WEB CONTENT MINING TECHNIQUES – A COMPREHENSIVE SURVEY

Niki R. Kapadia *

Kinjal Patel *

ABSTRACT

The World Wide Web (WWW) is rich source of information and continues to expand in size and complexity in order to get maximum details on-line. However, it is becoming challenging ask to retrieve the required web pages/information very effectively and efficiently on the web. The paper contains techniques of web content mining, review, various algorithms, examples and comparison. Web mining is one of the well-know technique in data mining and it could be done in three different ways (a)web usage mining, (b)web structure mining and (c)web content mining. Web usage mining allows for collection of web access information for web pages. Web content mining is the scanning and mining of text, pictures and graphs of web page to determine relevance of content to the search query. Web structure mining is used to identify the relationship between the web pages linked by information. The paper presents various examples based on web content mining techniques in detail, results and comparison to extract necessary information effectively and efficiently.

Keywords: *Data Mining, Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Clustering, Segmentation.*

* Govt. Engineering College, Modasa.

I. INTRODUCTION

Web mining is the data mining technique that automatically discovers/extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.

Mining techniques are in detail, results and comparison to extract necessary information effectively and efficiently.

II.WEB MINING PROCESS

Web mining process is shown in figure 2. And its steps are given below:

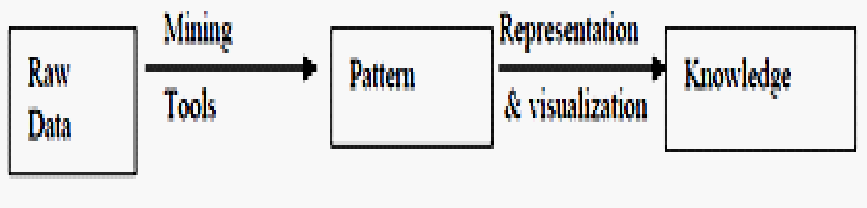

Figure 2 web mining process

Steps of web mining process:

a. Resource Finding: - It is the task of retrieving intended web documents.
b. Information selection and preprocessing:- Automatically selecting and pre-processing specific from information retrieved web resources.
c. Generalization:-Automatically discovers general patterns at individual web sites or multiple sites.
d. Analysis:-Validation and interpretation of the mined patterns.

III. WEB MINING CATEGORIES

Web mining can be categorized as below.

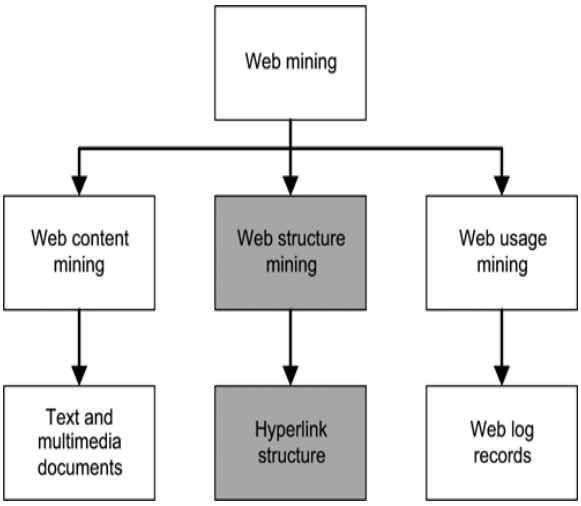

Fig 2. Web mining categories

a. **Web Content Mining:** - Web content mining is the process of extracting useful information from the contents of web documents. It is related to data mining. It is related to text mining because much of the web contents are text based. Text mining focuses on unstructured texts. Web content mining is semi-structured nature of the web. Technologies used in web content mining are NLP,IR.

b. **Web Structure Mining:** - tries to discover useful knowledge from the structure and hyperlinks. The goal of web structure mining is to generate structured summery about websites and web pages. It is using tree-like structure to analyze and describe HTML or XML.

c. **Web Usage Mining:** - Web usage mining is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. It focuses on technique that can be used to predict the user behavior while user interacts with the web. It uses the secondary data on the web. This activity involves automatic discovery of user access patterns from one or more web-servers. It consists of three phases namely: pre-processing, pattern discovery, pattern analysis. Web servers, proxies and client applications can quite easily capture data about web usage.

IV. WEB CONTENT MINING METHODS

With the rapid growth of the web, there is an increasing volume of data and information published in various web pages. Web mining refers to develop new techniques to effectively extract and mine useful information/knowledge from the web pages. There are various methods of web content mining.

a. Structured data extraction.

Structure data extraction is most widely used in web content mining. Structured data is easier to extract compare to unstructured data. There are several approaches to structure data extraction, called wrapper generation.

The **first** approach is to manually write an extraction program for each web site based on observable format patterns of the site. This approach is time consuming. It doesn't scale for large number of sites.

The **second** approach is wrapper induction/wrapper learning. The user first manually labels set of trained pages. A learning system then generates rule from the training pages. The resulting rules are then applied to extract target items from web pages. E.g.: WIEN, stalker, BWI, WL^2.

The **third** approach is automatic approach. Structured data objects on the web are normally retrieved from database and displayed in the web pages with fix templates.E.g. MDR, Roadrunner, EXALG etc.

b. Unstructured text extraction.

Most Web pages can be seen as text documents. Extracting information from Web documents has also been studied by many researchers. The research is closely related to text mining, information retrieval and natural language processing. Current techniques are mainly based on machine learning and natural language processing to learn extraction rules. Recently, a number of researchers also make use of common language patterns (common sentence structures used to express certain facts or relations) and redundancy of information on the Web to find concepts, relations among concepts and named entities. The patterns can be automatically learnt or supplied by human users. Another direction of research in this area is Web question-answering. Although question-answering was first studied in information retrieval literature. it becomes very important on the Web as Web offers the largest source of information and the objectives of many Web search queries are to obtain answers to some simple questions.

c. Web Information Integration.

Due to the sheer scale of the Web and diverse authorships, various Web sites may use different syntaxes to express similar or related information. In order to make use of or to extract formation from multiple sites to provide value added services, e.g. metasearch, deep Web search, etc, one needs to semantically integrate information from multiple sources. Recently, several researchers attempted this task. Two popular problems related to the Web

are (1) Web query interface integration, to enable querying multiple Web databases (which are hidden in the deep Web) and (2) schema matching, e.g. integrating Yahoo and Google's directories to match concepts in the hierarchies. The ability to query multiple deep Web databases is attractive and interesting because the deep Web contains a huge amount of information or data that is not indexed.

d. Building concept hierarchies.

Because of the huge size of the Web, organization of information is obviously an important issue. Although it is hard to organize the whole Web, it is feasible to organize Web search results of a given query. A linear list of ranked pages produced by search engines is insufficient for many applications. The standard method for information organization is concept hierarchy and/or categorization. The popular technique for hierarchy construction is text clustering, which groups similar search results together in a hierarchical fashion. Several researchers have attempted the task using clustering. A different approach is proposed which does not use clustering. Instead, it exploits existing organizational structures in the original Web documents, emphasizing tags and language patterns to perform data mining to find important concepts, sub-concepts and their hierarchical relationships. In order words, it makes use of the information redundancy property and semi-structure nature of the Web to find what concepts are important and what their relationships might be. This work aims to compile a survey article or a book on the Web automatically.

e. Segmenting Web pages & Detecting noise.

In web data mining, classification and clustering are used to remove noisy blocks and enables to produce much better results. Another application is web browsing using a small screen device called PDA.

f. Mining web opinion sources.

Web was available. Companies usually conduct consumer surveys or engage external consultants to find such opinions about their products and those of their competitors. Now much of the information is publicly available on the Web. There are numerous Web sites and pages containing consumer opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. This online word-of-mouth behavior represents new and measurable sources of information for marketing intelligence. Techniques are now being developed to exploit these sources to help companies and individuals to gain such information effectively and easily. For instance, proposes a feature based summarization method to automatically analyze consumer opinions in customer reviews from online merchant sites and dedicated

review sites. The result of such a summary is useful to both potential customers and product manufacturers.

V. ALGORITHMS OF WEB CONTENT MINING

There are various techniques available through which we can mine useful information. Here, In this paper, I am describing various clustering algorithms used to fetch information. Various clustering algorithms are described as below.

a. Hierarchical Clustering

Hierarchical Clustering is a method of cluster analysis which builds hierarchy of clusters. It is the collection of objects arranged in hierarchical fashion. They are available in types.

- Agglomerative:- It starts with each object being separate cluster itself, and merges groups according to distance measure. Clustering stops when all objects are in within single group. This method follows bottom-up merging.
- Divisive:-This follows the opposite strategy. They start with one group. Group contains all the objects. Then group is divided in to smaller groups, until each objects falls in one cluster. This approach is similar to Divide-and-conquer algorithms.

Most of the times, both approaches suffer from the fact that once a merge or a split is committed, it cannot be undone or refined.

Process of algorithms is as follow:

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

Disadvantages:

The main weaknesses of hierarchical clustering algorithms are:

- They do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
- They can never undo what was done previously.

b. Partitional Clustering Algorithm

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroid can be further clustered to produces hierarchy within a dataset.

Single Pass:

A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid. Otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

One of the issues with such algorithms is their high complexity, as some of them exhaustively enumerate all possible groupings and try to find the global optimum. Even for a small number of objects, the number of partitions is huge. Partitional Clustering algorithms try to locally improve a certain criterion. First, they compute the values of the similarity or distance, they order the results, and pick the one that optimizes the criterion. Hence, the majority of them could be considered as greedy-like algorithms.

K-means algorithm is partitioning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

Advantages:

- It is relatively efficient compared to other algorithms

Disadvantages:

- Need to specify k, the number of clusters, in advance
- Unable to handle noisy data and outliers
- Not suitable to discover clusters with non-convex shapes

IV. CONCLUSION

This Paper concludes that there are many existing methods to mine information. Hierarchical and partitioning methods are used commonly to mine information from the web. Each method has advantages and disadvantages. Regarding the future work, other algorithms of clustering to improve performance.

V. REFERENCES

1. http://en.wikipedia.org/wiki/Data_mining
2. http://faisalsikder.wordpress.com/2010/02/01/the-scope-of-data-mining-and-kdd/
3. Data Mining: Concepts and Techniques, Jiawei Han and Micheline Kamber, 2nd Edition.
4. http://www.thearling.com/text/dmtechniques/dmtechniques.htm
5. http://en.wikipedia.org/wiki/Web_mining
6. http://www.web-datamining.net/structure/
7. http://www.web-datamining.net/content/
8. http://www.web-datamining.net/usage/
9. http://www.web-datamining.net/
10. Markus Schedl1, Peter Knees1, Tim Pohle1, and Gerhard Widmer.(May 2011)."Towards an Automatically Generated Music Information System via Web

Content Mining". In Proceedings International journal of Information Processing & Management 47 in (2011), (PP.426-439)

11. Bing Liu, Kevin Chen-chuan Chang..."Editorial Issues on Web content Mining".SIGKDD Explorations –Volume 6.
12. Jing Li and C.I. Ezeife. "Cleaning Web Pages for Effective Web Content mining",