# WEB MINING: AN OVERVIEW

**P.N. Mulkalwar[1] & A.K. Shingarwade[2]**
[1]Department of Computer Science, Amolakchand Mahavidyalaya, Yavatmal (M.S.), India
pnm_amv@rediffmail.com
[2]Department of Computer Science, College of Management & Comp. Sci., Yavatmal, (M.S.), India
atul_shingarwade@rediffmail.com

## Abstract

*Web mining an interdisciplinary research area spanning several disciplines such as machine learning, database system, expert system, intelligent information systems and statistic. Web mining has evolved into an active and important area of research because of previously unknown and interesting knowledge from very large real-world database. Many aspects of web mining have been investigated in several related fields. Main but important aspect of the problem lies in the significance of needs to extend their studies to include the nature of the contents of the real world database. Web mining is also used to get social aspect of the individual to get track of its activities and other information by different ecommerce websites. In this paper we will review the concept of web mining.*

**Keywords:** *Data mining, web mining, E-commerce, Social e-commerce, Clustering.*

## Introduction

In present day human beings are used in the different technologies to adequate in there society. Every day the human beings are using the vast data and these data are in the different fields .It may be in the form of graphical formats, may be documents, may be the records (varying array), may be video .As the data are available in the different formats so that the proper action to be taken for better utilization of the available data. As and when the customer will require the data should be retrieved from the database and make the better decision.

This technique is actually we called as a data mining or Knowledge Hub or simply KDD (Knowledge Discovery Process).The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of “Data mining” is due to the perception of “we are data rich but information poor”. There is very large amount of data but we hardly able to turn them in to useful information and knowledge for managerial decision making. To generate information it requires very large database. May be it available in different formats like audio/video, numbers, text, figures, and Hypertext formats. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for extraction of the essence of information stored, automatic summarization of data and the discovery of patterns in raw data.

The World Wide Web (Web) is a popular and interactive medium to disseminate information today. The Web is dynamic, diverse, and huge and thus raises the temporal, multimedia data, and scalability issues respectively in database. Due to those situations, we are currently drowning in information and facing information overload. With the huge amount of online available information, the www is a fertile area for research in data mining. The Web mining research is at the cross road of research from several research communities, such as information retrieval, database, and within AI, especially the sub-areas of machine learning and natural language processing. Every E-commerce website in World Wide Web uses the web mining for the technique for the item navigation on the e-commerce site. In this paper we review the concept of web mining and different methods and algorithms used for the web mining and finds advantages and limitations of them.

## Web Mining

We are living in a world which is surrounded by infinite number of data. The world has accepted computers as the best means for storing data. This is because it is very it is, convenient easy to store data, anyone with access to a computer can do it, and most importantly, information stored can be shared among number of users or it can be transferred to other locations. However, as more text documents are stored in very huge database, it becomes very difficult to understand hidden patterns or relations in the data. Since text data is not in numerical format, it is not possible to analyze it with statistical methods.

However every day, people encounter a large amount of information and store or represent it as data, for further analysis and management. A considerably large portion of information present on the World Wide Web (WWW) today is in the form of unstructured or semi-structured text databases. The World Wide Web instantaneously delivers very high number of these files in response to a user request. However, because the lack of structure information, the users are at a loss to manage the information contained in these data efficiently. The WWW continues to grow at very fast rate and become an information gateway and as a medium for conducting business. Web mining is the extraction method of interesting and useful knowledge and implicit information from activity or artifact related to the World Wide Web.

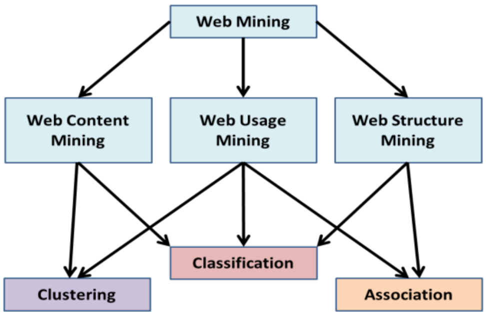

Figure 1: General relationship between web mining and its components

Figure 1 shows the basic flowchart of the web mining and its different components. Mainly there are three main components of web mining and they are web structure mining, web usage mining and web content mining. These components are then divided into different methods i.e. clustering, classification and association of data as show in figure 1.

## Review of Literature

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services (Etzioni, O., 1996). This area of research is so huge today partly due to the interests of various research communities, the ultimate growth of data information sources available on the Web and the recent interest in e-commerce websites. This creates confusion when we ask what constitutes Web mining and when comparing research in this area.

It is Oren Etzioni first proposed the term of Web mining in 1996. He claimed the Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. Many of the following researchers cited this explanation in their works. In the same paper, Etzioni came up with the question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has become the focus of quite a few research projects and papers. Some of the commercial consideration has presented on the schedule.

Oren Etzioni (1996) and Kosala, Blockeel (2000) suggested a similar way to decompose Web mining into the following subtasks:

a.Resource Discovery: the task of retrieving the intended information from Web.

b.Information Extraction: automatically selecting and pre-processing specific information from the retrieved Web resources.

c.Generalization: automatically discovers general patters at the both individual Web sites and across multiple sites.

d.Analysis: analyzing the mined pattern.

In brief, Web mining is a technique to discover and analyze the useful information from the Web data. Kosala, Blockeel (2000) claims the Web involves three types of data: data on the Web (content), Web log data (usage) and Web structure data. H. Michael Chung and Paul Gray (Chung H.M. et al., 1999) classified the data type as user profile data, structure data, usage data, and content data. M. Spiliopoulou et al. (1999) categorized the Web mining into modeling mining, Web text mining and Web usage mining; while today the most recognized categories of the Web data mining are Web usage mining, Web structure mining, and Web content mining (Kosala, R., 2000). It is clear that the classification is based on what type of Web data to mine.

The Web content mining is differentiated from two different points of view: Database View and Information Retrieval View. Kosala and Blockeel (2000) structured the research works done for unstructured data and semi-structured data from information retrieval. It shows that most of the researches use bucket of similar words, which is based on the statistics about single words in isolation, to represent unstructured take and text single word found in the training subjects as features. But for the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the data for data presentations. As for the database view, in order to have the better information querying and management on the Web, the mining always tries to infer the structure of the Web site of to transform a Web site to become a database.

Chakrabarti (2000) provides a in-depth survey of the research on the application of the techniques from statistical pattern recognition, data mining, and machine learning to analyzing hypertext. It's a good resource to be aware of the recent advances in content mining research.

Madria et al. (1999) gave a detailed description about how to discover interesting and informative facts describing the connectivity in the Web subset, based on the collection of interconnected web data. The information generated from the Web structure mining includes the follows: the information measuring the frequency of the local links in the Web tuples in a Web table; the information measuring the frequency of Web tuples in a Web table containing links that are interior and the links that are within the same document; the information measuring the frequency of Web tuples in a Web table that contains links that are global and the links that span different Web sites; the information measuring the frequency of identical Web tuples that appear in a Web table or among the Web tables.

Spiliopoulou (1999) abstract the potential strategic aims in each domain into mining goal as: prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users. There are no definite distinctions between the Web usage mining and other two categories. In the process of data preparation of Web usage mining, the Web content and Web site topology will be used as the information sources, which interacts Web usage mining with the Web content mining and Web structure mining. Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.

The last comprehensive survey on web usage mining has been done by Koutri, Avouris, and Daskalaki (2004), Pierrakos et al. (2003) earlier explored the area of web usage mining as a tool for the personalization process. Eirinaki, and Vazirgiannis (2000) also presented an excellent survey upon a overall area of web mining for personalization. Kosala and Blockeel (2000) explored the terms of web mining and the related research area earlier in their work. Since then web mining research has been somewhat traced in the annual WebKDD workshop. WebKDD 2008 was the tenth of a successful series of Workshops. Launched in 1999 by Brij Masand and Myra Spiliopoulou, the first workshop of the series WebKDD 1999 invited contributions on “Web usage mining”. In the 10 years that followed, the scope of WebKDD was broadened to cover the emerging KDD topics on the web. Together with this, historical study has been conducted by several researchers that are specialized in web mining techniques and several frameworks have already been explored. The result of these researches has lead to the development of lot of applications in the area of web mining and they are successfully applied in business and e commerce domain areas.

Few landmark researches have been followed here. Cooley et al. (1997) proposed a framework for web mining using various web mining task and implemented a prototype namely Webminer. It is implemented by applying a framework that performs cluster analysis on sequential pattern discovery and association rules. Author Zaiane (1998) proposed the idea of how to implement the OLAP technique on the Web mining. Their works on the multimedia data also provided a valuable solution for content mining. Cooley (2000) in University of Minnesota did in-depth research to all the procedure of usage mining. They proposed a mining prototype Web Miner and derived a system Web SIFT to perform the usage mining, which is practical possible. Lee and Liu (2004) proposed an intelligent multi-agent based environment known as intelligent Java Development Environment (iJADE) to provide an integrated and intelligent agent based platform in the e-commerce environment on online shopping. The different application on intelligent agent for helping users is applied in various applications and not only in e-commerce environment.

Mobasher et al. (2001) proposed effective and scalable techniques for Web personalization based on association rule discovery from usage data. Toolan and Kushmerick (2002) proposed techniques based on web usage mining to deliver personalized Site Maps that are specialized to the interest of each individual visitor. Applying the agent technology has improved the performance of web mining compared to traditional approach such as database approach for web mining. Besides that, Eirinaki and Vazirgiannis (2000) developed a module that comprises a web personalization system concerning the module on web usage mining. Also, Lu, Dunham, and Meng (2005) later proposed a technique to generate Significant Usage Patterns (SUP) and used them to acquire significant “user preferred navigational trails”. Falkowski et al. (2006) proposed two approaches to analyze the evolution of two different types of online communities on the level of subgroups.

## Clustering

Clustering can be considered the most important *unsupervised learning* problem occurred in web mining for e-commerce websites; so, all other problem of clustering is deals with finding a *structure* in a collection of unlabeled data A loose definition of clustering could be “the process of organizing objects into groups whose members are similar to each other”. A *cluster* is defined as the collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.
We can explain this with a very simple example as follows:

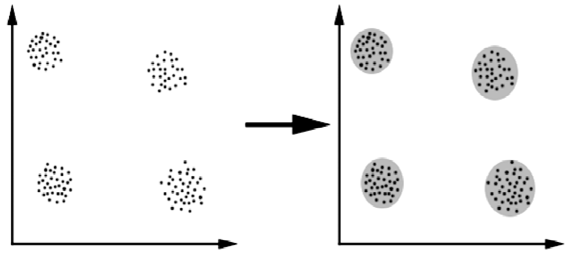

Figure 2 Graphical Presentation of clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are “close” according to a geometrical distance (i.e. given distance) is called a *distance-based clustering*. *Conceptual clustering is another kind of clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all objects. In another

words we can say that objects are collected according to their fit to descriptive concepts but noting according to similarity measures.
So, the goal of clustering is to determine the intrinsic grouping in a set of n labelled data. But how to decide what constitutes a better clustering? It is shown that there is nothing absolute “best” criterion which would be independent of the final aim of the clustering. Apparently, it is the user which must gives this criterion, According to the needs of the user to get the proper results.

There are different clustering algorithms proposed by different authors in the last few decades. Some of the areas follow:

## 1. K-Means Clustering Algorithm

The k-means is the simplest and most commonly used algorithm employing a squared error criterion. It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is met (e.g., there is no reassignment of any pattern from one cluster to another, or the squared error ceases to decrease significantly after some number of iterations). The k-means algorithm is very popular for clustering because its time complexity is *O*(n), where *n* is the number of patterns and it is easy to implement. The main problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function value if the initial partition is not properly chosen.
A partition clustering algorithm obtains a single partition of the data instead of a clustering method, such as the dendrogram produced by a hierarchical technique. Partition methods have advantages in applications involving large data sets for which the construction of a dendrogram is computationally prohibitive. A problem accompanying the use of a partition algorithm is the choice of the number of desired output clusters. A seminal paper (Dubes 1987) provides guidance on this key design decision. The partition techniques usually produce clusters by optimizing a criterion function defined either locally (on a subset of the patterns) or globally. Combine search of the set of possible labeling for an optimum value of a criterion is clearly computationally prohibitive. Therefore, the algorithm is typically run multiple times with the best configuration, and different starting states obtained from all of the runs is used as the output clustering.
The most intuitive and frequently used criterion function in partition clustering techniques is the squared error criterion, which tends to work well with compact and isolated clusters. The squared error for a clustering y of a pattern set x (containing *K* clusters) is

$$e^2(x,y) = \sum_{j=1}^{K}\sum_{i=1}^{n_j} ||x_i{}^{(j)} - c_j||^2 ,$$

where $x_i{}^{(j)}$ is the $j^{th}$ pattern belonging to the $j^{th}$cluster and $c_j$ is the centroid of the $j^{th}$ cluster.

## 2. Bisecting K-Means Clustering

Bisecting k-Means is like a combination of k-Means and hierarchical clustering. It starts with all objects in a single cluster.
The pseudo code of the algorithm is displayed below:

Basic Bisecting K-means Algorithm for finding K clusters

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic k-Means algorithm (*Bisecting step*)
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

The critical part is which cluster to choose for splitting. And there are different ways to proceed, for example, you can choose the biggest cluster or the cluster with the worst quality or a combination of both. Bisecting k-means clustering algorithm can be used to partition the item from the e-commerce site as per the client interests and his previous searches.

## Social Networking with E-Commerce

Online service providers, such as Facebook, Twitter, LinkedIn and Amazon, are beginning to collect various kinds of public and private data across the Web for the purposes of targeted marketing. Despite this explosion in social networking among businesses, there are still many who wonder how exactly social media directly affects a business' online commerce. In a report published back in April 2011, Forrester Research, challenged the idea that sites like Facebook are the future for driving e-commerce business, concluding that social media has little impact on online retail purchases. There are a lot of business marketing incentives to combine the data stored by these online services. Platforms have been introduced by these companies to better combine e-commerce and social networking.

A vast variety of data is collected about the user, and the current legal system has different legal standards for different kinds of data. This data can be used by e-commerce websites to give the options to their clients. For example if A and B are two friends in a Facebook social networking side, and if A purchased or view some item from the particular e- commerce website. When B comes or open the same e-commerce website products viewed and purchased by A are recommended to B because A and B are friends and they may like similar kind of products.

## Conclusion

In this paper we survey the research in the area of Web mining. We study the clustering algorithms used for the web mining to partition of the data. We try to cover maximum paper available on the web mining as possible. We also study the social networking and how it used with the e-commerce. In this paper we review the clustering technique used for the web mining.

## References

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

**Kosala, R., Blockeel, H. (2000).** Web mining Research: A Survey, ACM SIGKDD Explorations, 2(1), pp.1-15.

**Chung H.M., Gray, P. (1999).** Current Issues in Data Mining, Journal of Management Information Systems, vol. 16, no. 1, pp. 11-16,

**Spiliopoulou, M., Faulstich, L.C. & Winkler, K. (1999).** A Data Miner analyzing the Navigational Behaviour of Web Users. Proceeding of the workshop on machine learning in user modeling of the ACAI'99, international conference Creta, Greece.

**Chakrabarti, S. (2000).** Data Mining for hypertext: A tutorial survey, ACM SIGKDD Exploration. vol 1, issue 2, pp. 1-11.

**Madria, S. K., Bhowmick, S. S., Ng, W. K. & Lim, E. P. (1999**). Research issues in web data mining, in proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK'99, pages 301-312.

**Spiliopoulou, M. (1999).** Data mining for the Web. In Proceedings of Principles of Data Mining and Knowledge Discovery, Third European conference, PKDD'99, P588-589.

**Koutri, M., Avouris, N. & Daskalaki, S. (2004).** A survey on web usage mining techniques for web-based adaptive hypermedia systems, Adaptable and Adaptive Hypermedia Systems Idea, pp. 1-23.

**Pierrakos, D., Paliouras, G., Papatheodorou, C. & Spyropoulos, C. D. (2003).** Web usage mining as a tool for personalization: A survey, User Modeling and UserAdapted Interaction, Kluwer Academic Publishers, 13(4), pp. 311-372.

**Eirinaki M. & Vazirgiannis, M. (2000).** Web mining for web Personalization, ACM Transactions on Internet Technology, 3(1), pp. 1-27.

**Cooley, R., Mobasher, B. & Srivastava, J. (1997).** Web mining: information and pattern discovery on the World Wide Web', Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence, 97(2.1), pp. 558-567.

**Zaiane, O. R. (1998).** Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs, Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries ADL98, IEEE Computer Society, Santa Barbara, CA, pp. 19-29.

**Cooley, R. (2000).** Web Usage Mining: Discovery and Application of Interesting Patterns from Web data, PhD thesis, Dept. of Computer Science, University of Minnesota, USA.

**Lee, R. S. T. & Liu, J. N. K. (2004).** iJADE Web-Miner: An Intelligent Agent Framework for Internet Shopping, IEEE Transactions on Knowledge and Data Engineering, 16(4), pp. 461- 473.

**Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2001).** Effective personalization based on association rule discovery from web usage data', Proceeding of the third international workshop on Web information and data management WIDM 01, 9, USA, pp. 9-15

**Toolan, F. & Kusmerick, N. (2002).** Mining Web Logs for Personalized Site Maps, Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) - (WISEw'02) (WISEW '02). IEEE Computer Society, Washington, DC, USA, pp. 232-237.

**Lu, L., Dunham, M. H. & Meng, Y. (2005).** Discovery of Significant Usage Patterns from Clusters of Clickstream Data, Proceedings of the ACM SIGKDD Workshop on Knowledge Discovery in Web WebKDD05, Chicago, IL, USA.

**Falkowski, T., Bartelheimer, J. & Spiliopoulou, M. (2006).** Mining and Visualizing the Evolution of Subgroups in Social Networks, Proceedings of the 2006 IEEE WICACM International Conference on Web Intelligence, pp. 52-58.