2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation

Knowledge Discovery and Retrieval on World Wide Web Using Web Structure Mining

Sekhar Babu Boddu[1]
Asst.Professor
sekhar_9866@yahoo.co.in

V.P Krishna Anne[1]
Asst.Professor
krishnavpraveen@gmail.com

Rajesekhara Rao Kurra[2]
Professor
krr_it@yahoo.co.in

Durgesh Kumar Mishra[3]
Professor
durgeshmishra@ieee.org

[1]Department of Master of Computer Applications,
[2,3]Department of Computer Science and Engineering

[1,2]KLEF University, Vaddeswaram, Guntur, Andhra Pradesh, India
[3]Acropolis Institute of Technology & Research, Indore, MP, India

***Abstract:* The World Wide Web is nearing omnipresence. The explosively growing number of Web contents including Digitalized manuals, emails pictures, multimedia, and Web services require a distinct and elaborate structural framework that can provide a navigational surrogate for clients as well as for servers. Due to the increasing amount of data Available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing. In this paper, we provide an introduction of Web mining as well as a review of the Web mining categories. Then we focus on one of these categories: the Web structure mining. Within this category, we introduce link mining and review two popular methods applied in Web structure mining: HITS and Page Rank.**

***Key Words*– Web structure mining, Web mining, link mining.**

I INTRODUCTION

The World Wide Web is a collection of Web sites and its Web contents. The Web evolves continuously and Changes dynamically since new Web sites are born and the old ones disappear simultaneously, and contents of those Web sites are updated at any times. While the Web contains vast amount of information and provides an access to it at any places and any times, that is a prize beyond our reach without efficient searching tools for the Web. Efficient searching for Web contents becomes more important than ever before as the Web evolves and users increase explosively the World Wide Web has becoming one of the most comprehensive information resources. It probably, if not always, covers the information need for any user. However, the Web demonstrates many radical differences to traditional information containers such as databases, in schema, volume, and topic-coherence. Those differences make it challenging to fully use Web information in an effective and efficient manner. Web mining is right for this need [1].In fact, Web mining can be considered as the applications of the general data mining techniques to the Web. However, the intrinsic properties of the Web make us have to tailor and extend the traditional methodologies considerably. Firstly, even though Web contains huge volume of data, it is distributed on the internet. Before mining, we need to gather the Web document together. Secondly, Web pages are semi-structured, in order for easy processing documents should be extracted and represented into some format. Thirdly, Web information tends to be of diversity in meaning, training or testing data set should be large enough. Even though the difficulties above, the Web also provides other ways to support mining, for example, the links among Web pages are important resource to be used Besides the challenge to find relevant information, users could also find other difficulties when interacting with the Web such as the degree of quality of the information found, the creation of new knowledge out of the information available on the Web, personalization of the information found and learning about other users. Web mining techniques could be applied to solve, partially or completely, the above cited problems. However, Web mining techniques are not the only tools to solve those Problems. Other research communities such as database, machine learning and information retrieval, are also addressing the above mentioned difficulties. This situation creates confusion to determine what forms Web mining. This paper is structured as fallows:

In section II we provide an overview of web mining categories. In section III we go through web structure mining and introduce link mining. In section IV we review two well known algorithms: HITS and PageRank. In section V we discuss preprocessing. In section VI we look on pattern analysis. The statistics collected from the website that can be depend up on various factors to find the knowledge discovery is discussed in section VII. We address about the work on webpage segmentation in section VIII. Finally, conclusion of our paper in section IX

II. WEB MINING OVERVIEW

To clarify the confusion to determine what forms Web mining. Kosala and Blockeel [2] had suggested a decomposition of Web mining in the following tasks:

1. *Resource finding***:** the task of retrieving intended Web documents.
2. *Information selection and pre-processing:* automatically selecting and pre-processing specific information from retrieved Web resources.

978-0-7695-4062-7/10 $26.00 © 2010 IEEE
DOI 10.1109/AMS.2010.108

IEEE computer society

Downloaded from http://www.elearnica.ir

3. *Generalization:* automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. *Analysis:* validation and/or interpretation of the mined patterns. In general, Web mining tasks can be classified into three categories: *Web content mining, Web structure mining and Web usage mining.* However, there are two other different approaches to categorize Web mining. In both, the categories are reduced from three to two: Web content *mining and Web usage mining*. In one, Web structure is treated as part of Web Content [11]; while in the other, Web usage is treated as part of Web Structure [12]. All of the three categories focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the Web. Each of them focuses on different mining objects of the Web. Fig. shows the Web categories and their objects. As follows, we provide a brief introduction about each of the categories. ***Web content mining*** targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories [13]:
1. *Intelligent Search Agents*. These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information
2. *Information Filtering/ Categorization*. These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
3. *Personalized Web Agents***.** These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest. The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are *multilevel databases* and *Web query systems*. For further information about Web content mining please refer to [2; 5; 12]. ***Web structure mining*** focuses on the hyperlink structure of the Web. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models [8]. Two algorithms that have been proposed to lead with those potential correlations: HITS [14] and Page Rank [10], and Web structure mining itself will be discussed in the next section. ***Web usage mining*** focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining collects the data from Web log records to discover user access patterns of Web pages. There are several available research projects and commercial that analyzes those patterns for different purposes. The generated from this analysis can be classified as system improvement, site modification, business intelligence and usage characterization [3]. The challenges involved in web usage mining could be divided in three phases [11]:
1. *Pre-processing.* The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. It includes data cleaning, data integration, data transformation and data reduction.
2. *Pattern discovery.* Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns.
3. *Pattern Analysis.* This process targets to understand, visualize and give interpretation to these patterns. . Web usage mining depends on the collaboration of the user to allow the access of the Web log records. Due to this dependence, privacy is becoming a new issue to Web usage mining, since users should be made aware about privacy policies before they make the decision to reveal their Personal data. For further information about Web usage mining please refer to [3; 11; 13]. We should note that there is no clear boundary between the above categories. As we mentioned, the two or three category definitions are quite acceptable, showing that Web content mining, Web structure mining and Web usage mining could be used isolated or combine in an application.

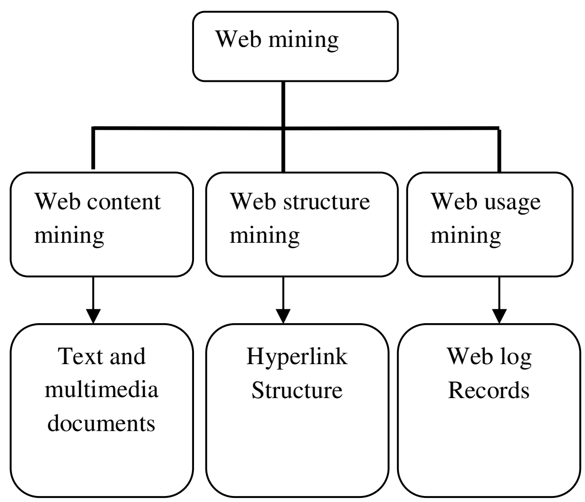


Fig. 1 Web mining categories and objects.

III. WEB STRUCTURE MINING

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [8], which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts [8]. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of link mining which are applicable in Web structure mining.

1. *Link-based Classification*. Link-based classification is the most recent upgrade of a classic data mining task to linked domains [7]. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the webpage.

2. *Link-based Cluster Analysis*. The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. *Link Type*. There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

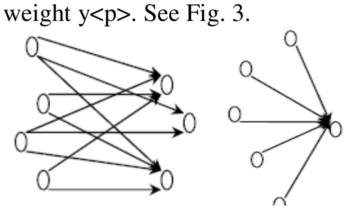
4. *LinkStrength*. Links could be associated with weights.

5. *Link Cardinality*. The main task here is to predict the number of links between objects. There are many ways to use the link structure of the Web to create notions of authority. The main goal in developing applications for link mining is to made good use of the understanding of these intrinsic social organization of the Web

IV. HITS CONCEPT AND PAGERANK METHOD

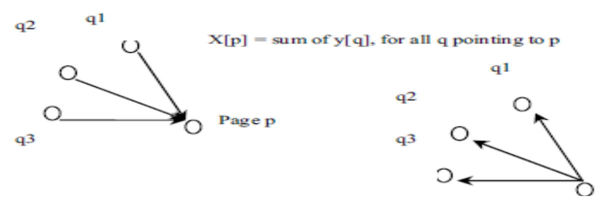
In this section we review two approaches: HITS concept and PageRank method. Both approaches focus on the link structure of the Web to find the importance of the Web pages.

A. HITS: Computing Hubs and Authorities In HITS concept, Kleinberg [14] identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to Kleinberg [14], "Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority *is* page that is pointed to by many good hubs". See Fig. 2. HITS associates a non-negative authority weight x<p> and a non-negative hub weight y<p>. See Fig. 3.



Hubs, Authorities Unrelated page of large in degrees

Fig: 2 A densely linked set of Hubs and Authorities



Y[p] =sum of X[p],for all q point to by p Page

Fig:3 Basic operations of HITS from [14]

According to Kleinberg [14], "Numerically the reinforcing relationship can be expressed as follows: if p points to many pages with large x-values, then it should receive a large y-value; if p is pointed to by many pages with large y-values, then it should receive a large x-value. Given weights x<p>,y<p>, then the x-weights and y-value are as follows see Figure 5

$$X^{<p>} \longleftarrow \sum_{q:\,(q,\,p)\in E} Y^{<q>} \qquad Y^{<q>} \longleftarrow \sum_{q:\,(q,\,p)\in E} X^{<p>}$$

Fig: 4 X weighted and Y weighted from [14]

Although HITS provides good search results for a wide range of queries, HITS did not work well in all cases due to the following three reasons [5]:

1. Mutually reinforced relationships between hosts. Sometimes a set of documents on one host point to a single document on a second host, or sometimes a single document on one host point to a set of document on a second host. These situations could provide wrong definitions about a good hub or a good authority.

2. Automatically generated links. Web document generated by tools often have links that were inserted by the tool.

3. Non-relevant nodes. Sometimes pages point to other Pages with no relevance to the query topic

B. PageRank Model

L. Page and S. Brin [10;15] proposed the Page Rank algorithm to calculates the importance of web pages using the link structure of the web. In their approach Brin and Page extends the idea of simply counting in-links equally, by normalizing by the number of links on a page. The Page Rank algorithm is defined as [15]: "*We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is adamping factor, which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C (A) is defined as* *the number of links going out of page A. The Page Rank of a page A is given as follows:*

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn)) (1)

Note that the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one." And "*The d damping factor is the probability at each page the "random surfer" will get bored and request another random page."* Note that the rank of a page is divided evenly among its out-links to contribute to the ranks of the pages they point to. The equation is recursive, but starting with any set of ranks and iterating the computation until it converges may compute it. Page Rank can be calculated using a simple iterative algorithm, and corresponds to the principal eigen vector of the normalized link matrix of the web. Page Rank algorithm needs a few hours to calculate the rank of millions of pages [15].

C. Applications

HITS was used for the first time in the Clever [17] search engine from IBM, and PageRank is used by Google [18] combined with other several features such as anchor text, IR measures, and proximity. The notion of authoritativeness comes from the idea that we wish not only to locate a set of relevant pages, but rather the relevant pages of the highest quality. However, the Web consists not only of pages but also of links that connect one page to another. This structure contains a large amount of information that should be exploited. PageRank and HITS belong to a class of ranking algorithms, where the scores can be computed as a fixed point of a linear equation. Bianchini [16] noted that HITS and PageRank are used as starting points for new solutions, and there are some extensions of theses two approaches. There are other link-based approaches to be applied on the Web. For further information please refer to [3, 5, 9, 13, 14, 15, 16]. Beside being used for weighting Web pages, link resource can also be used for clustering or classifying Web pages. The principle is based on the assumption that (1) if page *p1* has a link to page *p2*, *p1* should be similar to *p2* in content, and (2) if p1 and p2 are co-cited by some common pages, *p1* and *p2* should also similar. Web pages can be clustered into a lot of connected page communities with respect to their citation and co-citation strengths among the pages. In fact, Ziv Bar-Yossef and Sridhar Rajagopalan [19] put all the algorithms, which uses links, to three categories.

1. Relevant Linkage Principle: Links points to relevant resources.
2. Topical Unity Principle: Documents often co-cited are related, as are those with extensive bibliographic overlap. This idea is previous addressed by Kesseler for bibliographic information retrieval in [20].
3. Lexical Affinity Principle: Proximity of text and links within a page is a measure of the relevance of one to another. Even though those link algorithms can always provide a good support for Web information retrievals, clustering and knowledge discoveries on the Web, authors also find problems associated with those technologies [19, 21, 22,23]. Taher H. Haveliwala notices that original PangeRank algorithm, introduced by Page et al. [10], pre-compute ranking vector based on all the Web pages. This ranking vector is computed once and used for any queries later. The ranking is actually independent of the specific queries when using it. The authors tried to solve this limitation by computing a set of PageRank vectors, each biased with a different topic. In other words, for each topic, they assigned a weight for each page. Therefore, searches in different topics could select corresponding vectors for ranking. Ziv Bar-Yossef and Srihar Rajagopalan [19], Deng Cai et al. [22], and shaian-Hua Lin et al.[23] addressed problems caused by topic drift of Web pages. In other words, it is easy to find a Web page with multiple topics. In such cases, two links from the same Web page may lead to different semantics if they are anchored in the areas of different topics. Or maybe some links only links to some advertisements. One effective solution for topic drift problems is page segmentation.

Therefore, ranking based on the information content can be improved [22, 23]. Furthermore, Page Rank, HITS, and other link-based algorithms can be applied to page blocks. The basic ideas are: (1) Links from important blocks are assigned higher weights, [23] (2) a block (other than page) is semantically similar to a page if there is a link anchored in this block to the page, (3) two pages are similar if there are co-cited via some common Deng Cai et al .[29] defined the importance factors of a block with respect to the size of the block and its position in the screen when browsing. Their experimental result reveals that block-level PageRank and HITS can improve retrieval performance significantly.

V. PREPROCESSING

Real-world data tend to be dirty, incomplete and inconsistent. Data preprocessing techniques can improve the quality of the data, there by helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, since quality decisions must be based on quality data. Data used in preprocessing cover server log files, web page content, web site structure and hit counts of pages in the web site. Figure 5 shows model for amalgamation of web usage and web structure mining. Data cleaning removes entries unhelpful to data analyzing and mining. It has to remove log entries that have status code as "failure" or "error". Secondly some automatic search engines generate some access records, those have to be identified and removed from the log file. Some of other common indicators such as (a) the repeated request for the same URL from the same host;(b) a time interval between requests too short to apprehend the contents of a page; and (c) a series of requests from one host all of whose referrer URLs are empty.The dynamic behavior of web users under a particular session can be any of the following These behaviors can be used to construct more complex navigation behaviors in a single session. These four basic behaviors constructing complex navigations are given below: 1. A Web user can start session with any one of the possible entry pages of a web site. This behavior includes new page which is not requested by any other previous page accessed from the same domain in near-time 2. A Web user can select the next page having a link from the most recently accessed page.

3. A Web user can press the back button one more time and thus selects as the next page a page having a link from any one of the previously browsed pages (i.e., pages accessed before the most recently accessed one).

4. A Web user can terminate his/her session.

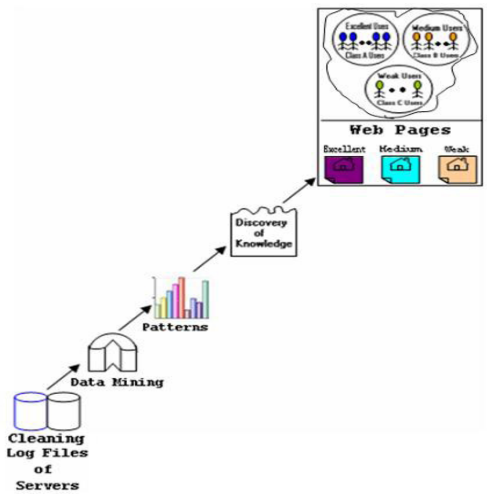


Fig 5 Amalgamation of web usage and web structure mining [30]

VI. PATTERN ANALYSIS

Challenge of pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. Visualization assists an analyst to better apprehend navigation patterns and to predicate trends of data. Knowledge about content and structure also contribute to filtering un-useful knowledge. Many web tools provide some objective criteria, supporting and confidence. Such criteria are helpful to manually filter some believed unimportant knowledge. WebViz tool has done some pioneering work in visualizing of access patterns. It displays access pattern of user as directed graph, with nodes representing page of the access pattern and links representing the hyperlinks between pages. Web pages in the web site can be classified in to Excellent –the web pages with highest hit counts. Medium - the web pages with average hit counts. Weak --- the web pages with least hit count. Web users who have visited web sites can be classified as Class A users (Excellent), Class B users (Medium) and Class C users (Weak users).

VII. KNOWLEDGE DISCOVERY

The statistics collected from the web site can help discovering the knowledge. This knowledge collected can be used to take decision on various factors like 1. The web pages with highest hit counts will be the popular pages. 2. What is possible the navigation patterns of users. 3. The time spent on each web page which tells about importance of the web page.4. If time spent on particular web page is negligible it indicates that the web page does not contain important information. 5. The web pages for which no user's request is there, indicates that page must be modified. 6. If log file entry says repeatedly for particular web page "redirect", it should be notified to web site designer/owner. After the pattern analysis is done on web pages, the important decision can be done regarding structure of the website. The Excellent web pages will be moved very near to the home page, at next level medium class web pages moved and so on… The pages with more hit count can be given the preference to be brought closer to the home page provided web site owner/designer agrees. The heap tree can be generated based on hit counts available in the log file during particular session. This heap tree generated will help us make decision about topology of web site during next interval so that the web pages which are more popular can be brought very near to the home/parent web page. With this restructuring, the web users can gain quick access to the web pages along with best utilization of bandwidth and server's memory space since every HTTP request will be entered into the log file of the server.

VIII WEB PAGE SEGMENTATION

In this section, we will discuss some works on Web page segmentations. Earlier works on document segmentation were on free texts and motivated by raising performances of the information retrieval. In information retrieval, documents are ranked with the values of similarities of the documents to the queries. One popularly used similarity model is calculated using cosine between vector of query terms and vector of document terms, as

$$\mathrm{sim}(q, d) = \frac{\sum W_{q,i} * W_{q,i}}{(\sum W^2_{q,j})^{1/2} (\sum W^2_{d,j})^{1/2}}$$

where $W_{q,i}$ and $W_{d,i}$ are the weights of term *ti* in query q and document d respectively. From this metric function, it is clear that mutual affection of several topics in the same document may cause lower ranking of the document even it may contains a passage which can well covers the user's need. For this reason, documents are partitioned into a sequence of passages with respect to topics. Text segmentation algorithms in general fall into three categories: The first one is to identify the locations of topic changes for text stream. Texts created from automatic speech recognition, newswire feeds, or television closed transcripts may contain cue-words as topic transitions [24]. However, general texts may not contain noticeable topic change words. The second method is to partition texts with fixed-length windows. Though it is simple, such method improves retrieval performance effectively [25]. But it can not be coupled with topics in the document. The third method is based on semantic coherence of the text [26,27,28] Such kind of methods partition texts by measuring semantic coherence between two consecutive blocks which are fine-grained text units such as sentences or even words. Web documents are more likely to contain multiple topics than traditional free texts. Helpfully, more separators (different HTML tags) can be used in segmentations of the Web documents. Once Web pages are partitioned into blocks, Web page ranking and indexing can be performed on the granularity of semantic blocks other than on the whole pages. Therefore, ranking based on the information content can be improved [22, 23]. Furthermore, Page Rank, HITS, and other link-based algorithms can be applied to page blocks. The basic ideas are: (1) Links from important blocks are assigned higher weights, [23] (2) a block (other than page) is semantically similar to a page if there is a link anchored in this block to the page, (3) two pages are similar if there are co-cited via some common Deng Cai et al .[29] defined the importance factors of a block with respect to the size of the block and its position in the screen when browsing. Their experimental result reveals that block-level PageRank and HITS can improve retrieval performance significantly.

IX. CONCLUSION

In this paper, we survey the research area of Web mining, focusing on the knowledge issues on web using web structure mining and introduced the Link mining, as well as block-level link mining. For this we had also reviewed two popular algorithms HITS and PageRank. These both methods focus on the link structure of the web and calculate the importance of webpage. We explain about data preprocessing is an important step in the knowledge discovery process. Preprocessing is internally depending up on those two algorithms, for which to know about website structure and hit count of the page in website. This method is most useful to find knowledge discovery. Another concept pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users, webpage in the website classified into different classes with respect to hit count of the webpage. This is also an important method in the process of knowledge discovery and retrieval on web. Web Structure Mining plays a vital role with various benefits including quick response to the web users, reducing lot of HTTP transactions between users and server. We hope the paper could provide useful information related to knowledge issues on web using different methodologies and this area as useful starting point for identifying opportunities for further research.

REFERENCES

[1] O. Etzioni. The World Wide Web: Quagmire or gold mine. *Communications of the ACM*, 39(11):65-68, 1996.

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

[3]Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan,Web Usage Mining: Discovery and Applications of Usage Patterns from WebData, *ACM SIGKDD Explorations Newsletter,* January 2000, Volume 1 Issue2.

[4] Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma, Liu Wenyin, Ranking User's Relevance to a Topic through Link Analysis on Web Logs, *WIDM'02*, November 2002.

[5] A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition, 2002.

[6] G. Piatetsky-Shapiro, and W.J. Frawley, Knowledge Discovery in Databases. *AAAI/MIT Press*, 1991.

[7] Q. Lu, and L. Getoor. Link-based classification. In *Proceedings of ICML-03*, 2003.

[8] L. Getoor, Link Mining: A New Data Mining Challenge. *SIGKDD Explorations*, vol. 4, issue 2, 2003.

[9] S. Chakrabarti, B. E. Dom, D, Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Mining the Link Structure of the World Wide Web. February, 1999.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. *Technical report*, Stanford University, 1998.

[11] Han, J., Kamber, M. Kamber. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2000.

[12] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. Web mining:knowledge discovery on the Web. *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference-* on Volume 2, Page(s):137 - 141 vol.2 - 12-15 Oct. 1999

[13] Cooley, R.; Mobasher, B.; Srivastava, J.; Web mining: information andpattern discovery on the World Wide Web. *Tools with Artificial Intelligence,1997. Proceedings., Ninth IEEE* International Conference. Page(s):558 – 567 -3-8 Nov. 1997.

[14]Kleinberg, J.M.,Authoritative sources in a hyperlinked environment.*In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998*,pages 668-677 – 1998.

[15] Brin, S.; Page,L. The Anatomy of a Large-scale Hypertextual Web 594 Search Engine. *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[16] Bianchini, M.; Gori, M.; Scarselli, F.; Inside PageRank. ACM Transaction on Internet Technology (TOIT), Volume 5 Issue 1 – February, 2005.

[17]http://www.research.ibm.com/topics/popups/innovate/hci/html/clever .h tml. Last accessed 15/04/2005.

[18]http://www.google.com/. Last accessed 15/04/2005.

[19] Ziv Bar-Yossef and Sridhar Rajagopalan. Template Detection via Data Mining and its Applications. In: Proceedings of WWW2002, May 7-11, 2002,Honolulu, Hawaii, USA. 580-591.

[20] M. Kessler. Bibliographic coupling between scientific papers. AmericanDocumentation, 14:10-25, 1963.

[21] Taher H.Haveliwala.Topic-Sensitive PageRank: AContext-Sensitive Ranking Algorithm forWeb Search. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No4, July/August 2003, 784-796.

[22] Deng Cai, Shipeng Yu, and et al. Block-based Web Search. In Proceedings of ACM SIGIR'04, July 25-29, 2004, Sheffield, South Yorkshire,UK. 465-463. In:Proceedings of ACM SIGKDD'02, July 23-26, 2002, Edmonton, Alberta, Canada. 588-593.

[23] Sian-Hua Lin and Jan-Ming Ho. Discovering Information Content Blocks from Web Documents. In: Proceedings of ACM SIGKDD'02, July 23-26, 2002, Edmonton, Alberta, Canada. 588-593.

[24] M. Hajime, H. Takeo, and O. Manabu. Text Segmentation with Mltiple Survace Linguistic Cues. In: Proceedings of COLING-ACL. 1998, 881-885.

[25] Callan, J. P., Passage-Level Evidence in Document Retrieval, In Proceedings of ACM SIGIR'94, Dublin, 1994. 302-310.

[26] Xiang Ji and Hongyuan Zha. Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. In: Proceedings of SIGIR'03, July 28-August 1, 2003, Toronto, Canada. 322-329.

[27] Marti A. Hearst. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics, 23(1): 33-64, 1997.

[28] T. Brants, F. Chen, and I. Tsochantaridis. Topic-Based Document SegmentationwithProbabilisticLatent Semantic Analysis. In: Proceedings of CIKM'02, November 4-9,2002, McLean,Virginia, USA. 211-218.

[29] Deng Cai, Xiaofei He, Ji-Rong Wen, and Wei-Ying Ma, Block-level Link Analysis, In Proceedings of ACM SIGIR'04, July 25-29, 2004, Sheffield South Yorshire, UK.4

[30]Geetha R.B, Shahikumar G, Prasad Reddy PVGD,Amalgamation of Web usage Mining and Web Structure mining of the ACEEE,may2009,volNo.2,279-2.