WEB STRUCTURE MINING USING PAGERANK, IMPROVED PAGERANK – AN OVERVIEW

V. Lakshmi Praba¹ and T. Vasantha²
¹Department of Computer Science, Government Arts College for Women, Tamil Nadu, India
E-mail: vlakshmipraba@rediffmail.com
²Manonmaniam Sundaranar University, Tamil Nadu, India
E-mail: vasanthasankarganesh@gmail.com

Abstract
Web Mining is the extraction of interesting and potentially useful patterns and information from a Web. It includes Web documents, hyperlinks between documents, and usage logs of web sites. The significant task for web mining can be listed out as Information Retrieval, Information Selection / Extraction, Generalization and Analysis. Web information retrieval tools consider only the text on pages and ignore information in the links. The goal of Web structure mining is to explore structural summary about web. Web structure mining focusing on link information is an important aspect of web data. This paper presents an overview of the PageRank, Improved Page Rank and its working functionality in web structure mining.

Keywords:
Web Structure Mining, Improved PageRank

1. INTRODUCTION

Web mining is the application of data mining and knowledge discovery techniques to extract information from Web. Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) are the three knowledge discovery domains in Web mining [3]. Web content mining is the process of extracting knowledge from the content of documents - including text, image, audio, video, etc. Research in web content mining includes resource discovery from the web, clustering such as document based clustering and Link based clustering, and information extraction from web pages. Web structure mining is the process of inferring knowledge from the Web and links between references and referents in the Web. Web usage mining involves the automatic discovery of user access patterns from one or more Web servers [3].

2. WEB STRUCTURE MINING

Web structure mining deals with the structure of the hyperlinks within the web itself. In this paper the structure of web as a directed graph with web pages and hyperlinks, where the web pages correspond to nodes and hyperlinks correspond to edges connecting the related pages has been focused. An overview of the two kinds of web structure namely, Link Structure, Document Structure has been considered.

2.1 LINK STRUCTURE

Link Structure is based on hyperlink structure. Hyperlink Structure is a structural unit that connects a place in a Web page to a different place, either within the same Web page or on a different Webpage. A hyperlink that connects to a different place within the same web page is called as Intra-Page Hyperlink, and hyperlink that connects a particular place in another web page is called as Inter-Page Hyperlink [3].

2.2 DOCUMENT STRUCTURE

Document Structure represents the designer’s view of the content organization within a web page. The content represented in the arrangement of HTML and XML tags within a page is called as Intra- page structure. Both HTML and XML documents can be represented in tree structure format.

3. WEB STRUCTURE ALGORITHMS

Web structure mining discovers the structure of web document as well as link structure of the hyperlinks. The hyperlink structure of the web is valuable for locating information. Based on the topology of the hyperlinks, it will categorize the web pages and generate the information, such as the similarity and relationship between different web sites. WSM has two influential hyperlink based search algorithms – PageRank Algorithm and Improved PageRank Algorithm.

3.1 PAGERANK ALGORITHM

PageRank was presented by Sergey Brin and Larry Page. The search engine Google was built based on this algorithm. PageRank has emerged as the dominant link analysis model for Web search, partly due to its query-independent evaluation of Web pages and its ability to combat spamming [4].

PageRank suggest that a link from page u to v implies the author of u recommends taking a look at page v. A page has a high PageRank value if there are many pages that point to it. PageRank is a static ranking of Web pages in the sense that a PageRank value is computed for each page off-line and it does not depend on search queries. The PageRank Algorithm represents the structure of the Web as a matrix, and PageRank values as a vector. The PageRank vector is derived by computing matrix-vector multiplications repeatedly. The sum of all PageRank values should be one during the computation [4].

3.2 PAGERANK COMPUTATION

The PageRank Algorithm is the most popular method to compute the PageRank value of web sites. The PageRank computation based on the concepts of back links. Considering d as a web page, the set of web pages that points to d are called as back links of d, denoted by Bd. Set of pages pointed by d are called as forward links, denoted by Fd. The formula to calculate the pagerank is as follows:
\[
PR(m) = \sum_{n \in \text{out}} PR(n)/N(n)
\]  

(1)

The web can be viewed as a directed graph whose nodes are the pages or documents and the edges are the hyperlinks between them. This directed graph structure in the web is called a web graph. A graph can be represented as \(G(V, E)\) consisting of a set of vertices and a set of edges denoted by \(V(G)\) and \(E(G)\) respectively. Let us assume the Hyperlink Structure for a web graph of eight pages (Fig.1), and the corresponding matrix for the designed Hyperlink Structure is represented in Fig.2. In this Hyperlink Structure pages are denoted by rectangular shape. The links are denoted by directed line from one page to another. Every page in this Hyperlink Structure has at least one outgoing edge.

![Fig.1. Hyperlink Structure for a web graph of eight pages](image)

![Fig.2. Matrix for Hyperlink Structure](image)

Entries in the square matrix are determined by link between the pages. If there is no link between the pages then the value of the entry is zero otherwise its value is \(1/P\) where \(P\) is number of outgoing edges[1],[4].

Each row of the square matrix indicate back link of every node and each column of the square matrix indicate forward link of every node. Using the values of matrix the rank can be calculated for all pages in the web graph. Back link of page 1 is represented by the row of the matrix and forward link of page 1 is represented by the column of the matrix. PageRank value of each page is obtained from summing the rank of Non zero values of its corresponding row in matrix \(M\).

An initial rank matrix \(Rank_0\) is defined as a \(N \times 1\) where \(N\) is the total number pages in the web graph All entries in \(Rank_0\) are set to 1/N. \(Rank_0 = (1/N, 1/N, 1/N, 1/N, 1/N, 1/N, 1/N, 1/N)^T\). To determine the PageRank value, a number of iterations are carried out until it converges to a fixed point. The converged \(Rank\) is the PageRank values of all pages[5],[6].

Compute new matrix \(M'\) by applying damping factor \(d\) with square matrix \(M\), where \(d\) is the probability that user will link to a random page and \((1-d)\) is the probability that user will follow the links consecutively, to trace the web. \(M'\) is expressed as Eq. (2).

\[
M' = (1 - d)M + dI/\sum N
\]

(2)

where \(N\) is the total number of pages and \([1/N] \times N\times N\) be an \(N\times N\) square matrix in which all entries are \((1/N)\). The constant \(d\) is less than one. According to the literature review, many researches (Sung Jin Kim and Sang Ho Lee, Hung-Yu Kao and Seng-Feng Lin) have suggested normally \(d\) is set to 0.15. Hyperlink Structure for Fig.1 illustrates no Dangling Pages and no RankSink problem. \(M'\) constructed for the Fig.1 using Eq. (2) is given below.

\[
M' = \begin{bmatrix}
0.0188 & 0.4437 & 0.4437 & 0.0188 & 0.4437 & 0.0188 & 0.0188 & 0.0188 \\
0.2313 & 0.0188 & 0.4437 & 0.1887 & 0.0188 & 0.0188 & 0.0188 & 0.0188 \\
0.2313 & 0.0188 & 0.0188 & 0.1887 & 0.0188 & 0.0188 & 0.0188 & 0.0188 \\
0.2313 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.1887 & 0.0188 \\
0.0188 & 0.4437 & 0.0188 & 0.4437 & 0.1887 & 0.0188 & 0.0188 & 0.0188 \\
0.2313 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.1887 & 0.0188 \\
0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.4437 & 0.0188 \\
0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188 & 0.0188
\end{bmatrix}
\]

Fig.3. Calculated Matrix Values of \(M'\) and initial rank value
The Eq. (3) is used to compute PageRank values of all pages as given below:

\[ \text{PageRank Calculation:} \]

\[ \text{Rank}_{i+1} = M'X \]

The PageRank calculation equations for all the eight pages are given below.

**Page 1 - PageRank Calculation:**

\[ \text{PR}[1] = ((1-d)(0)+d_1/2)^{\frac{1}{N}} + ((1-d)(0)+d_2/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_7/2)^{\frac{1}{N}} + ((1-d)(0)+d_8/2)^{\frac{1}{N}} + ((1-d)(0)+d_9/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{27}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{81}/2)^{\frac{1}{N}} \]

\[ + ((1-d)(0)+d_{82}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{146}/2)^{\frac{1}{N}} + ((1-d)(0)+d_{147}/2)^{\frac{1}{N}} = 0.0023 + 0.0023 + 0.0023 + 0.0023 = 0.0093 \]

**Page 2 - PageRank Calculation:**

\[ \text{PR}[2] = ((1-d)(0)+d_1/2)^{\frac{1}{N}} + ((1-d)(0)+d_2/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_7/2)^{\frac{1}{N}} + ((1-d)(0)+d_8/2)^{\frac{1}{N}} + ((1-d)(0)+d_9/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{27}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{81}/2)^{\frac{1}{N}} \]

\[ + ((1-d)(0)+d_{82}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{146}/2)^{\frac{1}{N}} + ((1-d)(0)+d_{147}/2)^{\frac{1}{N}} = 0.0289 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 = 0.1197 \]

**Page 3 - PageRank Calculation:**

\[ \text{PR}[3] = ((1-d)(0)+d_1/2)^{\frac{1}{N}} + ((1-d)(0)+d_2/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_7/2)^{\frac{1}{N}} + ((1-d)(0)+d_8/2)^{\frac{1}{N}} + ((1-d)(0)+d_9/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{27}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{81}/2)^{\frac{1}{N}} \]

\[ + ((1-d)(0)+d_{82}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{146}/2)^{\frac{1}{N}} + ((1-d)(0)+d_{147}/2)^{\frac{1}{N}} = 0.0289 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 = 0.0666 \]

**Page 4 - PageRank Calculation:**

\[ \text{PR}[4] = ((1-d)(0)+d_1/2)^{\frac{1}{N}} + ((1-d)(0)+d_2/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_7/2)^{\frac{1}{N}} + ((1-d)(0)+d_8/2)^{\frac{1}{N}} + ((1-d)(0)+d_9/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{27}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{81}/2)^{\frac{1}{N}} \]

\[ + ((1-d)(0)+d_{82}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{146}/2)^{\frac{1}{N}} + ((1-d)(0)+d_{147}/2)^{\frac{1}{N}} = 0.0289 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 = 0.0984 \]

**Page 5 - PageRank Calculation:**

\[ \text{PR}[5] = ((1-d)(0)+d_1/2)^{\frac{1}{N}} + ((1-d)(0)+d_2/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_7/2)^{\frac{1}{N}} + ((1-d)(0)+d_8/2)^{\frac{1}{N}} + ((1-d)(0)+d_9/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{27}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{81}/2)^{\frac{1}{N}} \]

\[ + ((1-d)(0)+d_{82}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{146}/2)^{\frac{1}{N}} + ((1-d)(0)+d_{147}/2)^{\frac{1}{N}} = 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 = 0.1462 \]

**Page 6 - PageRank Calculation:**

\[ \text{PR}[6] = ((1-d)(0)+d_1/2)^{\frac{1}{N}} + ((1-d)(0)+d_2/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_7/2)^{\frac{1}{N}} + ((1-d)(0)+d_8/2)^{\frac{1}{N}} + ((1-d)(0)+d_9/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{27}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{81}/2)^{\frac{1}{N}} \]

\[ + ((1-d)(0)+d_{82}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{146}/2)^{\frac{1}{N}} + ((1-d)(0)+d_{147}/2)^{\frac{1}{N}} = 0.0289 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.0023 = 0.1728 \]

**Page 7 - PageRank Calculation:**

\[ \text{PR}[7] = ((1-d)(0)+d_1/2)^{\frac{1}{N}} + ((1-d)(0)+d_2/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_7/2)^{\frac{1}{N}} + ((1-d)(0)+d_8/2)^{\frac{1}{N}} + ((1-d)(0)+d_9/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{27}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{81}/2)^{\frac{1}{N}} \]

\[ + ((1-d)(0)+d_{82}/2)^{\frac{1}{N}} + \ldots + ((1-d)(0)+d_{146}/2)^{\frac{1}{N}} + ((1-d)(0)+d_{147}/2)^{\frac{1}{N}} = 0.0023 + 0.0023 + 0.0023 + 0.0023 + 0.1086 = 0.1463 \]
Example of loop which consists of three pages in the loop are likely to have higher PageRank values than web pages and it act as RankSink [5]. Due to RankSink problem, the loop will never distribute any PageRank values. The loop accumulates PageRank but do not point to other web pages (that is a loop). Assume there are some web pages that point to one of pages in the loop.

By repeatedly doing the PageRank calculations using the Eqs. (4-11), the PageRank values are computed with no Dangling page and no RankSink[1][12] are given in Table 1.

3.2.1 RankSink:

RankSink is one of the problems of Page Rank Algorithm. Consider number of web pages that point internally to each other but do not point to other web pages (that is a loop). Assume there are some web pages that point to one of pages in the loop. Then, during the iteration, the loop accumulates PageRank values but never distributes any PageRank values. The loop forms a sort of trap, which is called as RankSink. Fig. 4 demonstrates a simple example of loop which consists of three web pages and it act as RankSink [5]. Due to RankSink problem, pages in the loop are likely to have higher PageRank values than the existence [10], [11].

For the illustration shown in Fig. 1, suppose the link from pages 1, 2, 4, 5 to pages 4, 5, 3, 1 respectively and a link from 1 to 6, which is an only link from group 2 to group 1. Existing PageRank Algorithm overcome the RankSink problem [1].
3.2.2 Dangling Page:

Major problem of original PageRank Algorithm is Dangling page. If a page in web graph has no forward link then the page is called as Dangling page. In Fig.6 the page 5 is a Dangling page as it has no outgoing edge, and the 5th column of corresponding matrix M is called as dangling column. It is represented as a zero vector (Fig.7). In general a matrix M is irreducible if and only if a directed graph is strongly connected. All columns in an irreducible matrix M have norms of value one. Here norm of 5th column of matrix M is zero. As a consequence of Dangling page, Rank, loses a part of its norm continuously during the iteration process. This event is called as norm- leak[1],[10].

3.3 IMPROVED PAGERANK ALGORITHM

Fig.6. demonstrates the Hyperlink Structure with Dangling page and its matrix is given Fig.7.

![Hyperlink Structure with Dangling page](image)

Using the Eq. (13)

$$M^* = (1 - d) M^* + d\left[1/\text{N}\right]_{\text{NXN}}$$

(13)

By using the Eq. (13), M* is constructed for Fig.6 is follows

Using Improved PageRank Algorithm, calculated PageRank values for each page is illustrate in Table.2. It shows that the result might converge to a fixed point. In an Improved PageRank Algorithm, the sum of all PageRank values are always maintained to be one [8],[9].

| Iteration | Page1 | Page2 | Page3 | Page4 | Page5 | Page6 | Page7 | Page8 | TotPR | PRLoss |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| 1         | 0.1250 | 0.1250 | 0.1250 | 0.1250 | 0.1250 | 0.1250 | 0.1250 | 0.1250 | 1.0000 | 0.0000 |
| 2         | 0.1383 | 0.1330 | 0.0798 | 0.1117 | 0.1595 | 0.1861 | 0.1595 | 0.0320 | 1.0000 | 0.0000 |
| 3         | 0.1261 | 0.1180 | 0.0841 | 0.1442 | 0.1903 | 0.2197 | 0.0819 | 0.0357 | 1.0000 | 0.0000 |

274
4. EVALUATION

PageRank Algorithm and the Improved PageRank Algorithm are implemented in C language and executed in Intel(R) Pentium (R) Dual CPU (200 Ghz) machine with 2GB of RAM. The number of nodes and the number of links are entered as input for both algorithms and the PageRank values are obtained as output. The elapsed times for PageRank Algorithm and Improved PageRank Algorithm to compute PageRank values are 0.44 ms and 0.49 ms respectively. For the demonstrated web graph, Fig.6 which has one Dangling Page and no RankSink, the obtained results are illustrates in Fig.9. The algorithm is executed for 25 iterations and it is observed that the Rank, and Rank_{i+1}, never converged for a web graph with Dangling pages.

![Fig.9. Result of PageRank Algorithm](image)

Fig.9 shows the result of the Improved PageRank Algorithm. The Improved PageRank Algorithm is executed for same web graph. It is observed that both Rank, and Rank_{i+1}, converged at the earlier stage of the iteration itself, which then remain unchanged.

![Fig.10. Result of Improved PageRank Algorithm](image)

Fig.10 shows the result of the Improved PageRank Algorithm. The Improved PageRank Algorithm is executed for same web graph. It is observed that both Rank, and Rank_{i+1},

5. CONCLUSION

The main purpose of this study is to explore the hyperlink structure and to understand the web graph. This study focused on Improved PageRank Algorithm which solves the norm-leak phenomenon. In the Improved PageRank Algorithm, pages have the exact PageRank values. This study included a number of pre-processing operations too. The evaluation results clearly indicated that the rank of Improved PageRank Algorithm always converged to a fixed point which is not the case with PageRank Algorithm. This Improved PageRank Algorithm finds applications in search engines.

Further research can be focused by analyzing the links between web sites. Algorithms like HITS (Hyperlink Induced
Topic Search) and Weighted PageRank can also be evaluated in similar line.

REFERENCES

[1] F.Crestani, M.Girolami and C.J.Van Rijsbergen (Eds.): ECIR 2002, LNCS 2291, pp.73-85, 2002 ©Springer-Verlag Berlin Heidelberg.

[2] Chakrabarti, S., B. Dom, D. Gibson, J. Kleinberg and R. Kumar et al., 1999, “Mining the link structure of the world wide web”, IEEE Comput., Vol. 32: pp. 60-67.

[3] T. Srivastava , Prasanna Desikan, Vipin Kumar, 2005, “Web Mining – Concept, Applications and Research Directions”, Studies in Fuzziness and Soft Computing, Vol. 180, pp. 275-307.

[4] Bing Liu, 2007, “Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data”, ©Springer-Verlag Berlin Heidelberg.

[5] K. Bharat and M. Henzinger, 1998, “Improved Algorithms for Topic Distillation in Hyperlinked Environments”, Proceedings of the 21st ACM SIGIR Conference, pp.104-111.

[6] S.Brin and L.Page, 1998, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Proceedings of WWW7, pp.107-117.

[7] J.Dean and M. R. Henzinger, 1999, “Finding Related Web Pages in the World Wide Web”, Proceedings of WWW8, pp.1467-1479.

[8] D.Gibson, J.Kleinberg, and P. Raghavan, 1998, “Inferring Web Communities from Link Topology”, Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, pp.225-234.

[9] Y.Ng, A.X.Zheng, and M.I. Jordan, 2001, “Stable Algorithms for Link Analysis”, Proceedings of the 24th ACM SIGIR Conference, pp.258-266.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, 1998, “The PageRank Citation Ranking: Bringing Order to the Web”, unpublished manuscript, Stanford University.

[11] T.H.Haveliwala, 1999, “Efficient Computation of PageRank, unpublished manuscript”, Stanford University.

[12] http://www.google-pagerank-improve.com.