

# AN UPGRADED FOCUS ON LINK ANALYSIS ISSUES IN WEB STRUCTURE MINING

Dr K. Vivekanandan<sup>1</sup>, A. Pankaj Moses Monickaraj<sup>2\*</sup>, K. Prabhu<sup>3</sup>

---

<sup>1</sup>Professor, BSMED,  
Bharathiar University, Coimbatore-46.

<sup>2</sup>Doctoral Scholar, Department of  
Computer Science, Bharathiar  
University, Coimbatore -46.

**Email Id:**

[pankajmoses@hotmail.com](mailto:pankajmoses@hotmail.com)

## **Abstract**

Web mining branches to three Text, Structure and Usage Mining. Structure Mining focuses on the arrangement of web pages and effective retrieval of them. Each and every page in web are inter connected through web links either pointing to the same site or pointing to another side. This can be studied through various researches in this area which helps to personalize user, retrieve page and accurately pick pages for users. This area has their co branches in Opinion Mining and mostly the concept of graph theory plays a vital role in the algorithms. This paper focuses research issues in the area of Link analysis.

**Keywords-** Web Mining, Link Analysis, Graph Mining, Opinion Mining.



**Pubicon**

## Introduction

Early search engines retrieved pages for the user based primarily on the user query and the indexed pages of the search engines. The retrieval ranking and ranking algorithm were simply direct implementations of those from information retrieval. This can no longer be sufficient due to two reasons. First, Content Similarity methods are easily spammed. A page owner can repeat some important words and remotely related words in their pages to boost the ranking.<sup>1</sup> Second, the number of pages grew rapidly so the relevant pages were huge. The abundance of information causes a major problem for ranking.

Researchers began to work on these problems. Web pages are connected through hyperlinks, which carry important information. Few hyperlinks are used to organize large amount of information at the same website.<sup>4</sup> Thus, some links point to pages in the same site and others point to other website. Such outgoing links often indicate an implicit conveyance of authority to the pages being pointed to.

Through the overview of the hyperlinks connectivity in the web, led them to face few core areas of research in Link Analysis.

1. Social Network Analysis
2. Co-Citation and Bibliographic coupling
3. Page Rank
4. HITS
5. Community Discovery

## HYPERLINK CONNECTIVITY IN WEB

### Social Network Analysis

Social Network is the study of people in an organization called actors, their interactions and relationship (study of social entities).<sup>2</sup> These can be represented with a network or graph where each vertex (node) represents actor and each link represents a relationship. We can also find various kinds of sub-graph (communities).

This Social network Analysis is of two types' centrality and prestige, which has its impact on hyperlink analysis and search of the web. Both measure the degree of prominence of an actor in a network.

### Centrality

A central actor is one involved in many ties (link). Important or Prominent actors are those that are linked with other actors extensively. There are several types of links that causes several types of centrality.

Degree centrality – Central actors are the most active actors that have most links with other actors. Consider degree of an actor “i”, is simply the node degree (the number of edges) of the node.

Closeness centrality – This view is based on the closeness or distance (i.e.) the particular actor is central if it can easily interact with all other actions. Through the distance we can compute this measure.

Betweenness centrality – If two non-adjacent actors “j” and “k” want to interact and actor “i” is on the path between j and k, then “i” may have some control over their interactions.

### Prestige

Here first there is distinguishing between ties sent and ties received.<sup>3</sup> To compute the prestige of a node, the ties directed to the actor is focused. They are of some its kind, they are

Degree Prestige-The actor is prestigious if it receives many in- links. The value ranges from 0 and 1.

Proximity Prestige- The degree of an actor “i” only considers the actors that are adjacent to “i”. It defines as closeness of other actor “i”.

Rank prestige - A company CEO voting for a person is much more important than a worker voting for the person. Thus one’s prestige is affected by the ranks of involved actors.

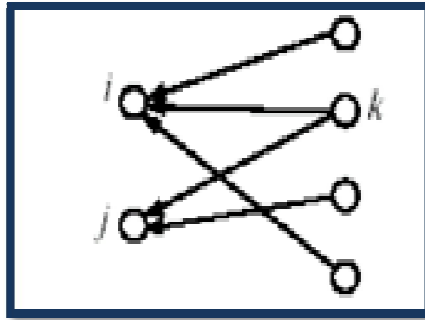
### Co-Citation and Bibliographic coupling

Citation analysis is an area of bibliometric research, which studies citation to establish the relationships between authors and their work. When publication cites another publication, a relationship is established between the publications. Citation analysis uses these links to performance various types of analysis.

This area focuses on two types’ co-citation and bibliographic coupling.

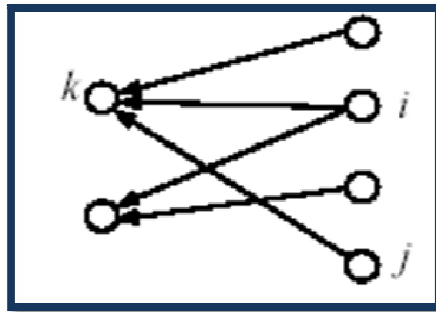
### Co-citation

Co-citation is used to measure the similarity of two documents.<sup>5</sup> If papers i and j are both cited by paper k, then they may be related in some sense to one another. If papers i and j are both are cited together by many papers, it means that i and j have strong relationship or similarity.



### Bibliographic coupling

This operates on a similar principle, but in a way it is the mirror image of co-citation.<sup>7,10</sup> Bibliographic coupling links papers that cite the same articles so that if both papers  $i$  and  $j$  both cite paper  $k$ , they may be said to be related, even though they do not directly cite each other.



### Page Rank

Page rank was presented by surgery Brin and Larry Page at Seventh International World Wide Web Conference (WWW7) in April, 1998.<sup>7</sup> It relies on the democratic nature of the Web by using its vast link structure as an indicator of an individual page's PageRank relies on the democratic nature of the Web by using its vast link structure as an indicator of an individual page's value or quality PageRank interprets a hyperlink from page  $x$  to page  $y$  as a vote, by page  $x$ , for page  $y$ .<sup>11</sup>

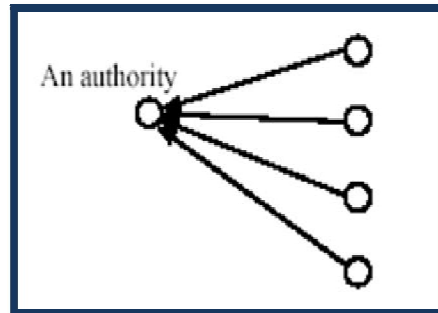
However, Page Rank looks at more than the sheer number of votes; it also analyzes the page that casts the vote. Votes casted by "important" pages weigh more heavily and help to make other pages more "important." This is exactly the idea of rank prestige.

A hyperlink from a page to another page is an implicit conveyance of authority to the target page.<sup>6</sup> The more in- links that a page  $i$  receives, the more prestige the page  $i$  has. Pages that point to page  $i$  also have their own prestige scores. A page of a higher prestige pointing to  $i$  is more important than a page of a lower prestige pointing to  $i$ . In other words, a page is important if it is pointed to by other important pages.

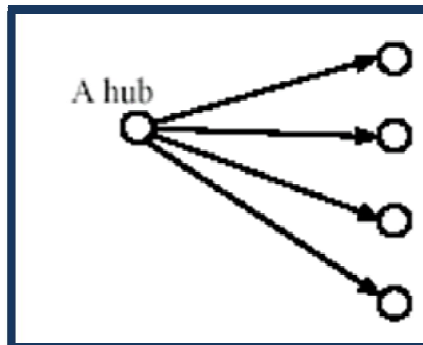
## HITS

HITS stands for Hypertext Induced Topic Search is query dependent. HITS first expands the list of relevant pages returned by a search engine and then produces two ranking of the expanded set of pages, authority ranking and hub ranking.

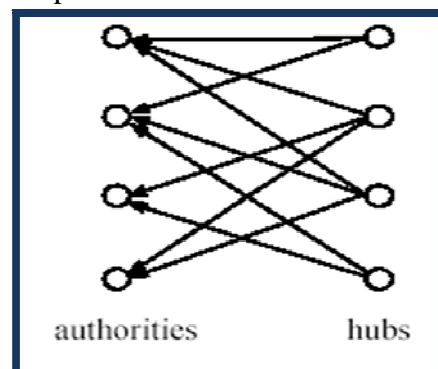
Authority Ranking - An authority is a page with many in- links. The view is that the page may have good or authoritative content on some topic and thus many people trust it and link to it.



Hub Ranking – A hub is a page with many out-links.[8] The page serves as an organizer of the information on a particular topic and points to many good authority pages.



The Core idea of HITS is that a good hub points many good authorities and a good authority is pointed to by many good hubs. Thus, authorities and hubs have a mutual relationship.



## Community Discovery

A Community is a group of entities that shares a common interest or is involved in an activity or event. The communities are represented by dense bipartite sub graphs.

Given a finite set of entities  $S=\{s_1,s_2,s_3,\dots,s_n\}$  of the same type, a community is a pair  $C=(T,G)$ , where  $T$  is the community theme and  $G$  belongs to  $S$  is the set of all entities in  $S$  that shares the theme  $T$ . [8] If  $s_i$  belongs to  $G$ ,  $s_i$  is said to be a member of the community  $C$ .

Communities have Hierarchical Structures namely: Sub- Community, Super-Community, and Sub-theme.

### Example:

A community  $(T,G)$  may have a set of sub-communities

$\{(T_1,G_1),\dots,(T_m,G_m)\}$ , where  $T_i$  is a sub-theme of  $T$  and  $G_i$  belong to  $G$ .  $(T,G)$  is also called Super community of  $(T_i,G_i)$ . In the same way each community can be further decomposed to give Community Hierarchy.

Community Manifestation is usually done through dataset which can be set of web pages, a collection of emails or set of text documents.

Web Pages – Hyperlinks, Content words

Emails – Exchange between entities, exchange words. Text Document – Co-occurrence

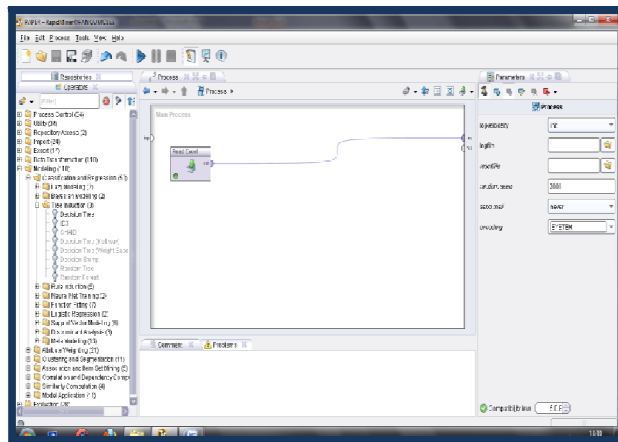
Overlapping of Communities comprises of Building a link graph, finding all triangles, finding community cores, Clustering around cores.

## IMPLEMENTATION

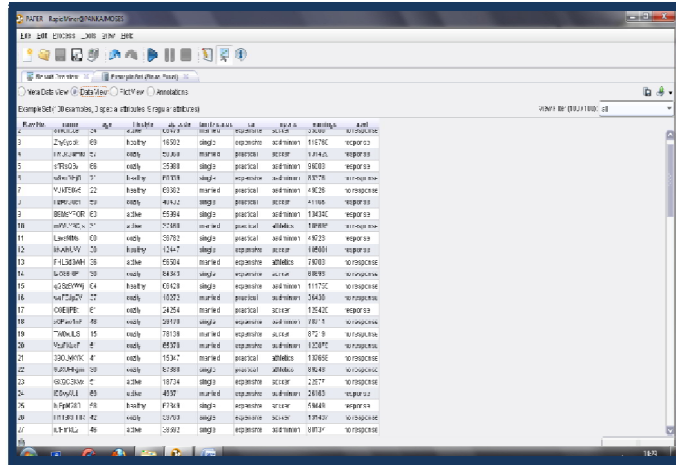
Implementation provide a complete overview about graph linking, mailing churn dataset with 11 columns and 101 rows, records of 1111 and attributes of name, age, lifestyle, zip code, family status, car, sports, earning, label and implementation done in Rapid Miner.

To start with a decision tree, it is categorised into response and no response which protruded into sports followed by what kind of sports they play.

Importing Dataset:

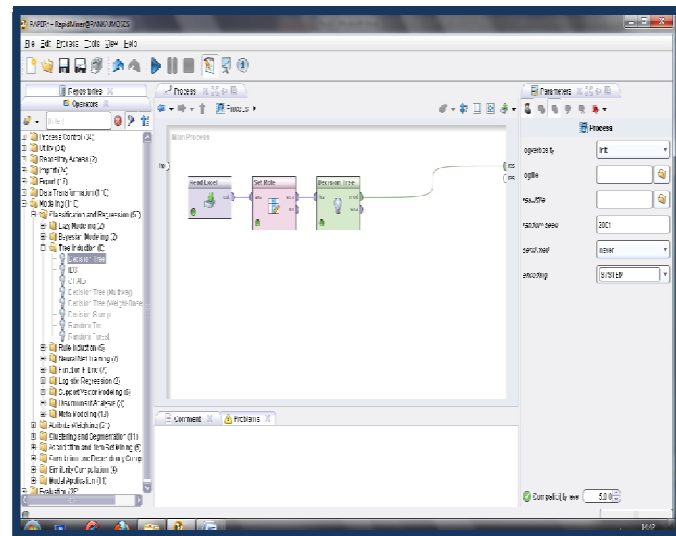


### Data View:

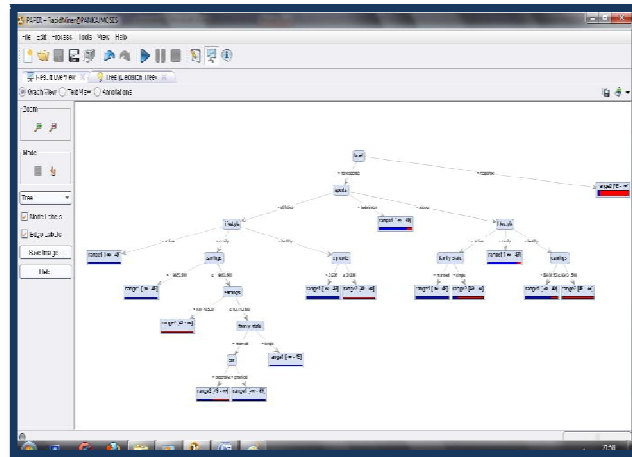


Number	Name	Age	Gender	Marital Status	Occupation	Education	Income	City
1	Abraham	42	male	married	engineer	graduate	15700	newyork
2	Zeynep	69	female	single	doctor	postgraduate	11790	newyork
3	Melissa	51	female	married	teacher	graduate	13400	newyork
4	Yehuda	65	male	single	engineer	graduate	9600	newyork
5	John	37	male	single	engineer	postgraduate	10700	newyork
6	Vijaya	22	female	married	engineer	postgraduate	4320	newyork
7	Joseph	59	male	single	engineer	graduate	11100	newyork
8	Robert	63	male	single	engineer	graduate	13600	newyork
9	William	71	male	married	engineer	postgraduate	10700	newyork
10	Lawrence	60	male	single	engineer	graduate	4320	newyork
11	Isabella	25	female	single	engineer	graduate	16600	newyork
12	Fredrick	55	male	married	engineer	graduate	7800	newyork
13	Robert	50	male	single	engineer	graduate	10800	newyork
14	Joseph	64	male	single	engineer	postgraduate	11700	newyork
15	William	57	male	married	engineer	postgraduate	5640	newyork
16	Joseph	61	male	married	engineer	graduate	12400	newyork
17	Joseph	68	male	single	engineer	postgraduate	7800	newyork
18	William	15	male	married	engineer	graduate	870	newyork
19	William	4	male	married	engineer	postgraduate	13200	newyork
20	Joseph	47	male	married	engineer	graduate	12900	newyork
21	William	68	male	single	engineer	graduate	10400	newyork
22	Joseph	4	male	single	engineer	graduate	2570	newyork
23	Joseph	65	male	single	engineer	postgraduate	2680	newyork
24	William	68	male	single	engineer	graduate	4840	newyork
25	William	42	male	single	engineer	graduate	9200	newyork
26	William	46	male	single	engineer	postgraduate	9700	newyork

### Decision Making:



### Obtained Tree:



## CONCLUSION

The latest trends in Link Analysis include the above 5 areas of which the key base core starts from mathematics area of graph theory which is revised as Graph mining. In this paper, Social Network Analysis, Co-citation, PAGERANK, HITS, Community Discovery are given as a view and their emergence from Graph Theory Mathematics (Decision tree) is implemented. Implementation supports the classification through graph, along with the pictorial view.

These are the more recent current research issues involved in Link analysis.

## References

1. I.G Webb Discovering Association with Numerical Variables. In Proc. of the SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.(KDD,01), pp:383-388,2001
2. C.L.Borgam, (ed.) Scholarly Communication and Bibliometrics. Sage Publications, Inc., July 1, 2001 71: 79-91
3. L.Page, S. Brin , R. Motwami, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report,Computer Science department, Stanford University. January 29, 1998
4. K.Bharat and M. Henzinger. Improved Algorithms for Topic Distillation in Hyperlinked Environments, .In Proc of Conf. on Research and Development, ACM New York, NY, USA, 1998
5. R. Weiss, B. Velez, M. Sheldon, C. Nemprenpre, P. Szilagyai and D.K. Giffor, HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering, in: Proc. of the 7th ACM Conference on Hypertext, 1996.
6. T. Haveliwala Extrapolation Methods for Accelerating PageRank Computations. In. Proc. 16th World Wide web Conf, WWW2003, May 20–24, 2003, Budapest, Hungary.ACM 1581136803/03/0005
7. S.Fortunato, A.Flammini , and F. Menczer. Scale-Free Network Growth by Ranking. *Phys. Rev. Lett.* APS Journals, Volume 96 , Issue 21, February 2006; published 31 May 2006.
8. Ntoulas, J.Cho, and C. Olston. Whats is new on web? The Evaluation of the web from a Serach Engine Analysis Perspective. In. Proc. of the World Wide web conference. WWW2004, May 17–22, 2004, ACM New York, NY USA.
9. M. Toyoda and M. Kitsuregawa, “Extracting Evolution of Web Communities from a Series of Web Archives,” Proc. of the 14th ACM Conf. on Hypertext and Hypermedia (Hypertext 03), 2003.



10. S.Pandey, S. Roy, C. Olston. Shuffling a stacked Deck: The case for Partially Randomized Ranking of Search Engine Results. In. Proc of Very Large Database. VLDB '05 Proceedings of the 31st international conference on Very large data bases, Pages 781 – 792,2005
11. X.Li, B.Liu and P.S. Yu. Time Sensitive Ranking with Application to Publication Search. *icdm*, pp.893-898, 2008 Eighth IEEE International Conference on Data Mining, 2008
12. V.Vapnik. The Nature of Statistical Learning Theory. JohnWiley, 2nd ed. 2000, XIX, 314 p.
13. B. Mobasher and S.S. Anand (Eds.): Intelligent Techniques for web Personalization. In Intelligent Techniques for Web Personalization ITWP 2003, LNAI 3169, pp. 1–36, 2005. Springer-Verlag Berlin Heidelberg 2005.
14. D.A Grossman, and O. Frieder. Information Retrieval: Algorithms and Heuristics, Springer, 20-Dec-2004.
15. T.Mitchell. Machine Learning. Mc.Graw Hill. 1997.
16. O.Parr Rud. Data Mining Cookbook. JohnWiley & Sons. Edition 1 , November 3, 2000