International Journal of Modern Trends in Engineering and Research (IJMTER)
Volume 02, Issue 05, [May – 2015] **ISSN (Online):2349–9745 ; ISSN (Print):2393-8161**

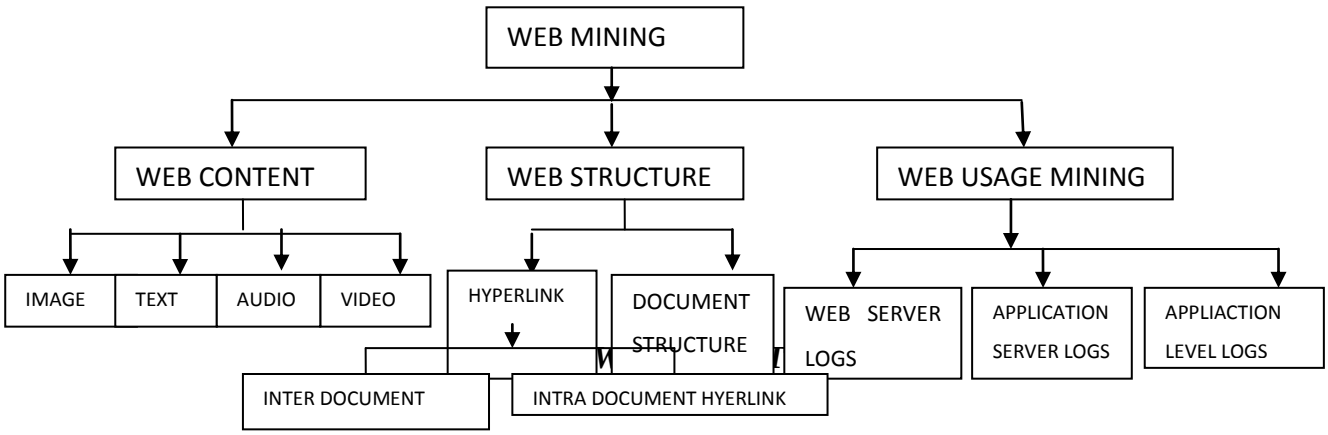
A STUDY ON WEB CONTENT MINING AND WEB STRUCTURE MINING

Ms. B.Nagarathna[1], Dr.M.Moorthi [2]
[1]*Research Scholar, Kongu Arts and Science College, Erode, TN, India*
[2]*Associate Professor, Kongu Arts and Science College, Erode, TN, India*

Abstract --From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. Web mining – i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage is the collection of technologies to fulfil this potential. Interest in Web mining has grown promptly in its short existence, both in the research and practitioner communities. This paper deals with a preliminary discussion of Web content mining, web structure mining contributions in the field of web mining, the prominent successful tools and algorithms.
Keywords: Web mining, Web content mining, Web structure mining

I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Internet has became an indispensable part of our lives now a days so the techniques which are helpful in extracting data present on the web is an interesting area of research. These techniques help to extract knowledge from Web data, in which at least one of structure or usage (Web log) data is used in the mining process (with or without other types of Web).In general, Web mining tasks can be classified into three categories [2,3]: *Web content mining, Web structure mining and Web usage mining.*



All of the three categories focus on the process of knowledge discovery of implicit, previously un-known and potentially useful information from the Web. Each of them focuses on different mining objects of the Web. Figure 1 shows the Web categories and their objects.

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 multi-media 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 [1]:

***Intelligent Search Agents*:** These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
Information Filtering/ Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
***Personalized Web Agents*:** These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

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 sovereign 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 [4].

Web usage mining focuses on techniques that could predict the behaviour of users while they are interacting with the WWW. Web usage mining collects the data from Web log re-cords to determine user access patterns of Web pages. There are several available research projects and commercial products that analyze those patterns for different purposes. [3].

II. WEB CONTENT MINING

The Web content mining refers to the discovery of useful information from web contents which include text, image, audio, video, etc. The mining of link structure aims at developing techniques to take lead of the collective verdict of web page quality which is available in the form of hyperlinks that is web structure mining [5]. It includes extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge amalgamation [1].

2.1 Web Content Mining Strategies

Web Content Mining Approaches: Two approaches used in web content mining are Agent based approach and database approach [6, 7]. The three types of agents are intelligent search agents, Information filtering/Categorizing agent, and personalized web agents [19]. Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine information. Adapted web agents learn user preferences and discovers documents related to those user profiles [6, 7].

Web content mining has the following approaches to mine data

(1) Unstructured text mining,
(2) structured mining,
(3) Semi-structured text mining, and
(4) Multimedia mining. [8]

Unstructured Text Data Mining: Most of the Web content data is of unstructured text data. Content mining requires application of data mining and text mining techniques [9]. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Some of the techniques used in text mining are

- Information Extraction,
- Topic Tracking,
- Summarization, Categorization,
- Clustering and
- Information Visualization [8].

Structured Data Mining: The Structured data on the Web represents their host pages. Structured data is easier to extract when compared to unstructured texts. The techniques used for mining structured data are

- Web Crawler,
- Wrapper Generation,
- Page content Mining.[8]

Semi-Structured Data Mining*:* Semi-structured data evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intra-document structure. The techniques used for semi structured data mining are

- Object Exchange Model (OEM),
- Top Down Extraction, and
- Web Data Extraction language.[8]

Multimedia Data Mining: The techniques of Multimedia data mining are;

- SKICAT,
- Color Histogram Matching, Multimedia Miner and Shot Boundary Detection.

2.2 Web Content Mining Algorithms

There are two common tasks involved in web mining through which useful information can be mined. They are Clustering and Classification. Here various classification algorithms used to fetch the information are described

i) Decision Tree: The decision tree is one of the powerful classification techniques. Decision trees take the input as its features and output as decision, which denotes the class information. Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5. The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. This split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned [15].

ii) k-Nearest Neighbour: KNN is considered among the oldest nonparametric classification algorithms. To classify an unknown example, the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation [15].

iii) Naive Bayes: Naive Bayes is a successful classifier based upon the principle of Maximum A Posteriori (MAP).Given a problem with K classes {C1, . . . , CK} with so called prior probabilities P (C1), . . . , P(CK), can assign the class label c to an unknown example with features such features $x=(x_1,...,x_N)$ such that **$c=argmaxcP(C=ckx_1.....x_N)$,** is choose the class with the maximum a posterior probability given the observed data. This posterior probability can be formulated, that is choosing the class with the maximum a posterior probability given the observed data. This posterior probability observed data. This posterior probability can be formulated,

$$P(C=ckx_1,..,x_N)=p(C=c)P(x_1,....,x_NkC=c)P(x_1.....x_N)$$

As the denominator is the same for all classes, it can be dropped from the comparison. Now, we should compute the so-called class conditional probabilities of the features given the accessible classes. This may be quite difficult taking into account the dependencies between features. This approach is to assume conditional independence i.e. x_1,. , x_N are independent. This simplifies numerator as P(C = c) P (x_1kC = c). P (x_NkC = c), and then choosing the class c that maximizes this value over all the classes c = 1. . . K [15].

iv)Support Vector Machine: Support Vector Machines are among the most robust and successful classification algorithms. It is a new classification method for both linear and nonlinear data and uses a nonlinear mapping to transform the original training data into a higher dimension. Among the new dimension, it searches for the linear optimal separating hyper plane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a adequately high dimension, data from two classes can be partitioned by a hyper plane [15].

v) Neural Network: The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. It contains an input layer, one or more hidden layers and an output layer. The basic unit in a neural network is a neuron or unit. The inputs to the network correspond to the attributes measured for each training tuple. The inputs fed simultaneously into the units making up the input layer. It will be weighted and fed simultaneously to a hidden layer. Number of hidden layers is arbitrary, although usually only one. Weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction [15]. As network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer.

vii) Cluster Hierarchy Construction Algorithm (CHCA)

The algorithm takes a binary matrix (a table) as input. The rows of the table correspond to the objects we are clustering. Here we describing this algorithm with web pages, but the method is applicable to other domains as well. The columns correspond to the possible attributes that the objects may have (terms appearing on the web pages for this particular application). When row i has a value of 1 at column j, it means that the web page corresponding to i contains term j. From this table, which is a binary representation of the presence or absence of terms for each web page, we create a reduced table containing only rows with unique attribute patterns (i.e., duplicate rows are removed). Using the reduced table, we create a cluster hierarchy by examining each row, starting

with those with the fewest terms (fewest number of 1's); these will become the most general clusters in our hierarchy.

The row becomes a new cluster in the hierarchy, and we determine where in the hierarchy the cluster belongs by checking if any of the clusters we have created so far could be parents of the new cluster. Potential parents of a cluster are those clusters which contain a subset of the terms of the child cluster. This comes from the notion of inheritance discussed above. If a cluster has no parent clusters, it becomes a base cluster. If it does have a parent or parents, it becomes a child cluster of those clusters which have the most terms in common with it. This process is repeated until all the rows in the reduced table have been examined or we create a user specified maximum number of clusters, at which point the initial cluster hierarchy has been created.

The next step in the algorithm is to assign the web pages to clusters in the hierarchy. In general there will be some similarity comparison between the terms of each web page (rows in the original table) and the terms associated with each cluster, to determine which cluster is most suitable for each web page. Once this has been accomplished, the web pages are clustered hierarchically. In the final step we remove any clusters with a number of web pages assigned to them that is below a user defined threshold and re-assign the web pages from those deleted clusters.

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 [16], 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.

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 [16].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:** the most recent upgrade of a classic data mining task to linked Domains. 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 web page.
2. **Link-based Cluster Analysis** The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the 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. **Link Strength.** Links could be associated with weights.
5. **Link Cardinality**. The main task here is to predict the number of links between objects. page categorization

- finding related pages
- finding duplicated web sites and to find out similarity between them

3.1 Web structure Mining Algorithms

3.1.1 HITS: (Hyper-link Induced Topic Search) Algorithm

In HITS concept, Kleinberg [17] 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).
According to Kleinberg [17], "*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 a page that is pointed to by many good hubs*".

Let **a** is the vector of authority scores and **h** be the vector of hub scores.
a=[1,1.....1],h=[1,1.....1];
do
a=A^Th;
h=Aa;
Normalize **a** and **h;**
While a and h do not converge (reach a convergence threshold)
a*=a**;**
h*=h**;**
return a*,h*
The Vectors **a*** and **h*** represent the authority and hub weight

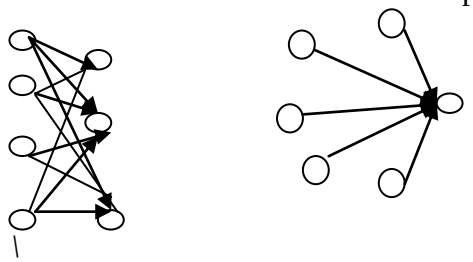
HITS Algorithm

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 [18]:

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.

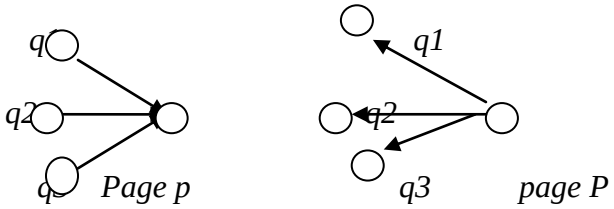
2. Automatically generated links: Web document generated by tools often have links that were inserted by the tool.

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



Hubs Authorities' unrelated page of large In-degree

Figure 2: A densely linked set of Hubs and Authorities



y[p]=sum of x[q],for all q pointing to by p

Figure.3 The basic operations of HITS

See Figure. 2. HITS associates a non-negative authority weight x and a non-negative hub weight y. See Figure. 3. The weights of each type are normalized so that their squares sum to 1.

3.2.2 Page Rank Model

L. Page and S. Brin [21] proposed the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. In their approach Brin and Page ex-tends the idea of simply counting in-links equally, by normal-izing by the number of links on a page. The Page Rank algorithm is defined as [21]: "*We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85.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)

3.2.3 Weighted page rank algorithms

Wenpu Xing and Ali Ghorbani proposed a Weighted Pagerank algorithm which is an extension of the Pagerank algorithm [20]. This algorithm assigns a larger rank values to the more important pages rather than Dividing the rank value of page evenly among its outgoing linked pages, each outgoing link gets a value proportional to its importance. In this algorithm weight is assigned to both back link and forward link. Incoming link is defined as number of link points to that particular page and outgoing link is defined as number of links goes out. This algorithm is more efficient than pagerank algorithm because it uses two parameters i.e. back link and forward link. The popularity from the number of in links and out links is recorded as Win and Wout respectively.

3.2.4 Weighted page content rank algorithm

Weighted Page Content Rank Algorithm (WPCR) [20] is a proposed page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of in links and out links of the page. If a page is maximally matched to the query, that becomes more relevant. This algorithm is better than the pagerank as well as weighted pagerank algorithm because its complexity is less than both the algorithm and is < (Ologn).

3.2.5 Topic sensitive page rank algorithm

In this algorithm, different scores are computed, multiple important scores for each page under several topics that form a composite Page rank score for those pages matching the query [19]. At query time, the similarity of the query is compared to each of these vectors or topics and subsequently, instead of using a single global ranking vector, the linear combination of the topic-sensitive vectors is weighed using the similarity of the query to the topics. This method yields a very accurate set of results relevant to the context of the particular query. For each web document query

sensitive importance score. The results are ranked according to this composite score *s*. It provides a scalable approach for search rankings using Link analysis. At query time, these importance scores are combined based on the topics of the query and associated context to form a composite Page rank score for those pages matching the query.

IV. CONCLUSION

The World Wide Web is the universe of network-accessible information, an embodiment of human knowledge. The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. Thus various Data mining techniques and web content mining, web structure are used to extract useful information or knowledge from web page contents. By these techniques we can make our search of contents over the web faster and exact.

REFERENCES

[1] Han, Kamber, M. Kamber. "*Data mining: concepts a techniques*" .Morgan Kaufmann Publishers, 2000
[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, and Web Usage Mining: Discoveryand Applications of Usage Patterns from Web Data, *ACM SIGKDD Explorations Newsletter*, January 2000, Volume 1 Issue 2.
[4] O. Etzioni. *The World Wide Web: Quagmire or gold mine*. Communications of the ACM, 39(11):65–68, 1996.
[5] G. Srivastava, K. Sharma, V. Kumar," *Web Mining: Today and Tomorrow"*, in the Proceedings of 2011 3rd International Conference on Electronics Computer Technolgy (ICECT), pp.39943April2011
[6] J. M. Kleinberg. *Authoritative sources in a hyperlinked environment.* In Proc. of ACM- SIAM Symposium on Discrete Algorithms, pages 668–677,1998
[7] R.Cooley,. B.Mobasher,.; J.Srivastava,.; *"Web mining: information and pattern discovery on the World Wide Web*". In Proceedings of Ninth IEEE International Conference. pp. 558 – 567, 3-8 Nov. 1997.
[8] F.Johnson., S.K.Gupta,., *Web Content Minings Techniques: A Survey*, International Journal of Computer Application. Volume 47 – No.11, p44, June(2012).
[9] Darshna Navadiya, Roshni, *Web Content Mining Techniques-A comprehensive Survey*, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181
[10] QZhang,., R.S.Segall,., *Web Mining: A Survey of Current Research Techniques, and Software*, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).
[11] Mozenda,*http://www.mozenda.com/web- mining-software Viewed 18 February 2013.*
[12] screen-scraper,*http://www.screen- scraper.com* Viewed 19 February 2013
[13] Web Content Extractor help.WCE,*http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February 2013*.
[14] Automation Anywhere Manual. AA, *http://www.automationanywhere.com*Viewed 06 February 2013.
[15] Darshna Navadiya, Roshni Patel, *Web Content Mining Techniques-A Comprehensive Survey*, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue10,December- 2012 ISSN: 2278-0181
[16] Getoor, Link Mining: A New Data Mining Challenge. *SIGKDD Explorations*, vol. 4, issue 2, 2003
[17] J.M.Kleinberg,., Authoritative sources in a hyperlinked environment i*n Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998*, pages 668-677 – 1998.
[18] 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.
[19] G. Piatetsky-Shapiro, and W.J. Frawley, Knowledge Discovery in Databases. *AAAI/MIT Press*, 1991
[20] Q. Lu and L. Getoor. Link-based classification. In *Proceedings of ICML-03*, 2003

