Rank calculation:

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

b)PR(P)=
$$(1-d)+d[\sum_{V \in B} PR(V)\ W^{in}(P,V)W^{out}(P,V)\ ](CW+PW)$$
c) Output PR(P) i.e. the Rank score

3. If the keyword has already calculated the rank by using Weighted Page Rank in such case it simple use fuzzy algo formula to calculate cluster value.

Cluster value is

$$k \approx \sqrt{n/2} \quad \ldots\ldots (7)$$

where n is the number of data points.

4. Based on the cluster value then that page will look after in that cluster and display on the screen.

5. Apply FCM algorithm to cluster the scattered data.

FCM is based on minimization of fuzzy c means functional formulated as

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \quad \ldots (8)$$

a) Initialize U=[uij] matrix, *U(0)*

b) At k-step, calculate the centers vectors C(k) = [cj] with U(k)

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad .. (9)$$

c) Update U (k), U (k+1)

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|^2}{\|x_i - c_k\|^2}\right)^{2/m-1}} \quad \text{.....(10)}$$

d) If || U (k+1) - U (k) ||<ε then STOP, otherwise return to step 2.

6. If the keyword search by user is a new keyword and the corresponding link is existing in database and then Weighted Page Content Rank will compute to calculate the rank of the page and simultaneously apply the fuzzy c means formula to calculate its cluster value.

VI Experimental Results

S.No.	No. of Links	WPCR algo (Time Taken in ms)	Improved WPCR algo (Time Taken in ms)
1.	8	100472	250
2.	13	112235	297
3.	18	128069	281

Table 2: The time taken to process the web pages.

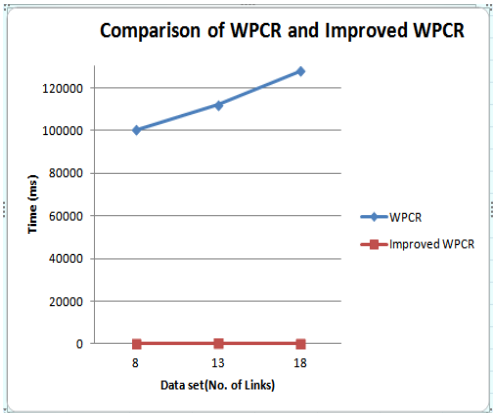

Fig.4. Performance of WPCR and improved WPCR in terms of time and dataset.

The figure shows that when Weighted Page Content Rank Algorithm is implemented, the processing time is too much. As the number of links increased it also increases a time. To improve the performance of WPCR Algorithm Fuzzy Algorithm can be used. In this the time taken to process the Web Pages is so small. As the number of web links is increased, the time is not increased. So, this helps to improve the processing time of Web links.

VII Conclusion and Future Work

During this research numerous research areas are covered, leading to the formulation of a number of conclusions. The rapid growth of the Web and its increasingly wide accessibility has created the need for better and more accurate search capability. Search engines are required to produce the web pages that the users are searching for within the first pages of results. In this setting, the role of ranking becomes critical. In this research, Page Rank and Weighted Page Rank algorithms are used by many search engines but the users may not get the required relevant documents easily on the top few pages. To resolve this problem, a new algorithm Weighted Page Content Rank Algorithm has been studied. This algorithm is aimed at improving the order of the pages in the result list so that the user may get the relevant and important pages easily in the list. But when user search a keyword the WPCR algorithm takes too long time to process the Web pages. To solve this problem, WPCR algorithm is integrated with clustering algorithm. With the use of Fuzzy logic the processing time is very small. So, based on Fuzzy logic approach improves relevancy factor. This technique keeps the related documents becomes more efficient in terms of time complexity. The results can be further enhanced by using other methods of clustering. It can also improve the relevancy factor to retrieval the Web documents which can further improve the result set.

References

1. Chintandeep kaur, Rinkle Rani Aggarwal "*Web Mining Tasks and Types: A Survey*" Volume 2, Issue 2, Feb. 2012.
2. Neelam Tyagi, Simple Sharma, *"Weighted Page rank algorithm based on number of visits of Links of web page"* International Journal of Soft Computing and Engineering Volume 2, Issue-3, July 2012.
3. Laxmi Choudhary and Bhawani Shankar Burdak, *"Role of Ranking Algorithms for Information Retrieval"*, International Journal of Artificial Intelligence & Applications (IJAIA), Volume 3, No.4, July 2012.
4. T.Munibalaji, C. Balamurugan "*Analysis of Link algorithms for Web Mining*" International Journal of engineering and Innovative Technology, Volume 1, Issue 2, Feb 2012.
5. R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, *"Low complexity fuzzy relational clustering algorithms for web mining"*, IEEE Trans. Fuzzy Systems 9 .595–607, Aug. 2011.
6. Tamana Bhatia, "*Link Analysis Algorithms for Web Mining*" IJCST Volume 2, Issue 2, June 2011.
7. Rekha Jain, Dr. G. N. Purohit *"Page Ranking Algorithms for Web Mining"* International Journal of Computer Applications Volume 13, No.5, Jan 2011.
8. J.C. Bezdek, *"Pattern Recognition with Fuzzy Objective Function Algorithms",* Plenum Press, New York, Volume 3, June 1981.
9. Vaibhav C. Naik "*Fuzzy C-Means Clustering Approach*" Department of Industrial and Management Systems Engineering, University of South Florida, Volume 2, 8July, 2004.