WEB STRUCTURE MININGALGORITHMS: A REVIEW

Harshita Mishra*, Rajat Verma**
Department of Computer Science & Engineering
Amity School of Engineering & Technology
Amity University, Lucknow
Harshu1995@yahoo.com,rajatverma310795@gmail.com

***Abstract*—** The Web structure mining is used for obtaining the structural synopsis of webpages and websites. It also deals with the discovery of structure of web document. This helps in navigation purposes and also the comparison or integration of webpage schema. By web structure mining, one can access information in webpages by providing reference schema. In web structure mining, the information gathered is about frequency measurement of local links in the web tuples in a web table. Other objectives of web structure mining are to know the hierarchy or network of hyperlink in the webpage of a domain and helps in generalizing the flow of information of website which belongs to a domain. This makes query process easy and efficient. This paper deals with the extraction of relevant data from webpage. Different algorithm which can be implemented in web structure mining from extraction of data are- Hypertext Induced Topic Search (HITS), PageRank and Weighted PageRank etc.

Keywords: *Hypertext Induced Topic Search, PageRank, Weighted PageRank etc.*

I. INTRODUCTION

Web structure mining is mainly used for linking structure of the webpages and websites. To find similarities between sites or finding web communities, structure mining can be used. Example of webpage structure is-

<Html>
. . .
link

</html>

To study nodes and connection structure of website, Web Structure Mining uses graph theory. Web structure mining is classified as:

1) Configuration extraction with respect to hyperlink.
2) Formation of document is mined

Hyperlinks are analyzed in Web Structure Mining for calculating the website’s rank. Structuring the hyperlink is the main issue within webpage itself. To recover closely related data analysis of link formation in the webpage in helpful for users.

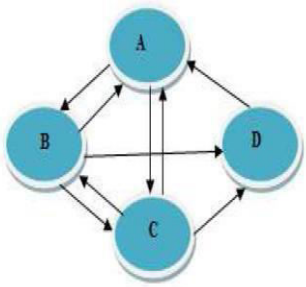

Fig 1: four webpages and their hyperlink structure

1. **Configuration extraction with respect to hyperlink:**

Hyperlink consists of structured components in which webpage that are located at different location.

a. **Formation of document is mined:**Analysis of webpages constructed as tree is explained using Markup Languages such as Hypertext Markup Language and Extensible Markup Language.

Major technique used-
1) PageRank Technique
2) Weighted PageRank Technique(WPR)
3) Hyperlink Induced Topic Search Technique(HITS)

Technique such as-
- Link Editing
- Topological Frequency
- Utility Mining

are used for web structure mining.

II. CLASSIFICATION OF CONNECTION

1. **In-connections or inbound connections:** Connections present in website from the outside are called in link or inbound connection.

2. **Outbound connection:** Connection that are from one webpage toother webpage in a website or another website are called outbound connection.
3. **Suspended connections:** Links that point to any webpage with noother links are called suspended connection.

III. EXPLANITION OF ALGORITHM

1- Ranking of Pages:

Ranking of pages is calculated at indexing time instead of query time. Voting process is done for the ranking of pages and rank is called vote. In web structure mining two parameters are used-

1. Forward link
2. Backward link.

2- Weighted Page Rank:

Values with greatest rank are analyzed by using remarkable pages in making deciding returns of high rank. In-connections and out-connections of webpages are used for calculation of remarkable pages. Weighted PageRank technique is preferred over PageRank technique.

3- Hyperlink Induced Topic Search:

Connection analysis is done in this scenario. Webpage analysis is calculated by exclusion of in-connection and out-connection in page. Iterative calculation is done by two different ways. This algorithm has two values-

1- Authority value
2- Hub value

Indication of hyperlink is done by hubs and hyperlink is used to indicate authority of pages.

4- Connection Editing:

Important pages are those which are highin-degree. Each page is graded offline.

5- Topological Frequency Utility Mining: Thecomputation of each page is based on topology parameters, frequency and utility.

IV. PAGE RANK

Sergey Brin and Larry Page proposed PageRank. Webpages are represented as nodes in the graph and connections or hyperlink is represented as arcs. Points inside nodes are In-connection and point out from nodes is out-connection.

PageRank algorithm is given as:

PR(A)=(1-d)+d(PR(T1)/C(T1)

+ ....PR(Tn)/C(Tn))

PR(A)=Page rank of page A

PR(Ti)=Page rank of pages Ti which links to page A.

C(Ti)=Number of out-bound links on page Ti

d=Damping factor range between 0 and 1.

Formula can be simply represented as, (d=0.85) Many iteration are applied to get accurate value. Equation of PageRank is given as;

$$\Pi^T = \pi^T(\alpha S+(1-\alpha)E)$$

1- Summation Formula:

$$r(P_i) = \sum_{P_j \in Bp_i} r(P_j)/|P_j|$$

Bp_i: Set of pages pointing to P_i

$|P_j|$: Out-connection of the page is given as P_j

r (P_j): Unknown inputs in the beginning of the calculation

1/n is the given equal page rank n-number of pages in google's index.

$$R_{k+1}(P_i)=\sum_{p_j \in Bp_i} [r_k(P_j)]/|P_j|$$

2- PageRank Calculation:

PageRank calculation uses iterative algorithm and uses Eigen vector principle in connection matrix of webpages for normalization.

$$P(i) = \lim_{n\to\infty} [N(i, n)]/n$$

3- Matrix model:

The construction of connection in the matrix model can be given as-

Example-

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Then iterative value of PageRank r can be given as-

$$\mathbf{r}_{(k+1)}^T = \mathbf{r}_{(k)}^T \mathbf{Q} \qquad , k = 0, 1, \ldots$$

4- Random walker:

The process in which random page is selected is called random walker. From random page out-connection is selected randomly. The pages are now known as asymptotic probabilities that are surfer.

This is given by-

$$\hat{\mathbf{Q}} = \mathbf{Q} + \frac{1}{n}\mathbf{d}\mathbf{e}^T$$

5- Sub graph: stuck scenario:

In Stuck in a sub graph we can move from any page which has small probability to link structure of any other page.

$$\hat{\hat{\mathbf{Q}}} = \alpha\hat{\mathbf{Q}} + (1-\alpha)\frac{1}{n}\mathbf{e}\mathbf{e}^T$$

6- Practical calculations of Page Rank:

In Practical calculations of PageRank, one is the largest Eigen value in irreducible column-stochastic matrix and a non-negative elements are present at right eigenvector.

Formula is given as-

$$\hat{\mathbf{Q}}^T \mathbf{r} = \mathbf{r}$$

The link structure can be freely illustrated by the link matrix q which is sparse in nature (sparse matrix means most of the element are 0)

7- Damping factor:

It is represented by d that is considered as a click-through probability, and is added to block the pages with the links that are going outwards from "absorbing" the ranking of pages of those pages connected to the sinks. Chances of factor concerningd=0, then damping factors are renewed, which are uniformly scatteredby definition. The damping factor has a range greater than 0 but less than 1.It can be considered as an average scenario between the two extreme values.

The input scenarioof ranking of pages section is inbound connections and it is used by the google. The main aim of this procedure is retrieving the processed form of data as well ascomparisons.

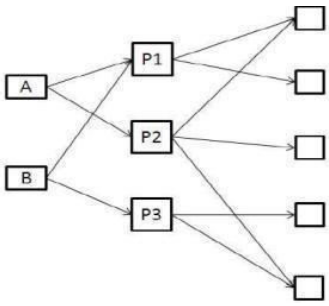

FIG 2: This figure illustrates how the linking procedure is done in a collection of webpages, primarily termed as a website.

It performs in a range, beginning from 0 and ending till 10.
It can be depicted as a log scale.
The tabular representation illustrates the values of ranking of web pages.

Toolbar PageRank (log base 10)	Real PageRank
0	0 – 100
1	100 - 1,000
2	1,000 - 10,000
3	10,000 - 100,000
4	and so on...

V. DIFFICULTY SCENARIO IN RANKING OF PAGES

a. Rank Sink:

In terms of ranking of pages, this difficulty is having a minor consideration. If we assume thata few pages on the web that correspond to a pages in the circle. During iteration process, circleassemble ranking of page inputs instead of ranking of page protocols are scattered. It is defined as the circle or the loop that illustrates as a trick type. The values concerning ranking of pages are more than the existence in the tenure ofthis problem.

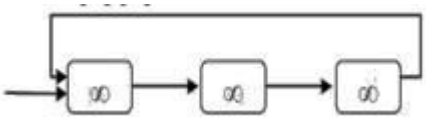

Fig 3: This figure illustrates the continuity sequence in terms of page sink.

b. Dangling page:

This problem considered as a more complex and having a major approach. When a web graph inside a page is not having a forward link it is considered as dangling. Throughout the repetition processes, the dangling page of $Rank_i$ willmislay its norm uninterruptedly.

How Ranking is calculated:
Procedure-

- In the initial step we need to discover all the back links of Q (set G).
- Next Step comprises of finding: PR (Q) = (1-factor of damping) +factor of damping.
- Final Step comprises of scoring of ranks.

VI. ADVANTAGES & DISADVANTAGES OF PAGE-RANK:

ADVANTAGES:

- Ability in coping up with pages that has the spam characteristic.
- Performs the estimation of the rank score (global).
- Performs the not dependent algorithm in concern to the query scenario.
- Effective computation (rapid).

DISADVANTAGES:

- Favors the older pages.
- Page Rank can be increased by a procedure known as “link farm”.
- Buying a link is also possible that increase the page rank.

VII. CONCLUSION:

Web Mining comes from the fifth section of knowledge discovery in database that concludes the concept of data mining. Data Mining is used for the extraction of useful and meaningful data from the huge amount of data basically from the central repository considered as the "Warehouse of Unprocessed raw facts and figures".

The Structural abstract of the web page in concern to obtaining procedures referring to sub category of obtaining useful data from the repository along with other factors such as the extraction of relevant unprocessed facts in terms of content phase as well as the usage scenario. The approach ofthe comprehensive review is to analyze the page rank algorithm. In future, analysis of the efficiency of this algorithm (ranking of pages)as well as comparison among the different procedures concerning to the structural abstract of extracting processes of relevant data or techniques that are based on performance will be conducted.

REFERENCES

[1]"Investigating Google's PageRank algorithm", Erik Andersson, Per-Anders Ekström, Report in Scientific Computing, advanced course - spring 2004.

[2]"Notes on PageRank Algorithm", ENGG2012B Advanced Engineering Mathematics, Kenneth Shum.

[3]"Web mining–concepts, applications and research Directions". Srivastava, T., Desikan, P., & Kumar, V. In *Foundations and advances in data mining* (pp. 275-307). Springer, Berlin, Heidelberg.

[4]"Web Mining Research: A Survey", Raymond Kosala, Hendrik Blockeel, ACM July 2000, Vol 2, Issue 1.

[5]"Web structure mining using page rank, improved page rank an overview", V. Lakshmi Praba and T. Vasantha ICTACT Journal on Communication Technology, March 2011, Vol 2, Issue 1

[6] "Analys.is of Link Algorithms for Web Mining", International Journal of Engineering and Innovative Technology (IJEIT). Volume 1, Issue 2, February 2012.

[7] "Optimizing web servers using page rank prefetching forclustered accesses." Safronov, V., & Parashar, M. *InformationSciences*, *150*(3-4), 165-176, 2005