ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.11

Discovering Web Communities in Web Structure Mining–An Overview

Gandhimathi .K
Assistant Professor
Department of Computer Science
Sri Jayendra Saraswathy Maha Vidyalaya, Coimbatore, India

Abstract:
In World Wide Web, web pages are connected together with hyperlinks. The web structure mining is based on the graph structure of hyperlinks and it extracts the useful information from the structure of web data. Web structure mining aims to generate structural summary about web sites and web pages. There are several goals for web structure mining such as ranking important web pages, discovery of web communities, and analysis of the web graph from macroscopic point of view, modeling and simulating the process of web graph generation. This paper presents an overview on existing approaches for the discovery of web communities in web structure mining.

Keywords: web mining, web structure mining, and web community.

I.INTRODUCTION

A World Wide Web (WWW) is becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the solutions for knowledge discovery on the web. The web mining is the application of data mining techniques to automatically discover and extract information from web documents and services. Web mining covers a wide area of research communities such as databases, IR, machine learning and NLP. Kosala et al., [1] has suggested a decomposition of web mining as, resource finding, information selection and pre-processing, generalization and analysis. Resource finding is used for retrieving intended web documents. The information selection and pre-processing is used for automatically selecting and pre-processing specific information from web resources. Generalization is used to automatically discover the general patterns at individual web sites as well as across multiple sites. Analysis is the process of validation and interpretation of the mined patterns. Web mining is classified into three areas of interest such as web content mining, web usage mining and web structure mining. Web content mining is the process of extracting the useful information from content of the web documents. The web documents may consists of text, images, audio, video and structured records like tables and lists. The web content mining is applied on the web documents, which is the result produced from a search engine. Web usage mining is the process of analyzing user's browsing behaviour. It consists of a three-phase process such as, data preparation, pattern discovery and pattern analysis. The applications generated from this analysis can be classified as personalization, system improvement, site modifications and business intelligence. Web structure mining is the structure of a web graph consists of web pages as nodes, and the hyperlinks as edges connecting related pages. The web structure mining is the process of discovering structure information from the web. The main purpose of web structure mining is to extract previously unknown relationship between the web pages. Web structure mining categorizes the web pages and generates the information, like similarity and the relationship between different web sites [2]. The mining can be performed at the document level or at the hyperlink level. Web structure mining focuses on the identification of authorities i.e. the pages that are considered as important sources of information from many people in the web community. Web structure mining can be divided into two categories; First category includes extracting the patterns from hyperlinks in the web. Second category consists of mining the document structure such as HTML or XML tags.

II. WEB STRUCTURE MINING AND WEB COMMUNITIES

Web structure mining is used to identify the relationship between web pages linked by information or direct link connection. This structure data is discovered by the provision of web structure schema through database techniques for web pages. The connection allows a search engine to pull data relating to a search query directly to the linking web page from the web site the content rests upon. The completion takes place through use of spiders scanning the web sites, retrieving the home page and linking the information through reference links to bring forth the specific page containing the desired information. The hyperlink hierarchy is determined to find the path related information within the sites of the competitor links, connection through search engines and third party co-links. The web structure mining involves in modeling web site in terms of link structures. The mutual linkage information is used to find relevant pages based on the similarity or relevance between different web pages. Lee Giles et al.,[3] suggested that the rapid growth of World Wide Web had made dilemma for search engine designers. First dilemma is that no search engine covers more than about 16% of the web. In second dilemma the search engine resides between the precision and recall of the query result. In order to overcome the above problem a web community is introduced to enable the web crawler to effectively identify related subset of the web and also enables search engines and portals to increase the precision and recall of search results. A web community is a collection of web pages in which each member page has more hyperlinks within the community than outside the community. A web community is based on the structure of hyperlinks. Kumar et al.,[4]

suggested three reasons why one should be interested in discovering these communities. First reason is that communities provide valuable and up-to-date information resources for a user. Second, the communities represent the sociology of the web, which is easy to learn and understand the web. Third, communities enable target advertising at a very precise level. Web communities can be characterized as a phenomenon manifested by both link proximity and content coherence, but there has been significant success in identifying communities based on the link structure alone.

III. BASICS OF WEB COMMUNITY IDENTIFICATION

A search engine finds relevant pages by consulting inverted index and return pages that match some or all query terms. The query results are often too large to be inspected by user. So, there is a need to sort according to relevance. Kleinberg (1999) [5] realized that there are two types of pages that could be relevant for a query.

Authorities: Pages that contain a lot of information about the query topic.
Hubs: Pages that contain a large number of links to pages that contain information about the topic.
Mutual reinforcement: A good hub points too many good authorities, a good authority are pointed to by many good hubs. A practical use of the relationship is done by associating each page x with a hub score h(x) and with an authority score a(x), which are computed iteratively.
Hub Scores h(p): Hub scores are updated with the sum of all authority weights of pages it points to

$$h(x) = \sum_{(x,y) \in E} a(y)$$

Authority Scores a(p): Authority scores are updated with the sum of all hub weights that point to

$$a(x) = \sum_{(y,x) \in E} h(y)$$

A simple approach to determine relevant pages is to sort query results according to the number of in-links. The drawback here is universally popular pages would be considered to be highly authorities for all search terms they contain.

IV. WEB COMMUNITY IDENTIFICATION – A BACKGROUND STUDY

There are several algorithms and methods used for finding web communities. Some of the methods are described below. Hugo Zaragoza et al.,[6] proposed the Hypertext Induced Topic Search (HITS) technique, which is based on the following two **intuitions:** First, hyperlinks can be viewed as topical endorsements. A hyperlink from a page u devoted to topic T to another page v is likely to endorse the authority of v with respect to topic T. Second, the result set of a particular query is likely to have a certain amount of topical coherence. Due to these reasons link analysis is not performed on the entire web graph, but on the neighborhood of pages contained in the result set, as the neighborhood is more likely to contain topically relevant links. HITS which is a query dependent ranking algorithm performs the following steps to calculate hubs and authority scores along with the webpage out degree.

(1) It collects the root set that contain first *t* hits from a conventional search engine.

(2) It construct a base set which include all pages the root set points to and include pages that point into the root set $R \subseteq V$.

(3) It construct a focused sub graph, a graph structure of the base set $B \subseteq V$ and delete intrinsic links i.e., links between pages in A link same domain. A link section predicate P takes an edge (u, v) Є E. In this study, we use the following three link section predicates:
all(u, v) ↔ true
ih(u, v) ↔ host(u) ≠host(v)
id(u, v) ↔ domain(u) ≠ domain(v)
(4) It iteratively computer hub and authority scores.

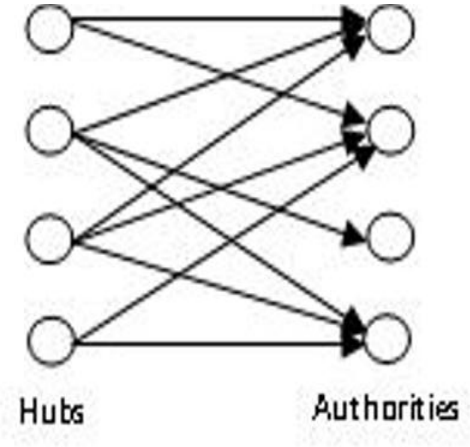

Figure.1. HITS Relevancy.

HITS have been used for identifying relevant documents for topics in web catalogues and for implementing "Related Pages" functionality. The main drawback of the HITS algorithm is that the hubs and authority score must be computed iteratively from the query result, which does not meet real-time constraints of an on-line search engine. Other problems are that it is not good enough to be applied in mining the informative structures, which converge into densely linked irrelevant pages called topic drift problem, which is notorious in the area of information retrieval. Kleinberg [2] suggested that the HITS algorithm could be used for finding related pages by providing the evidence that it might work well. Jeffery Dean et al.,[7] suggested two related page algorithms such as companion algorithm and cocitation algorithm, which is an extension of HITS algorithm to exploit not only the links but also the order of a page. The companion algorithm takes the input as a starting URL u and consists of the following steps,

(1) By building a vicinity graph for u.
(2)It contract duplicates and the near- duplicates in this graph.
(3)It computes edges of the weights based on host-to-host connections.
(4)It calculates a hub score and an authority score for each node in the graph and return the top ranked authority nodes. This algorithm is a modified version of the HITS algorithm. The cocitation algorithm finds pages that are frequently cocited with the input URL u. i.e., it finds other pages that are pointed to by many other pages that all also point to u. Two nodes are cocited if they have a common parent. The number of common parent of two nodes is their degree of cocitation. Cocitation algorithm first chooses B arbitrary parents of u. For each of these parents p, it adds a set S up to BF children of p that surrounds the link from p to u. Elements of the S are siblings of u. For each node s in S, it determines that degree of cocitation of s with u. The algorithm returns the 10 most

frequently cocited nodes in S as the related pages. The performances of both the algorithms are correlated, but cocitation algorithm is used extensively. Lee Giles et al.,[3] proposed the following methods for identifying the web community which is ideal community, approximate communities and Expectation Maximization (EM) algorithm. Ideal community is a undirected graph where each edge has unit capacity. Thus the graph induced from the web would have edge directions removed. Source link count is problematic if only a single source vertex is used so, it chooses virtual sink vertex. Approximate communities are a true web page that is not used as a sink, but as a artificial vertex to facilitate a connection by the vertices in graph that are most distant from the source. This method is used for identifying web communities that has only limited success when a small number of seed web pages are provided. The main drawback of this method is, only small subset of a community can be identified, to solve the expectation maximization algorithm is used. Expectation maximization algorithm used a two step process Expectation ("E") and Maximization ("M"). E corresponds to use the maximum flow algorithm to identify a subset of the community. The newly discovered web sites are relabeled as seeds, which are partially re-crawled from the new seeds to induce a new graph. The maximum flow procedure is executed, and the process gets iterated. A maximum flow-based web crawler uses a approximate community by directing a focused web crawler along link paths that are highly relevant. The EM approach incrementally improves the crawl results by re-seeding the crawler with highly relevant sites. The max-flow based community discovery can extract larger, more complete communities. However, it cannot find the theme, the hierarchy, and the relationships of web communities. Masashi Toyoda et al.,[8] proposed the Related Page Algorithm(RPA) that is applied on each seed and then it investigates how each seed derives other seeds as a related pages. It first builds a sub graph of the web around the seed, and it extracts authorities and hubs in the graph using HITS. Then authorities are returned as the related pages. To identify web communities and to deduce their relationship, first put focus on the relationship between a seed page and derived related pages by the algorithm. A page s is derives a page t as a related page, and the t also derives s as a related page. The both pages s and t are pointed to by similar sets of hubs. When applying related page algorithm to one of the fans, the page derives the original fan, because the fan pages are mutually linked by each other that is pointed to by similar sets of hubs. If each fan derives the other fans as related pages, acknowledge that these fans form a fan community. When applying RPA algorithm to the official page, it derives the official pages of other teams as related pages instead of the fan page. In this case, the official page is related to the fan community, but the page itself is a member of the community. This mechanism is used to find related communities. Symmetric derivation relationship is used for identifying communities. Georgious Paliouras [9] suggests that the user is not considered as an isolated individual any more, but as a member of the one or more communities. Communities arise in a number of different ways. The Social networking tools typically allow users to proactively connect to each other. Alternatively, the data mining tools discover communities of connected web sites or communities of web users. The community discovery has to be based on information provided by the users that may imply commonalities and associations among them. Typically, the information that is taken implies a common interest of users on the products or services provided by a particular web site. There are several observations that help us to infer the interest of a user in a particular item. Among them, most common ones are selection of the item for viewing, purchase of an item and explicit rating of the item.

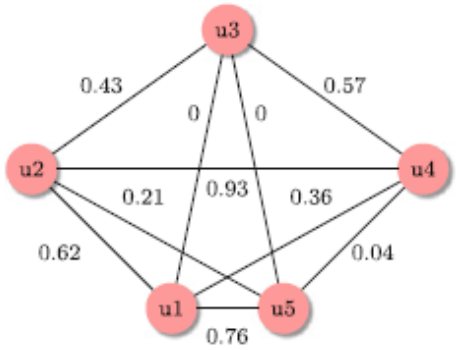

Figure.2. Weighted graph of a user.

To identify users interest an extended version of server logs is used to record the other information such as the id of the user, if the user has logged in, or if the referring web page that the user was viewing before the hit. In order to identify these communities, one needs to measure the degree of similarity between users, in terms of their expressed interest about items.

$$R_i(\mathbf{uc}_{il}, \mathbf{uc}_{im}) = \frac{\sum_{k=1}^{T_i} uc_{ilk} \times uc_{imk}}{\sqrt{\sum_{k=1}^{T_i} uc_{ilk}^2} \times \sqrt{\sum_{k=1}^{T_i} uc_{imk}^2}}$$

$$R_i(\mathbf{uc}_{il}, \mathbf{uc}_{im}) = \frac{\sum_{k=1}^{T_i} (uc_{ilk} - \overline{uc_{il}}) \times (uc_{imk} - \overline{uc_{im}})}{\sqrt{\sum_{k=1}^{T_i} (uc_{ilk} - \overline{uc_{il}})^2} \times \sqrt{\sum_{k=1}^{T_i} (uc_{imk} - \overline{uc_{im}})^2}}$$

Based on the similarity of the users, measured by *Ri* , one can construct a weighted graph. *UGi = (UCi, UEi, Ri)*, the vertices of which represent the users, and the edges *UEi*, weighted by *Ri* , denote the degree of similarity among the users. Having measured the similarity *Ri* among the users *UCi* of a site *si*, communities are defined simply as clusters of similar users. Therefore, generic clustering methods have been used, in order to discover the communities in usage data. From the above study it is made clear that detecting the web communities in web structure is great importance in the sociology, biology and computer science disciplines where systems are often represented as graphs. To detect communities many researchers have proposed different methods, each of which has advantage over the other depending on the requirements. Most of the recent research is based on community-driven personalization which creates more impact for larger web structure.

V. CONCLUSION

This paper presents an overview of various methods for discovering the communities in web structure mining. To utilize website as a business tool web structure mining is most essential. Discovering the communities is one of the goal in web structure mining. Community structures are quite common in real network. Many researchers proposed their algorithm only for identifying a small subset of community. In future, further study may be carried to identify the large set of communities.

VI. REFERENCES

[1]. Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", *ACM SIGKDD Explorations Newsletter*, June 2000, Volume 2 Issue 1.

[2]. Zhiguo Gong, "Web Structure Mining: An Introduction". Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China.

[3]. Gary William Flake, Steve Lawrence, C. Lee Giles, "Efficient Identification of Web Communities".

[4]. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "Mining the link structure of the worldwide web" *IEEE Computer*, 32(8):60–67, 1999.

[5]. J. M. Kleinberg. "Authoritative sources in a hyperlinked environment". In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, 1999.

[6]. Marc Najork, Hugo Zaragoza, Michael Taylor, "HITS on the Web: How does it compare".

[7]. Jeffrey Dean, Monika R. Henzinger, "Finding related pages in the World Wide Web". *Compaq Systems Research Center, 130 Lytton Ave., Palo Alto, CA 94301, USA*

[8]. Masaru Kitsuregawa, Masashi Toyoda, Iko Pramudiono, "Web community mining and web log mining: Commodity cluster based execution".

[9]. Georgios Paliouras, "Discovery of Web user communities and their role in personalization".

[10]. P Ravi Kumar, and Singh Ashutosh kumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.

[11]. X. He, C. Ding, H. Zha, and H.D. Simon. Automatic topic identi_cation using webpages clustering. Proc. IEEE Int'l Conf. Data Mining. San Jose, CA, pages 195{202, 2001.

[12]. D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In Proc. 9th ACM Conference on Hypertext and Hypermedia (HYPER-98), pages 225{234, 1998.

[13]. Suppawong Tuarob, Prasenjit Mitra, and C. Lee Giles, "Improving Algorithm Search Using the Algorithm Co-Citation Network".

[14]. T. Murata, "Finding Related Web Pages Based on Connectivity Information from a Search Engine," Poster Proc. of the Tenth Int. World Wide Web Conf. (WWW10), 2001.