Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 172 (2020) 165–171

Procedia Computer Science

www.elsevier.com/locate/procedia

9th World Engineering Education Forum, WEEF 2019

Influential analysis of web structure using graph based approach

Umadevi.K.S, Balakrishnan P, Geraldine Bessie Amali

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

Abstract

Web pages are documents commonly used as a container, which are stored in servers and accessible through the internet. This information can be retrieved through the use of browsers using mobiles, desktops and other computational devices. The challenge is these applications are they are expected to meet the user's requirement to provide the best information with minimum time consumption. Not only for sharing quality information, but these web pages are also used for marketing, banking, healthcare applications and so on. In the growing financial era, the influence of a web page for marketing another web page is driving a new culture in the IT sector. The influence of a web page is one of the major roles when it comes to linking one page to the other page. In this paper, a graph-based method is proposed for analyzing how one web page can influence access of other web pages. This algorithm is used to discover the information flow patterns and their influences on the network structure. The main motive of the proposed work is to analyze the above-said metrics and verify the key pattern of linking flow and their influence in the network structure. The analysis of the proposed algorithm focuses on network structure along with navigation using edges among nodes in a weighted graph and facilitates the best method of accessing the designated web page.

Keywords: Web structure; Graph Theory; Influence.

1. Introduction

Web is a massive, unstructured and dynamic data repository, which can deliver a large amount of information. Webpages are semi-structured and presents various information are presented to the user in a readable manner. Everyday people browse and visit a number of websites in the internet. The pillars that hold this together are the web pages. It is a document commonly written in Hypertext Mark-up Language (HTML) and is accessible through Internet using an Internet browser. A webpage is accessed by entering a URL address and may contain text, graphics and hyperlinks to other webpages and files [1; Fig. 1].

* Corresponding author.
E-mail address: umadeviks@vit.ac.in

10.1016/j.procs.2020.05.025

On most of the websites, the information contained is read and if there are any interesting hyperlinks, the link is followed by clicking on them to find more information or to perform a task. Web structure mining is a data mining technique to discover the summary of the Web site and how the different webpages are connected. The inner structure contains information about hyperlinks that can be navigated through this page and also compare with other schemas [2]. In general, a web page may be linked to other web page(s), so the aim of the work is to discover the relation amongst themselves.

Other than financial aids, humans are making use of several websites for a specific cause. Though variety of techniques explored in web mining, they are able to access and understood such a high level of information associated with it. But the existing techniques need to be explored further to facilitate them as they do [3]. Hence, the objective of this work to analyze the web structure mining by considering the nodes as web page in a graph, hyperlink as a directional link for their influential measure using their navigation pattern.

In this paper, an algorithm is developed to analyze the influence of webpages on other pages. The algorithm discovers the information flow patterns and the influences on the underlying network structure. A matrix is constructed that shows the frequency of visit on each page directed from another web page. Thus the most influencing path in the network to reach the designated page is found by analyzing the overall network structure.

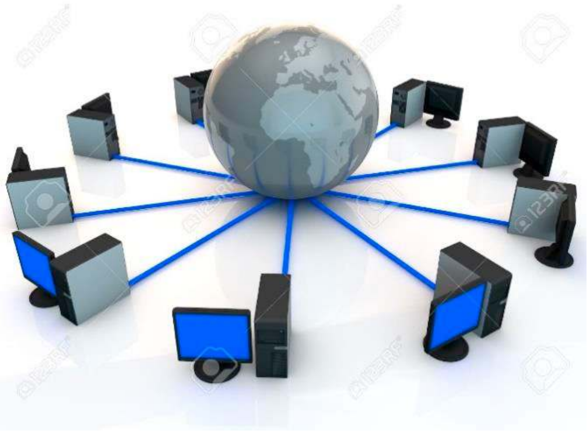

Fig. 1. World Wide Web

2. **Literature Review**

Usage of the graph theory based methods may also be helpful in examining the missing content and with respect to the context of the web structure [5]. While exploring the content using advanced graph theory concepts, it becomes feasible to predict the missing connections in the graph. The authors preferred Markovian chain model for their successful prediction by gathering information about the user navigation pattern at regular intervals.

A directed graph whose nodes are the webpages and edges pointing to the hyperlink between those pages can be used to study the behaviour for several reasons [4]

- Crawling statistics
- Sociology of the content in the web
- Behaviour of the algorithms used among the links
- Usage of web structure in navigation

For detailed analysis, researchers recommend two methods: Host level and Pay-level domain information [6]. First level of analysis contains the feature related information like degree of distribution, structure of the connected components either weak/strong and so on. In addition to these metrics, the proposed work focuses how these metrics would be helpful in influencing the access to its neighbor nodes for e.g. hyperlinked to other webpages. Because social influence always depends on some criteria such as strength of relationship, the network distance between users, temporal effect, characteristics of networks and individuals in the network. Researchers came up with large scale measurements and analysis of multiple online social networks, where they explored sites which provide the opportunity to study the characteristics of social network structures in order to identify the social influencing neighboring nodes.

In addition to various graph oriented techniques many other methods used for prediction and analysis as like Markovian Model addressed in the previous session. To overcome the issues associated with vector space model, authors prefer k-means algorithm to obtain information with minimum time requirement. The requirement for a framework with enhanced and ideal portrayal of web structure attracts much research for their contribution [7]. Graph theoretic technique framed the stage for the examination of protein structure data using the concept of sub-graph and the other properties associated with it to give a most effective powerful meaning for the relationship [8].

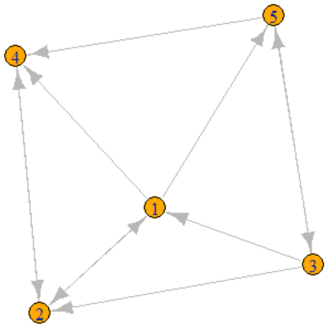

Fig. 2. Graph using PageRank Algorithm

Though the article considers a similar approach like page rank algorithm, it is quite different from PageRank algorithm [9]. PageRank counts the number of visits of users and estimate how important this website is. The Table 1, contains Rank for the nodes in the graph (Fig. 2) in terms of percentage. Even though, Page 2 and Page 4 has got the equal number of incoming edges Page 2 has got more percentage. It provides the mode of addressing the importance of a page with respect to the total users visiting the webpage.

Table 1. PageRank values for Fig. 2

Node	Rank
1	19.7
2	34.2
3	7.5
4	27.7
5	10.7

The objective of the proposed work is

- Identify the hyperlink in the webpages, convert the webpages and hyperlinks into graph and
- Calculate the centrality measures and Record the navigation pattern
- Using navigation pattern and Centrality measures, identify the influencing node.

3. Proposed System

Consider that, there are many websites that are connected and flow of each website said to be valid. A website in a network can have links to other sites. The visitor who visits the page will be navigated to the next page and from that page to the next and so on. There are a number of visitors who visit the same webpage regularly and may follow the same path to reach the same site. Thus these webpages are said to influence that particular website in the network.

Same page can direct to a site through different paths. When the user clicks on link on one page he will be navigated to the next page that contains some links to other page. Most of the users in the web wish to follow that path that navigates to the website as fast as possible. Different paths can be used to reach a website and the most influencing path in the network has the large number of visitors. We select a particular website in the network and find the different paths through which this network can be reached. The number of visitors of each page in the path is recorded and that represent the frequency of flow in the network.

Collecting the navigation pattern is used to assign a rank leading to the right destination. The algorithm keeps track of the frequency of flow paths in the network graph. The graph is converted into an adjacency matrix which is analyzed to find the most influencing path in the network. Consider Figure 2, to reach page 2 from page 5, there are several links connecting the nodes like 5-4-2, 5-3-2, 5-3-1-2, 5-3-1-4-2. But among these either reaching via page 4 or page 3 would be an ideal choice. Due to the conflict in accessing via either page 4 or page 5, this work suggest the usage of a navigation pattern to reach the designated web page

3.1. Statistical Measurements based on Web Structure

Using statistical measurements, a node's in-degree, out-degree, in-closeness, out- closeness and betweenness are evaluated. Globally, some web sites can have intrinsically some kind of higher influence than the others due to network structure and its behaviour. The influence is measured in terms of the frequency factor of each page in the network that is determined by the maximum number of visitors directed to a particular site in the network. The standard measures for analysis include the centrality of the network. The centrality indicator identifies the most influential nodes and understands the node importance in the graph. A node with high centrality value is considered to be the highly influencing node in the network. Centrality can be measured by connecting one node to other nodes.

Degree centrality is the number of total links of one node to other nodes in the graph. Nodes which have more links may be directed towards by more visitors and thus is relatively advantaged. Degree centrality for a directed graph or network has one of two measures:

- In-degree is amount of links that point towards the node of the web structure.
- Out-degree of the web structure has high out degree centrality, which has the ability to compare the information.

Closeness of the web structure can be found by measuring the degree of each node in the web structure. Betweenness can be analysing through the connected node which is extend to other node both are not connected to each other. After measuring each node it will take the absolute value and percentage of maximum possible node.

3.2 Frequency Flow Mine Algorithm for finding number of visitors

Input: Graph G, Frquency Flow
Output: P, Node influence value for all nodes
Algorithm: FrequencyFlowMine()

```
begin
```

```
        for each node i
                P[1,i] = { }
        if node i has at least f visitors then
                P[1,i] = {i};
        end
        k=1
        while P[k,i] is not empty do
                for each i generate G[k+1,i] from P[k,i] by adding each high frequency neighbour node of i to the
                paths in P[k,i] and thus keeps only high frequency paths with all individual nodes
                                k=k+1
        end
        k=1
        while k is not equal to '0' do
                Calculate the centrality of each n node in G[k,i]
                Calculate the betweenness of nodes i and its adjacent node in   P[k,i]
                Calculate the closeness value of  the nodes I and i+1 in P[k,i]
        end
        Generate the analysis graph containing nodes from the influencers set
 end
```

Algorithm considers all the possible information flows that are generated and are analysed based on the weight of each nodes. The nodes in the networks are dynamically connected or disconnected based their influence and centrality of each node, the in degree and out degree values. It also calculates the betweenness and closeness of the nodes in each flow pattern of the network. The weighted values of the network are represented in the form of an adjacency matrix and the degree values generated is also in the form of matrix. These values are inputted and a graph is generated reflecting the influence of the actors in the network. The generated graph is analysed to find the most influencing information flow path in the network.

4. Results and Discussion

The efficiency of the network is evaluated by calculating the betweenness, in-degree, in-closeness and out-closeness values of the nodes in the network. Different graphs are constructed from the network by using the FrequencyFlowMine() algorithm which represents different flow patterns. The graph is analysed by calculating the betweenness, degree of distribution, Closeness PageRank and Predicted Influencer.

Table 2. Predicted Result

Node	Degree of Distribution	Closeness	Betweeness	PageRank	Predicted Result
1	0	0.111	0.33	0.073	3
2	0	0.125	5.33	0.104	5.33
3	0	0.111	10.33	0.342	NA
4	0.33	0.142	9	0.17	3.33
5	0.1667	0.1	0	0.05	5
6	0.33	0.07	1	0.254	NA

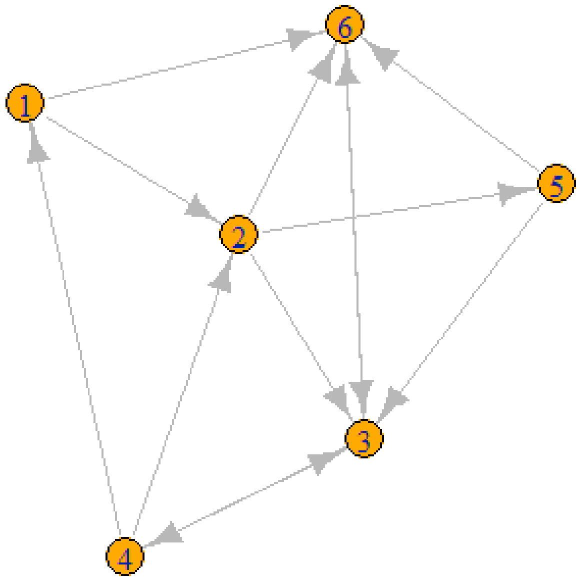

Fig. 3. Output Graph

The values are considered for analyzing the influence of flow frequency in the network. Hence, the predicted results depict that though there would be more usage in terms of in degree, this won't be able to influence other web pages thus resulting 0. The above table 2 contains all the values. Since node 2 is the ultimate influencer because it is connected to all the other node, the results show that it can be effectively used for navigating from source to destination node [Fig 3]. The frequency flow of network has different influence on the basis of all factors in the network.

5. **Conclusion**

Influence analysis of one web structure to another web structure aims at qualitatively and quantitatively measuring the visits with minimum possible time. The number of visits for the websites can be calculated through the graph-based structure with the possibility of maximizing the frequency of usage. In this paper, it introduced the web page with maximum influences by other web pages that will have minimal time to access. Then, it proposed to find the degree of the network structure, betweenness of the network structure and closeness of the network structure. The result shows that the effectiveness of finding patterns using influence and analyzing mining applications. The experimental results clearly show that the proposed frequency flow algorithm is extremely efficient and is well suited for several application areas such as the influence of the web site to other web sites. Web structure may also be used for mining the information about the hierarchy of relations through their hyperlinks.

6. REFERENCE

[1] Berners-Lee, T. J. (1989). Information management: A proposal (No. CERN-DD-89-001-OC).

[2] S.K.Madria, S.S.Rhowmich, W.K.Ng, and F.P.Lim. Research issues in Web data mining. *In Proceedings of First International Conference on Data Warehousing and Knowledge Discovery*, pp. 303-312, 1999.

[3] Keller, M., and Nussbaumer, M. Beyond the web graph: mining the information architecture of the www with navigation structure graphs. *In Proceedings of International Conference on Emerging Intelligent Data and Web Technologies*, pp. 99-106, 2011

[4] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., and Wiener, J. Graph structure in the web. *Computer networks*, Vol. 33, No. 1-6, pp. 309-320, 2000.

[5] Clauset, A., Moore, C., and Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature*, Vol. 453, No. 7191, 2008.

[6] Meusel, R., Vigna, S., Lehmberg, O., and Bizer, C. The graph structure in the web: Analyzed on different aggregation levels. *The Journal of Web Science*, Vol. 1, No. 1, pp. 33-47, 2015.

[7] Marcov A. Last M. and Kandel A., Model-Based Classification of Web Documents Represented by Graphs, *Proceedings of WEBKDD'06*, Philadelphia, Pennsylvania, USA, 2006.

[8] Artymiuk P. J., Spriggs R. V. and Willett P., Graph theoretic methods for the analysis of structural relationships in biological macromolecules*, Journal of the American Society for Information Science and Technology*, Vol. 56, pp 518 – 528, 2005.

[9] Li, Yanhong "Toward a qualitative search engine". Internet Computing, IEEE. *IEEE Computer Society*. Vol. 2, No. 4, pp. 24–29, 2002.