www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 9 September 2014 Page No. 8000-8003

STUDY ON WEB STRUCTURE MINING AND ITS PAGE RANKING ALGORITHM

T. Shanmugapriya[1], K.Kalaiselvi[2]

[1] Second M.E(CSE), Department of Computer Science and Engineering,
[2]Assistant Professor, Department of Computer Science and Engineering,
SNS College of Engineering, Sathy main road, Coimbatore-641035, Tamil Nadu, India
E-mail: priyasthiyagu@gmail.com
info.kalaiselvi@gmail.com

ABSTRACT:

Web Structure Mining deals with the hyperlink structure of the document in the web. The various Web Structure Mining algorithm are page rank, weighted page rank, hyper induced topic search(HITS),Link Editing, Topological Utility Frequency Mining. The study focuses on the page rank algorithm. The following section describes the page rank algorithm structure, computation, problems, pros and cons.

I.INTRODUCTION:

Web structure mining focuses on the link structure of the web. This can help in discovering similarities between sites or discovering web communities. Structure of web page is in the form:

<html>

. . .

link

</html>

Web Structure Mining is the process of using graph theory to analyze the node and connection structure of website. There are two kinds of web structure mining: 1)Extracting pattern from hyperlinks in the web. 2)Mining the document structure. Web Structure Mining analyzes the hyperlink for calculating the rank of the website. The challenge is to deal with structure of hyperlink within the web itself. It helps the user to retrieve the relevant document by analyzing the link structure of web.

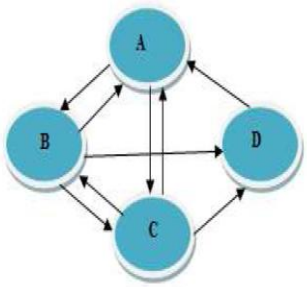

fig 1: Hyperlink structure of four pages

1.Extraxcting pattern from hyperlink in the web:

Hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: Analysis of tree like structure of page structure to describe HTML or XML tag usage.

The three main important algorithms in web structure mining are page rank algorithm(PR), Weighted page rank algorithm(WPR), Hyper Induced Topic Search(HITS). The other algorithm

of web structure mining are Link Editing, Topological Frequency Utility Mining.

II.VARIOUS KIND OF LINKS:

1.Inbound links or in-links: It refers that the link that are into the site from the outside.

2.Outbound link: It refers that the link from a page to other page in a site or other site.

3.Dangling links: These links points to any page with no outgoing links.

III.ALGORITHM:

1.Page Rank: The page rank for each page is computed during indexing but not during query time. Page Rank is a "vote", by all the other pages on the web, about how important the page is. A link to a page counts as vote of support. There are two parameters are used to mine the structure of web. They are 1.Forward link and 2.Backward link.

2.Weighted Page Rank: The larger rank values are decided based on the significant of web page. The significant of web page is calculated based on the number of in-links and out-links of pages. This algorithm is more efficient than page rank algorithm.

3.Hyperlink Induced Topic Search: It is a link analysis algorithm. Analysis of web page is calculated by dispensation in-links and out-links of the web page. Two different way of iterative calculation is performed. They are value of authority and value of hubs. The hub is the web page pointed to many hyperlinks. The authority of web page is pointed by many hyperlinks.

4.Link Editing: The grade for each page is computed offline. The pages with high in-degree and more time spent are important pages.

5.Topological Frequency Utility Mining: Based on frequency, utility along with topology parameters each page is computed.

IV.PAGE RANK: Page Rank was proposed by Sergey Brin and Larry Page. Nodes in the graph are webpages and arcs represents the link or hyperlinks. In-links are point into a node and out-link are point out from nodes.

The algorithm of page rank is given by

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.

A simple way to representing formula is,(d=0.85)

Accurate value are obtained only through many iterations.

The page rank equations are as follows:

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

1.Summation Formula:

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

Where Bp_i-The set of pages pointing to P_i

$|P_j|$-The number of out-links from page P_j

$r(P_j)$-Value is unknown in the beginning of the calculation

1/n is the given equal page rank

n-number of pages in google's index.

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

A simple iterative algorithm is used to calculate the page rank and corresponds to the principle eigen vector of the normalized link matrix of the web.

2.Calculating the page rank

Start with random web page, say i. Suppose this page has out going links to pages $j_1,j_2,j_3 \ldots j_{in}$. A simple random walk would choose each of those links with equal probability.

$$P_{ij}=\{(1/i_n \text{ if } j \in \{j_1,j_2,\ldots j_{in}\}\}$$

i- number of times it is traversed in a very long random walk:

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

3.Matrix model :

In the matrix-formulation, this link structure will be written as:

eg:

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

The iteratively calculated PageRank r could then be written as:

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

4.Random walker: This random walker (or surfer) starts from a random page, and then selects one of the out-links from the page in a random fashion. The Page Rank (importance) of a specific page can now be viewed as the asymptotic probability that the surfer is present at the page.

In matrix formulation, this can be written as:

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

5.Stuck in a subgraph: The ability to jump, with a small probability, from any page in the link structure to 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: An irreducible column-stochastic matrix has 1 as the largest eigen value and its corresponding right eigenvector has only non-negative elements. The final formula becomes:

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

Sparse link matrix Q, which was initially created to describe the link structure together with two more sparse matrices, as in equation:

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

7.Damping factor: The damping factor d, which is the click-through probability, is included to prevent sinks (i.e. pages with no outgoing links) from "absorbing" the Page Ranks of those pages connected to the sinks. f the click-through probability is d=0, then all clicks are random restarts, which are uniformly distributed (the 1/N coefficient in the first term) by definition. So, a damping factor 0<d<1 is a sort of weighted average between the two extremes.

The input parameter of page rank is inbound link used by the search engine google. The purpose of algorithm is used for information retrieval and compare those algorithm.

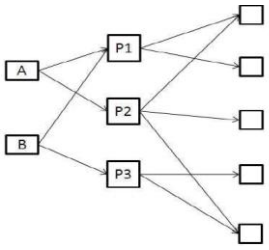

FIG 2:Link of a website

The page rank values from 0-10 and seems to be like a logarithmic scale. The below table shows the rank value 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.PROBLEMS IN PAGE RANK:

1.Rank Sink: It is a minor problem of page rank algorithm. Consider some web pages that points to one of the pages in the loop. While performing the iteration, the loop accumulates page rank values but never distribute any page rank values. Rank Sink is defined as the loop that forms a sort of trap. Page rank values are higher than the existence during the Rank Sink problem.

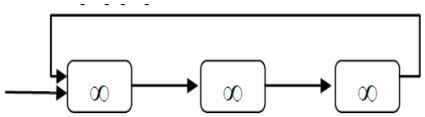

Fig 3:Loop of Rank Sink

2.Dangling page: It is the major problem of the page rank algorithm. The page contains a web graph and if it has no forward link then it is said to be dangling page. During the iteration the dangling page of $Rank_i$ is loses its norm continuously.

Rank calculation:
a) Find all back links of P (say set B).
b) PR (P) = (1-d) +d

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

VI.PROS AND CONS:

Pros:

1.Capable to fight with spam web pages.

2.Global rank score computation.

3.Query independent algorithm.

Efficient and fast computation.

Cons:

1.Recency Search.

2.Topic drift.

3.Link spamming (Page Cheating).

VIII.CONCLUSION:

Web mining comes from the data mining technique which is used to retrieve the knowledge from web. Web Structure Mining is the category and also an issue in a web mining. This study analyses the page rank algorithm. My future work is to analyse the performance of page rank algorithm and to analyse the comparison between various web structure mining based on performance.

ACKNOWLEDGEMENT:

I would like to thank my guide Ms.K.Kalaiselvi, Assistant Professor in Department of Computer Science and Coordinator Mr.D.Jabakumar Immanual M.E,Department of CSE.

REFERENCE:

[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", Jaideep Srivastava, Prasanna Desikan, Vipin Kumar.

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

[5]" web structure mining using page rank, improved page rank an overview", ictact journal on communication technology, march 2011, vol: 02, issue: 01

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

[7]"The Page Rank Citation Ranking: Bringing Order to the Web" Jan 29,1998

[8]" Comparative Study of Web Page Ranking Algorithms" International Journal of Emerging Technologies in computational and Applied Sciences.