Web Mining for Social Network Analysis: A Review, Direction and Future Vision

[1]Verma Gulhati Manjula, [2]Yumnam Jayanta Singh
Dept of Computer Science & Engineering and IT,
School of Technology, Assam Don Bosco University,
Airport Road, Azara, Guwahati - 781017, Assam. INDIA.
manjulaverma[@]gmail.com, Jayanata[@]dbuniversity.ac.in

Abstract:*Although web is rich in data, gathering this data and making sense of this data is extremely difficult due to its unorganised nature. Therefore existing Data Mining techniques can be applied to extract information from the web data. The knowledge thus extracted can also be used for Analysis of Social Networks and Online Communities. This paper gives a brief insight to Web Mining and Link Analysis used in Social Network Analysis and reveals the algorithms such as HITS, PAGERANK, SALSA, PHITS, CLEVER and INDEGREE which gives a measure to identify Online Communities over Social Networks. The most common amongst these algorithms are PageRank and HITS. PageRank measures the importance of a page efficiently with the help of inlinks in less time, while HITS uses both inlinks and outlinks to measure the importance of a web page and is sensitive to user query. Further various extensions to these algorithms also exist to refine the query based search results. It opens many doors for future researches to find undiscovered knowledge of existing online communities over various social networks.*

Keywords: Web Structure Mining, Link Analysis, Link Mining, Online Community Mining.

1. INTRODUCTION

1.1 Data Mining

The evolutionary path of database technology, led to the need of multidisciplinary field of Data Mining which emphasises on data analysis. Hence Data Mining is the process of extracting information from huge sets of data. The information or knowledge thus extracted can then be used for Market Analysis , Fraud Detection, Customer Retention, Production Control, Science Exploration, etc [25]. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining as Descriptive Functions and Classification & Prediction Functions. The descriptive function deals with the general properties of data in the database. Some of these functions are Class/Concept Description, Mining of Frequent Patterns, Mining of Associations, Mining of Correlations, and Mining of Clusters while Classification is the process of finding a model that describes the data classes or concepts. The derived model can be presented as Classification (IF-THEN) Rules, Decision Trees, Mathematical Formulae, Neural Networks, etc [26].

1.2 Knowledge Discovery

Knowledge discovery has wider scope than data mining [36].Data mining forms an essential step in the process of knowledge discovery. The steps used for knowledge discovery process are Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation.

1.3 Data Mining in Web

J Han and M Kamber[1] have stated that this extracted information or knowledge can also be used in various applications. One of them is online data or Internet web surf aid. M Júnior and Z Gong [2] in their work stated that the World Wide Web is a popular mode of publishing in current times. Due to which it has become a huge resource of information for each and every user. Yet, it is different from traditional databases in terms of its size, complexity, schema and dynamic nature. These differences make the publications on web extremely unorganised because of which this information cannot be used effectively and efficiently. Thus, existing data mining techniques can be used to automatically discover and extract information from the varied web resources. This extracted information can further be used to study the online communities existing over the social networks. Today, Social Network extends a platform for social interactions between individuals or group of individuals. This grouping of individuals for common interest or topic gives birth to online communities over Social Networks. The most common Social Networks being used for Online Community existence are FaceBook, Twitter and Blogger. Through this paper we have tried to surface the existing algorithms used to analyse the existing web structures in Social Networks in identifying the Online communities existing over them. This knowledge can then be used to discover interesting patterns on online communities over different Social Network Websites. For example the problem of finding importance of a web page was solved by a graph mining based algorithm PAGERANK which was used by Google. This paper is structured into various sections for the ease of study. In Section 2, we present an overview of Web Mining, mentioning the various theories in this area of research. In Section 3, we progress into the depth of

Web Mining discussing Web Structure Mining. Later, in Section 4 and 5, we give an understanding of the web graph and web model which is eventually used for representing Social Networks over the Web. Consequently, in section 6 and 7 we discuss Hyperlink Analysis and link algorithms, to brief the reader with the model and the algorithms used for building the application. Later in section 8, we have review the interesting structures in web graph which help to identify an online Community over social network using the link algorithms and then section 10 gives insight into the future research areas.

2. DIFFERENT THEORIES OF WEB MINING

The study of Chakraborti et al. [6] state that we interact with the web for the following reasons:

- To find relevant information using a keyword query at a search engine without low precision and low recall.
- Discovering new knowledge from the web data using data mining techniques
- Synthesis of personalized web pages for each individual based on their preferences like style, content or presentation.
- Learning about individual users to know their habits as to what they want and what they do.

These above mentioned problems can be solved using Web Mining technique in two ways.

- Firstly, they can be used to provide direct solution to the given problem.
- Secondly, they can be integrated with bigger applications to address these problems.

The other solutions can be from different research areas like databases, information retrieval, Artificial Intelligence and Natural Language processing. It is emphasised that three broad web mining operations used to discover data from the web are Clustering, Associations and Sequential Analysis. Clustering can be used to group users or pages based on specific characteristics, Associations helps to understand which web pages will be accessed together and Sequential Analysis specifies the order in which web pages are accessed.

In 1996, O Etzioni [4] first mined the term web mining. Their study is divided into subtasks which were further reviewed by Kosala and Blockeel [7] as:

i. Information Retrieval (or Resource Finding)
ii. Information Extraction(or Information Selection and Pre processing)
iii. Generalization
iv. Analysis

1. Information Retrieval

It is the process of locating unfamiliar web Sources. Web Sources constitute of web documents and web services on the web. This searching is done with the help of indices. Some of the popular indices being used are created by Web Crawler, Alta Vista, Meta Crawler. These softwares can scan millions of web documents and store a respective index in the document. On the other hand Meta Crawler's query language allows searching for phrases with respect to a geographic region or internal domain. In Parallel it posts the keyword of the query to nine indices, collect and prunes the responses returned and hence provide the user with high quality information. Today, the resource discovery systems make use of automatic text categorisation technology for retrieval of information from the web.

2. Information Extraction

This is the process of automatically extracting relevant information from the retrieved web resources. Initially, hand coded wrappers were written to access resources but with time this information is dynamically extracted reducing the need of hand coding. For eg.Harvest, FAQ-Finder, Internet Learning Agent (ILA), Shopbot etc.

3. Generalization

Once the information has been extracted the general patterns in it are discovered using Generalization. A major problem in learning about web is the fact that web is unorganised. There is abundant data on web but it is unlabelled. Also, a lot of data mining techniques require labelled inputs e.g. Whether a webpage is a home page or not can be specified with a label as positive or negative. Yet, techniques such as sampling, clustering reduces the need of labelled data while Ahoy! gives a solution to the problem of labelling based on its heuristics filtering algorithm.

4. Analysis

It refers to the validation and interpretation of the mined data. Interactive query triggered knowledge discovery is important part of this step.

According to Green et. al [5] web mining can be viewed from the perspective of agent paradigm with various agents as:

a. Filtering Agents
b. Software Agents
c. Intelligent Agents

Filtering Agents are also referred to as Distributed Agents and are used for knowledge discovery in data mining process. *Software Agents* comprise of User Interface Agents and Mobile Agents. User Interface Agents helps to increase the productivity of user's interaction with the system by personalization with the help of IR Agents, Information filtering Agents and personal assistant agents. There exist two approaches used for designing an *Intelligent Agent.* First is the Content based approach which helps the user to search he information based on analysis of content for user preferences and is generally used in web content mining. Second is the Collaborative approach which helps the user to search the

information based on the searches made by similar users. It analysis the user profiles, sessions and transactions to find user with same interest and give recommendations.

As per the study of Madira et. al [8], web mining techniques can be categorised into 3 areas based on which part of web to be mined. They are:

a. Web Content Mining
b. Web Structure Mining
c. Web Usage Mining

a. Web Content Mining

It is the discovery of useful information from the web contents or documents or data. Kosala and Blockeel [7] stated that there is a lot of data available on the web which is of different type like Government data, data in digital libraries, e-commerce data, business transaction data, customer information data etc. Apart from these, there is hidden data existing on the web which cannot be indexed and searched by the web crawlers because it is either generated dynamically or reside in database or is private. Also, the data on the web can range from textual, image, video, metadatato hyperlink data. The different forms in which web content data can exist are:

i. Unstructured Data is the free text data available on the web.
ii. Semistructured Dataconsist of the HTML documents on the web.
iii. Structured Datacomprises of the data generated from the tables or databases in form of HTML Pages.

Based on the different types of web content data, the various instances of web content mining are [7]:

i. Web Text Mining used for mining unstructured web data,
ii. Web Multimedia Mining used for mining multimedia data, and
iii. Web Hypertext Mining used for mining Semi-structured data.

The research done in web content mining can be viewed from two perspectives [7]:

i. IR View, which filters the information based on user's profile. It is used for mining semi-structured and unstructured data. General mining methods used here are Naïve Bayes, Induction Logic Programming, Rule learning, Classification Algorithm, Association Rule etc.
ii. DB View, models the web data based on queries and is generally applied to semi-structured and structured data. Methods are Proprietary algorithm, Association Rules etc.

b. Web Structure Mining

It is used to study the topology of hyperlinks [25]. We shall discuss web structure mining in depth in the following section.

c. Web Usage Mining

It helps to mine the secondary data generated by the user's interaction over the web like web surfer's session or behaviour and is available from proxy server log, web server access logs, mouse clicks etc. The different phases in which web usage mining can be divided are [1]:

i. Pre-processing: Here, the noisy, incomplete and inconsistent data is treated for future usage. It includes data cleaning, data integration, data transformation and data reduction.
ii. Pattern Discovery: To identify different user patterns various methods and algorithms are applied like machine learning, data mining, pattern recognition, statistics etc.
iii. Pattern Analysis: the pattern obtained are understood, visualized and interpreted in this phase.

Here, we like to note that there is no clear boundary between the above three categories of web mining. These could be used in isolated fashion or combined with each other in an application [2].

3. WEB STRUCTURE MINING

As mentioned by Kosala and Blockeel [7] and Pujari [19], Web Structure Mining tries to discover the underlying link structures of the web by building a model which can be used to categorize web pages and generate important information like similarity or relationship between web pages. Apart from this, it can also be used to find authorities and hubs. Algorithms used to model web topology are HITS, PAGERANK and CLEVER. Improvement of HITS by adding content information to link structure or by using outlier filtering, can be done for web page categorization and discovering micro communities on web.

According to Pujari [19], the web structure mining can be used for:

- Finding quality of page in terms of Authority of a Page and Ranking of a Page.
- Finding interesting web structures like graph patterns for Co-citations, social choice etc.
- Classifying web pages according to various areas of interests.

The research at hyperlink level is also known as Hyperlink Analysis.

According to Júnior and Gong [2], the challenge for web structure mining is to deal with the structure of the hyperlinks within the web itself which started with link analysis and gradually resulted into link mining [27]. This research area is at the intersection of link analysis, web hypertext mining, relational learning and inductive logic programming and graph mining. The web can be viewed as a collection of objects with

no unifying structure. The objects in the web are web pages, and links are in, out and co-citations [2]. Attributes include HTML tags, word appearances and anchor texts. Because of the diverse nature of these objects, the traditional data mining technique of Information retrieval cannot be used directly hence some changes in the traditional data mining technique were applied to perform link analysis. The tasks of link mining are [2]:

a. Link based classification: This task is to predict the category of a webpage based on words in the page, links between the pages, anchors, HTML tags etc.
b. Link based Cluster Analysis: This task is to find sub classes or groups or clusters. Here, similar objects are grouped together and dissimilar objects are grouped together into different groups which can then be used to discover the hidden patterns from the web.
c. Link Type: This task predicts the type of link between two objects or predicts the purpose of a link between the two objects.
d. Link Strength: This task emphasises on the importance of a link by associating links with weights.
e. Link Cardinality: This task is to predict the number of existing links between objects.

4. THEORIES ON GRAPH MINING AND IT'S APPLICATION

According to Broder et al [9], Web can be represented as Graph. The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting two related pages. We can view any collection V, of hyperlinked pages as a directed graph

$$G = (V, E) \quad (1)$$

Where,

V is the collection of hyperlinked pages(or nodes), and

E is the collection of Edges.

The nodes corresponds to the pages, and a directed edge from p to q is represented as

$$(p, q) \in E \quad (2)$$

This directed edge indicates the presence of a link from p to q. The outdegree of a node p is the number of nodes to which it has links, and the indegree of p is the number of nodes that have link to it.

Web Structure Terminology:

- Web-Graph: A directed graph that represents the web.
- Node: Each Web Page is a node of the web graph
- Link: Each hyperlink on the web is directed edge of the web graph
- In-Degree: The In-Degree of a node, p is the number of distinct links that point to p.
- Out-Degree: The out-degree of a node p, is the number of distinct links originating from p, pointing to other nodes
- Directed Path: A sequence of links, starting from p that can be followed to reach q.
- Shortest Path: of all the paths between nodes p and q which has the shortest length.
- Diameter: The maximum of all the shortest paths between the pair of nodes p and q, for all pairs of nodes p and q in the web-graph.

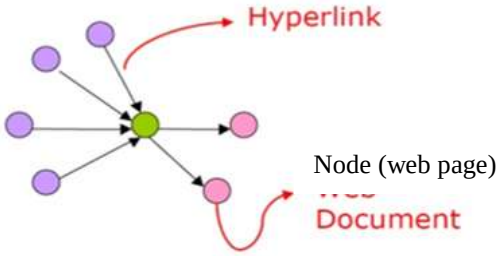

Figure 1: A Web Graph [28]

Figure2: Degrees of Nodes [28]

5. HYPERLINK ANALYSIS TECHNIQUE

The three areas that form the fundamental blocks for building various applications based on hyperlink analysis are Knowledge Models, Analysis Scope and Properties, Measures and Algorithms [10 and 28].

Knowledge Models are the underlying representatives that form the base to carry out the application specific task. The *scope of analysis* specifies if the task is relevant to a single node or set of nodes or the entire graph. The *properties* are the characteristics of single node or the set of nodes or the entire web. The *measures* are the standards for the properties such as quality, relevance or distance between the nodes. Algorithms are designed for efficient computations of the measures

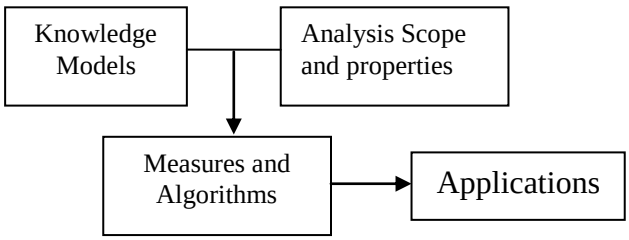

Figure 3: Link Analysis Model [28]

Basic knowledge models are needed to start research in the field of hyper link analysis , different measures are applied on this model to generate algorithms for achieving targeted application objective . Different kind of models, based on graph structure statistical methods or network flow have been proposed, some of them are:

- Graph Structure Model

There are various graph structures used for mining the web, Each of these represents certain concepts and information needed for mining and may comprise of single node, multiple nodes or the whole set of nodes that constitute the graph.

1. Single Node Model

Single Node Models are graph structures that consists of a single node and the links pointing to or away from it..

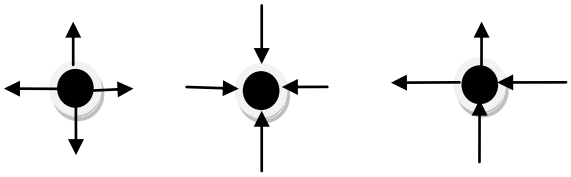

Figure 4: Single Node Model [36]

2.. Multiple Nodes Model

Multiple nodes models deals with graph structure which contains a number of interconnected nodes and their links. The concepts they reflect are:

Direct reference, Indirect reference, Mutual Reference, Co-Citation, Co-Reference, Directed Bipartite Gaph, Complete Bipartite Gaph, Bipartite Core and Community.

Direct reference is a concept where a node directly points to its adjacent node while for indirect reference a node is pointed to by another node which is further pointed to by an adjacent node. In case of mutual reference two nodes directly point to each other. Also, in Co-citation a node points to two other nodes indicating a similarity between pages while in Co-reference two nodes points a node indicating similarity between them.

3. Whole Grap Structure

The Web Model (The Bow-Tie Model)

As indicated by Johannes Furnkranz [18] the analysis of the structure of the web graph looks like a giant bow tie, which has strongly connected core component (SCC) of 56 million pages in the centre and two components with 44 million pages on either side, first is referred to as IN set containing pages by which the SCC can be reached and other is OUT set containing pages that can be reached from the SCC . Also, there are 'tubes' that reach the OUTset from the IN set without passing through the SCC and tendrils that connect the IN set (or the OUT set) to the other components.. Lastly, there are several components that cannot be reached from any point in this structure.

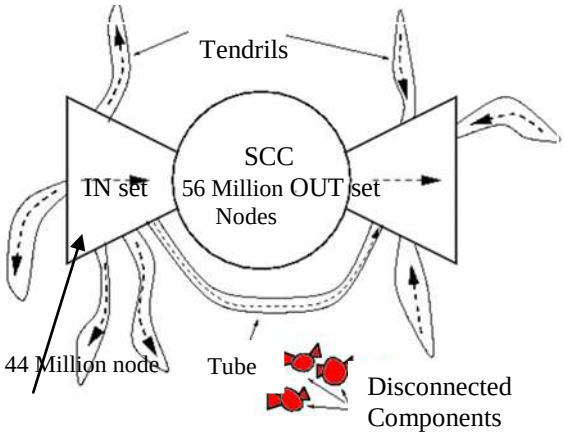

Figure 5: The Bow-Tie Model of the Web [28]

- Markov Models

According to Desikan et al [36], the underlying principle of an 'm' order Markov chain is that given the current state of system, the evolution of the system in the future depends only on the present state and the past 'm-1' states of the system.

- Maximal Flow Models

The s-t maximal flow problem can be described as: Given a graph G=(V,E) whose edges are assigned positive flow capacities, and with a pair of distinguished nodes s and t, the problem is to find the maximum flow can be routed from s to t. s is known as source node and t as sink node.

- Probabilistic Relational Models

A probalistic relational model is a combination of the relational model and Bayesian belief network. Accordingly the relationship among attributes within a class and the relationship of attributes across different classes can be modelled by assigning different probability distributions.[36]

- Other Models

Beyond the above mentioned models, other models can also be developed and used for hyperlink analysis.

Due to the dynamic nature of World Wide Web, 'Stability' becomes an important feature of Link Analysis Algorithms. The link analysis to give a robust view of authoritativeness of pages, should be stable under small perturbations on the web. Alternatively, a small change in web topology should not affect the overall link structure, and hence stability should reflect this insight correctly. Also, there exist some practical implications of stability issues like 'link spamming', where a good link analysis algorithm is resilient to any malicious attempt of web designers to promote the rank of their pages by adding or removing few links to or from the page. Some of the Link Analysis Algorithms are:

- *INDEGREE*

This is a simple algorithm proposed by Upstillet al [17]. According to them, the pages which have more incoming links

are more popular than the other pages. The major problem with this algorithm was that it could not capture the authority of a node. In a graph G: for each node i,

$$Ai=|Bi| \quad (3)$$

- *PAGERANK*

The PAGERANK Algorithm proposed by L Page and S Brin [11] calculates the importance of web pages using the link structure of web graph. Accordingly a page will have high page rank if there are many pages that point to it, or if there are some high rank pages that point to it. The PAGERANK algorithm is defined as, "We assume a page has pages $P_1,P_2,....P_N$ which point to it. The parameter d is damping factor which can be set between 0 and 1 and is usually 0.85." The out_deg(P) denotes the number of links going out of page P. It is being used by Google [2].The PAGERANK of a page P is given shown as follows [28] :

$$PR(P) = \frac{d}{N} + (1-d)\left(\frac{PR(P1)}{OutDeg(P1)} + \frac{PR(P2)}{OutDeg(P2)} + \frac{PR(P3)}{OutDeg(P3)}\right) \quad (4)$$

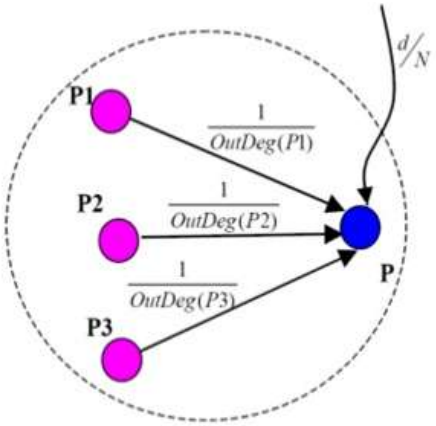

Figure 6: PAGERANK [28]

The formula used for calculating the PAGERANK of a page is recursive, starting with any set of rank and iterating the computations till it converges. The algorithm for PageRank is as follows [30]:

Step 1. Initialise the rank value of each page by 1/n where n is the total number of pages to be ranked.

Step 2. Consider some value of damping factor 'd' such that 0<d<1 e.g. 0.85,0.15 etc.

Step 3. Let PR be an array of elements representing PageRank for each web page. Then Repeat for each node i where 0<i<n.

PR[i]←1-d

For all pages Q that have a link to i, compute

PR[i]←PR[i]+d*A[Q]/Q_n

Where Q_n=number of outdegree of Q

Step 4. Update the value of A[i]=PR[i] for 0<i<n

Repeat from Step 3 till PR[i] value converges i.e. value of two consecutive iteration is similar.

Although PageRank is stable on the class of all directed graphs, it also provides a good support for web information retrieval, clustering and web knowledge discoveries. Yet,there were some problems noticed with this algorithm. T Haveliwala [30] noticed that original PAGERANK algorithm pre-computed ranking vector based on all pages. This ranking vector is computed and used later for queries. The ranking was independent of specific queries using it. This limitation was resolved by assigning weight for each page, for each topic. Weighted PageRank takes into account the importance of both the indegrees and outdegrees of pages and distributes rank based on the popularity of the pages. This is better than standard Page Rank as it returns larger number of relevant pages for a given query [35].

- *HITS (Hyperlink Induced Topic Search)*

It is a Link analysis algorithm that rates web pages on the concept of authority and hub. It was proposed by Jon Kleinberg [12] and is based on the principle that a document has a high weight authority if it is pointed to by many document with high hub weight and vice versa, and a document has a high hub weight if it points to many documents with high authority weight and vice versa.

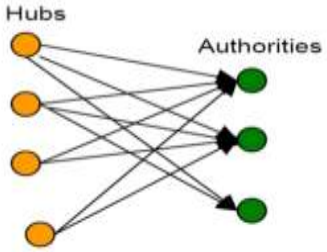

Figure 7: Hub and Authorities [28]

The steps in this algorithm are:

1. A subgraph (R) is created based on the query (keywords) by the algorithm.
2. Then, the weight of hubs and authorities for each node are calculated.
3. Further, expand R to a base set B, of pages linked to or from R.
4. Again, calculate weights for authorities and hubs.
5. Pages with highest ranks in B are returned.

HITS provided good results. However, it is not stable on the class of authority connected graphs and did not work well in few cases:

a) At sometimes a set of documents on one host points to a single document on another host or a single document on one host points to a set of document on another host. These situations may give a misleading result for a good hub or a good authority.
b) Automatically generated links by the tools may again provide wrong definition of good hub or good authority.

c) Sometimes pages point to other pages which are non-relevant to the query topic. This may lead to wrong results for hub and authorities.

The solution to the above mentioned problems was given by Randomized HITS and CLEVER.

- *CLEVER*

This algorithm was also given by Kleinberg. It identifies authoritative and hub pages, considering Authoritative Pages to be highly important pages as these Pages are the best source for requested information while HubPages only contain links to highly important pages. It was first used by IBM search Engine[2].

- *SALSA (Stochastic Approach for Link Structure Analysis)*

It was proposed by Lempel and Morgan [14] and is an improvement to HITS with the fact that authoritative page must be visible for thousands of pages of data set; hence it works on the principle of assigning weighted in/out degree values to the pages. This algorithm assigns high scores to hub and authority web pages based on the quantity of hyperlinks among them. It bears a close resemblance to HITS as this algorithm also works on the focused subgraph and assigns two scores to the web page: hub score and authority score. On the other hand it is alike PAGERANK as this algorithm also computes the scores by simulating a random walk through a Markov chain that represents the web graph, yet SALSA however works with two different Markov chains: a chain of hubs and a chain of authorities. It proceeds with two random walks on web pages; one by following a backward-link and second by following the forward-link alternately, or one by following a forward-link and then following a backward-link alternately. During the former random walk, authority weights are defined, and during the latter random walk, the hub weights are defined. Thus, SALSA assigns separate hub and authority scores to each page. SALSA is stable on the class of authority connected graphs but unstable on the class of directed graphs [31]. A solution to this problem was given by Randomised SALSA.

- *HUBAVG(Hub Averaging)*

This algorithm is combination of HITS and SALSA. It was proposed by Borodin et al [15] to remove the limitations of HITS. This algorithm tries to reduce the instability effect of HITS. Here, the calculation of scores of authority is same as HITS but the hub scores is the average of scores of authority. The basic principle of this algorithm is that a page is considered to be a good hub if it links to good authorities and its hub scores are calculated by considering only the scores of authority that are greater than or equal to the average score and vice versa for a good authority. The limitation of this algorithm is that a hub is scored low as compared to a hub pointing to equal number of equally good authorities if an additional link of low quality authority is added to it [32].

- *PHITS*

It is given by Cohn and Chang [16].They proposed a statistical algorithm to determine the two categories Authorities and Hubs. This model explains two types of variables, the quotes 'c' of a document 'd', based on a small number of common variables 'z' also called as aspects or factors. The model is then described statistically. HITS can estimate the probabilities of authorities while HITS can only provide the scalar magnitude of authority [33].

6. DISCUSSIONS AND RECOMMENDATIONS

6.1 SOCIAL NETWORK ANALYSIS

According to Pujari [19], Social Network Analysis is yet another way of studying the web link structure. It uses an exponential damping factor in the algorithms. The social network studies the ways to measure the relative standing or importance of individuals in a network. The basic premise here is that if a webpage points a link to another web page, then the former is endorsing the importance of the latter in some sense. Also, if there exists a link from a node to the other node and back from the latter to the former, it signifies some kind of mutual reinforcement while links from one node to different nodes shows existence of Co-Citation.yet another important premise says that if many nodes points to a node but there is no link moving out of the latter node then it is a social choice and there is possibility of existence of Online Community in part of the web graph.

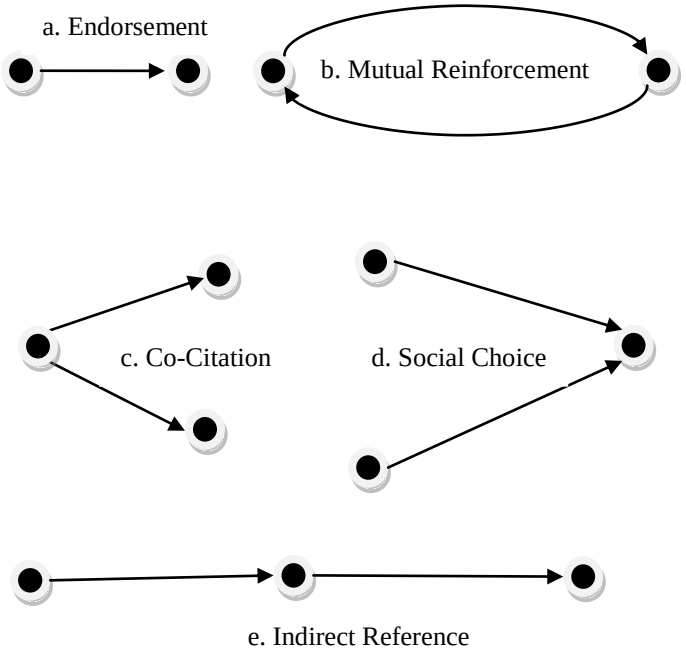

Figure 8 : Interesting Web Structures

Kautz et al [21] in his work "The Hidden web" proposed a measure *Standing of a Node* based on path counting. For nodes p and q, let $P_{pq}^{(r)}$ denote the number of paths of length exactly r from p to q. Let b<1, be a constant small enough so that equation number 3 converges [21].

$$Q_{pq} = \sum_p B_r P_{pq}^{(r)} \quad (5)$$

Here, b_r is the damping factor which varies with the length of the path. Then σ_q standing of the node q, is defined as

$$\sigma_q = \sum Q_{pq} \quad (6)$$

Kleinberg [22] discusses a heuristic method of giving weights to the links. A link is said to be transverse link if it is between pages with different domain names and intrinsic if it is between pages with the same domain name. Intrinsic links convey less information about the page than transverse links so they are not taken into account and hence deleted from the graph.

Botafogo [20]proposes another way of ranking pages through the notion of Index Node and Reference Node. An Index Node is one whose outdegree is significantly larger than the average outdegree of the graph. A Reference Node is one whose indegree is significantly larger than the average indegree of the graph.

For determining collection of similar pages, we need to define the similarity measure between the pages. The two similarity functions are:

- Bibliographic Coupling: For a pair of nodes p and q, the bibliographic coupling is equal to the number of nodes that have links from both p and q.
- Co-citation: For a pair of nodes p and q, the Co-citation number is the number of nodes that point to both p and q.

6.2 ANALYSIS OF GROUP OR COMMUNITIES ON SOCIAL NETWORKS

According to Mellah et. al [23] a community is a part of a graph where the nodes are strongly related together compared to the other nodes of the graph. The discovery of communities is an important problem in social network analysis, where the goal is to identify the groups (communities). Gibson et al [12] used hyperlink for identifying communities. Numerous approaches to detect communities were proposed in the past. Some of them are:

- PAGERANK,
- HITS
- PHITS
- SALSA

Apart from these, Dourisboure et al [24] identified a graph of web communities as dense subgraphs of web graph by using the heuristic algorithm on bipartite graph which has Authorities on one end and Hubs on the other.

7. CONCLUSION

In this paper we survey the research area of Web Mining emphasising on Web Structure Mining which is used for social network analysis. Initially, we have discussed Web Mining, followed by a discussion on Web Structure Mining leading to Link Analysis. As web mining is the use of data mining techniques to discover and extract information from the web documents and services, web structure tries to study the topology of hyperlinks while link analysis helps to reach the desired objective using an algorithm and a model. Amongst the various algorithms used for Link Mining, Social Network Analysis can be performed using two important algorithms PAGERANK and HITS. PAGERANK algorithms states that a page will have high page rank if thereare many pages that point to it, or if there are some high rank pages that point to it. HITS on the other hand rates web pages on the concept of authority and hub. It states that a page will have a high weight authority if it is pointed to by many document with high hub weight and vice versa, and a document will have a high hub weight if it points to many documents with high authority weight and vice versa.

Since web has a huge dataset and extends its support for Social Networks which in turn act as platform for existence of Online Communities, this paper provides the research community with great opportunities of research in the field of Social Network Analysis and Online Communities using graph theory and an extension of these algorithms. Some of the research areas for the same can be:

- Analysis of Different Groups (online communities) with similar interests on same Social Network (Facebook).
  Due to the fact that online communities consist of members from all age groups, a research study can be conducted to analyse the patterns and growth of such communities over Social Networks. This can be achieved using a link analysis algorithm along with heuristic approach.
- A comparative study of online Communities over various Social Networks (Twitter, Facebook)
  Since web provides a strong base for the existence of many online communities over different sites, a comparative study of usage for these online communities can be carried out in future research work.

The finding of this study will help many researchers in finding new undiscovered knowledge related to online communities existing over social networks.

Therefore, in this paper we discuss the role of data mining over the web, for studying various link structures with the help of algorithms based on graph theory. The structure of web can be represented as a collection of nodes and edges where nodes are the pages and the edges represents the link between the pages. Using this concept various web models have been developed like Bow-Tie Model, Maximal flow Model, Probabilistic Relational Model etc. These models serve as the base for generation of link analysis algorithm. The same knowledge can then be applied to analyse the existing online communities over various Social Networking sites.

REFERENCES

[1] J. Hui and M. Kamber, *Data Mining Concepts and Techniques*, USA: Morgen Kaufmann Publishers, 2001.

[2] M. Júnior and Z. Gong, "Web Structure Mining: An Introduction". *Proceedings of 2005 IEEE International*

Conference on Information Acquisition, China, June 27-July 2005.
[3] R. Kimball and M. Ross, *The data warehouse toolkit*,2nd edition.Wiley,2011
[4] O. Etzioni ,"World Wide Web: Quagmine or Gold mine?", *Communications of the ACM*, Jan. 1997
[5] S. Green, L. Hurst, B. Nangle, P. Cunningham, F. Somers and R. Evans, "Software Agents: A Review ". Technical Report TCD-CS-1997-06, University of Dublin, 1997.
[6] S. Chakrabarti, "Data Mining for Hypertext: A Tutorial Survey". *ACM SIGKDD Explorations*, 1(2) , Nov. 2000.
[7] R. Kosala and H. Blockeel , "Web Mining Research :A Survey". *ACM SIGKDD Explorations*, 2(1), July 2000.
[8] S. Madria, S. Bhowmick, W. Ng and E. Lim ,"Research Issues in Web Data mining". In *Proceedings of Data Warehousing and knowledge Discovery*, 1st International Conference, DaWak'99, pp. 303-312, 1999.
[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A Tomkins and J Wiener, Graph Structure in the Web. *Computer Networks, Proceedings of the 9th International World Wide Web Conference(WWW-9)*, 33(1-6), 2000.
[10] M. henziner ,*Hyperlink Analysis Technique* ,2009.
[11] L. Page, S. Brin , R. Motwani and T. Winograd, "The pagerank citation ranking: Bringing Order to the Web". Technical report, Stanford Digital Library Technological Project, 1998.
[12] D. Gibson, J. Kleinberg and P. Raghvan, "Inferring Web Communities from Link Topology". *HYPERTEXT'98: Proceedings of the ninth ACM conference on hypertext and hypermedia,ACM*, New York, 1998.
[13] S. Kleinberg et al, "Hypersearching the Web". Feature article, *Scientific American*, June 1999.
[14] R. Lempel and S. Moran, "The Stochastic Approach for Link Structure Analysis and the KTC effect". *Computer Networks*, 2000.
[15] A Borodin, G Roberts, J Rosenthal and P Tsaparas, "Finding authorities and hubs from link structures on the world wide web". *Proceedings of the international world wide web conference (WWW)*, 2001.
[16] D. Cohn and H. Chang, "Learning to Probabilistically identify Authorative Documents". *Proceedings of the 17th international conference on machine learning*, Morgan Kaufmann, San Francisco, 2000.
[17] T. Upstill, N. Craswell and D. Hawkins, "Predicting fame and fortune: PageRank or InDegree?". *Proceedings of the 8th Australasian document computing symposium*, Canberra, Dec. 2003.
[18] J. Furnkranz , "Web Structure Mining, Exploiting the Graph Structure of the World Wide Web".
[19] A. Pujari, *Data Mining Techniques*, 3rd edition. Hyderabad: Universities Press, 2013 .
[20] R. Botafogo, "Cluster Analysis for Hypertext Systems". *ACM-SIGIR*, 1993.
[21] H. Kautz, B. Selmen and M. Shah, "The Hidden Web". *AI Magazine*, 18(2):27-36, 1997.
[22] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment". *In Proceedings of ACM SIAM Symposium on Discrete Algorithms*, 1998.
[23] M Mellah et. al,"Link Analysis for Communities Detection on Facebook". *International Journal of data mining and emerging technologies*, vol. 4, no 1, May 2014.
[24] Y. Dourisbourne, F. Gerarci and M. Pellegrini, "Extraction and classification of dense communities in the web". *Proceedings of the 16th International Conference on WWW, ACM*, New York, 2007.
[25] http://tutorialspoint.com/datamining
[26] Sudha and Rani ,*Applications of Data Mining*. New Delhi: Discovery Publications,2008.
[27] L. Getoor, "Link Mining: A new Data Mining challenge". *SIGKDD Explorations*, vol. 4, issue 2, 2003.
[28] A Fahmideh, R Baettela and S Asadpoor ,http://www.slideshare.net/AmirFahmideh/web-mining-structure-mining, Z. Yoseef and S. Rajagopalan, "Template Detection via Data Mining and its Application". *Proceedings of the 11th international Conference on World Wide Web, USA: ACM 2002.*
[29] *P. Devi, A. Gupta and A. Dixit; "Comparitive study* of HITS and PAGERANK Link based Ranking Algorithms". *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, Issue 2, February 2014.
[30] H. Lee and A. Borodin; "Perturbation of the hyper-linked environment". *Proceedings of 9th Annual International Conference, COCOON 2003 Big Sky*, USA: Springer July 25-28 2003.
[31] P. Saxena, J. Gupta and N Gupta; "Web Page Ranking Based On Text Content Of Linked Pages". *International Journal of Computer Theory and Engineering*, Vol. 2, Issue 1, February 2010.
[32] D. Sharma and A. Sharma; "A Comparative Analysis of Web Page Ranking Algorithms". *International Journal on Computer Science and Engineering*, Vol. 02, Issue No. 08, 2010.
[33] A. Borodin et al ;"Link Analysis Ranking: Algorithms, Theory, and Experiments". ACM Transactions on Internet Technology, Vol 5, Issue 1,2005.
[34] W. Xing and A.Ghorbani; "Weighted PageRank Algorithm". *Communication Networks and Services Research, Second Annual Conference, IEEE* May 2004
[35] P. Desikan, J. Srivastava, V. Kumar, and P. Tan," Hyperlink Analysis: Techniques and Applications". University of Minnesota
[36] R. Elmasri and S. Navathe; "Database Systems: Models, Languages, Design and Application Programming" , 6th Edition, Pearson, Delhi.

Author Profile

Manjula Gulhati Verma is a student of Assam Don Bosco University, Guwahati, She is currently pursuing PhD in the field of Computer Science and Technology. Along with this, she is also working at Don Bosco College, Panjim, Goa as Assistant Professor in the department of Computer Applications. She received degree of Masters in Information Technology from University of Rajasthan followed by Diploma in Advanced Computing and Diploma in Information System Audit.

Jayanta Yumnam is working as Associate Professor and Head of Dept of Computer Science & Engineering and IT, School of Technology, Assam Don Bosco University, Guwahati,. He completed his Ph.D. in Computer Science and Technology from Dr. B. Ambedkar Marathwada University, India in the year 2004 . He has a rich experience of 14 years in industry & academia along with this, he has published many research papers in this area. He also holds prestigious memberships of IJET, IACSIT, IETE, IASTED, EUROSIS and IAENG