A Survey on Web Data Mining Applications

Dhanashree S. Deshpande
K.K. Wagh College of
Engineering, Nashik

ABSTRACT
The web has become an important medium to conduct business, communicate ideas and entertainment. It is essential to collect data from various data sources. The process of finding quality and useful information is tedious and a frustrating experience. Assembling and managing data from various sources produces reliable database in data warehouse. Data mining techniques are used in web mining applications to find interesting and potentially useful knowledge from web data. Data mining tools can mine the required information from mass web data and it is helpful for organizations to improve work efficiency. The aim of this paper is to give overview of web data mining categories and some of its applications.

Keywords
Web mining applications; data mining; web content mining; web usage mining; web structure mining

1. INTRODUCTION
In recent years, everyone found web as the main source of information retrieval & further utilization to produce quality output in the new generation of choosy & criticizing customer. In coming future years, competition will produce tremendous run of market position by using techniques of web data mining. Web data mining is defined as searching useful hidden information & patterns on the web. Web mining is the application of data mining techniques. Web has grown exponentially along with its strengths as well as its weakness. The strength is that one can find out information on just about anything even if the quality varies. The weakness is that there is the problem of abundance of information. For example searching for phrase "web data mining' on yahoo found more than 2 million pages. How does one plough through such a large collection of documents to find those that are more useful? How will search engine provide ranking for such large number of documents? It is the aim of web mining to find out solution for problem like this. Huge collection of diversified data in various formats on web requires techniques to organize data in proper manner so that any relevant data get accessed. The user always wants to search relevant information on classical web search engines with the method of keyword text-matching. But due to tremendous data user gets confused and may divert somewhere else. The web data mining technology helps to exactly match user's meaning and increases network consumption.

Data mining mainly deals with structured form of data organized in a database while web mining is unstructured form of data. So mining of web data is one of the most challenging tasks for the data mining [2]. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process. The paper is organized as follows. Section 2 presents types of web mining. Section 3 presents web data mining applications. Section 4 presents web data mining in e-commerce which explains detail process.

2. TYPES OF WEB MINING
Web mining is divided into several types:

2.1 Web Content Mining
It deals with discovering important and useful information or knowledge from web page contents. It contents unstructured information like text, image, audio, and video. Search engines, subject directories, intelligent agents, cluster analysis are employed to find what a user might be looking for. We can automatically classify and cluster web pages according to their topics [14]. We can discover patterns in web pages to extract useful data such as descriptions of products, postings of forums etc.

2.2 Web Structure Mining
It deals with discovering and modeling the link structure of the web. Web information retrieval tools make use of only the text available on web pages but ignoring valuable information contained in web links. Web structure mining aims to generate structural summary about web sites and web pages. The main focus of web structure mining is on link information [2]. 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. This can help in discovering similarity between sites or in discovering important sites for a particular topic.

2.3 Web Usage Mining
It deals with understanding user behavior in interacting with the web site or with web. Web usage mining mines weblog records to discover user access patterns of web pages. The logs include the web server logs, proxy server logs and browser logs. A web server usually registers a web log entry for every access of a web page. It includes the URL requested, the IP address from which the request originated and a timestamp. Information is also collected from cookie files. Raw weblog data need to be cleaned, condensed and

transformed in order to retrieve and analyze significant and useful information. Web structure mining shows that page A has a link to page B, web usage mining shows who or how many people took that link, which site they came from and where they went when they left page B.

Web Logs
172.158.133.121..[10/dec/2011:21:40:00-0800].Test-1.pdf
172.158.134.122..[11/dec/2011:15:25:00-0400].Test-2.pdf
172.158.134.123..[12/dec/2011:01:15:00-0200].Test-3.pdf

Fig 1 : Web Log Example

Web log information can be integrated with web content and web linkage structure mining to help web page ranking. The categories discussed above are dependent as web structure mining is closely related to web content mining and both are related to web usage mining. Research scope does exist in web usage mining. Recognizing useful information or patterns, classification of data, personalization of data, preprocessing of data are the research areas.

3. WEB DATA MINING APPLICATIONS

3.1 Intelligent Intrusion Detection System

Intrusion detection is a technology which actively protects the host from attack. We can solve classical problem in intrusion detection system by applying web data mining technology and intelligent agent technology. Network security becomes important issue as millions of users are using web daily for various purposes. The network problems like tampering websites, webpage hang horse, phishing, Trojan horse infected host have been flourished [13]. The technologies like firewall, anti-virus software, data encryption are applied to network security. These technologies have few limitations like firewall does not check whether the data streams passed to it have malice codes or not. For this purpose intrusion detection system is useful. Intrusion detection system provides security to the host from attack. The system consists of three modules data acquisition module, intrusion analyzing engine and emergency response. Data acquisition module consists of log files and web, abnormal program execution, network traffic, abnormal directories or file change. Intrusion analyzing engine uses method of pattern matching, statistical analysis and integrity analysis to analyze collected data and then identifies whether it is intrusion and then respond to it. Emergency response module produces emergency actions. For example, stopping the network connections, stopping the process and start backup, shutting down the web service. A classifier can then be derived to detect known intrusions. Association mining can be applied to find relationships between system attributes describing the network data.

Problems in Intrusion Detection System

1. The intrusion detection system can not detect few complicated attacks by some well trained attacker.

2. The ability of adaption and self-learning is not sufficient to form collaborative defense system [13].

The techniques of web data mining and intelligent agent are the basis of solving these problems. Web data mining has different advantage to acquire unknown knowledge by finding out noisy attacks and discovering dangerous and hidden intrusion.

3.2 Diet Recommender System

In current competitive world, fast food becomes important and popular for most of the people. But later it impacts on health of human being. Web Data Mining helps to solve this problem and produces solution in a greater extend. This system tracks eating habit of a person and recommends how it would be dangerous for health and gives information about diseases it will raise [7]. Also it recommends various food types to build your health. It also gives information to individual like what you lack in you body, what you have in greater quantity and related illness etc. It uses association rule mining to recommend quality dishes according individual likes. The system contains three modules: eating data acquisition, data mining process and healthy recommendation.

1. Data Acquisition - Food name, time of eating, food material, amount etc.

2. Data Mining - Data mining algorithm like clustering, classification, association rules etc are applied on this data to find out how much nutrients in each kind of food, fat, energy and vitamin calculation in your dish. It gives output whether your diet is proper or not.

3. Recommendation of Food - The system compares your health with standards. By data mining process we get information like in which nutrients you lack in, which are the factors too much in you, possible diseases, suitable exercise. It also gives recipe, tips by using association rules mining etc. The system helps to improve your health.

Challenges in Designing Healthy Eating Recommender System

1. Domain expertise knowledge is required. For example, to match recommendations produces by the system with standards it requires knowledge of doctor, to recommend suitable exercise knowledge of physicist is required, sheaf's knowledge to prepare best recipe for healthier diet.

3.3 Distance Learning

Data mining methods are essential to evaluate student performance and to improve courseware. E-learning environment transforms large amounts of useless data into an intelligent monitoring and recommendation system applied to the learning process [13].

3.4 Customer Relationship Management

By studying browsing and purchasing patterns on web stores, companies can tailor advertisements and promotions to customer profiles, so that customers are less likely to be annoyed with unwanted mass mailings or junk mail. These actions can result in substantial cost savings for companies.

4. WEB DATA MINING IN ELECTRONIC COMMERCE

The web data mining process can be divided as the five functional processes [7]. Those are data acquisition, data pre-processing, data mining, analysis evaluation and knowledge formulation modules. The functioning of each one is shown in Figure 2.

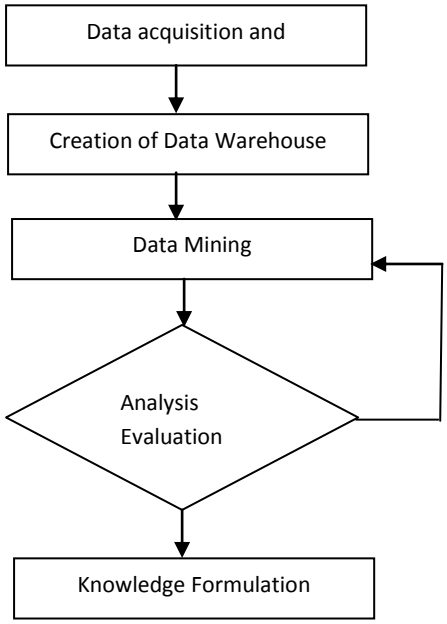

Fig 2 : Web data mining process

Now a days, organization are moving towards a web-enabled economy and increasingly rely on online information sources for a variety of decision support applications, data mining techniques are growing towards electronic commerce application. The goal of data mining is to extract qualitative models, translate into business patterns and visual representations. In electronic commerce, data mining techniques plays vital role in providing companies relevant and useful information. In this section, we survey data mining implementation in e-commerce.

4.1 Data Acquisition and Preparation

The process of gathering the requirements for data analysis is critical to the eventual success of any data mining project. One can not use data mining without a good idea of what kind of outcomes the enterprise is looking for. The organization experts need to present target which is to achieve by data mining process. It is essential process to acquire data related to customer buying habits, retention, analyzing customer segments, advertising effectiveness. Various ways are there to fetch the data in e-commerce. Web log files, cookies are primary sources of collecting data. Web server logs are used to record and accumulate users' interaction data [14]. It is quite difficult to fetch product details along with customer buying behavior. The web logs are generated mainly for the purpose of debugging the web server. So they are poor in information. The main purpose in this step is to acquire purchase data and path traversal patterns of past users which drive to find out navigation behavior of future users. According to number of visitors, web site owner can improve design and development to attract more visitors.

In e-commerce form data is a big source of data errors. Appropriate form validation can save a lot of time needed for data cleansing and later data analysis. For example data types can be validated include date, time, phone numbers, postal addresses and age. For domain data types, use drop down lists instead of free text fields. For example, use a drop down list containing "Male" and "Female" for gender.

4.2 Creation of Data Warehouse

A data warehouse can enhance business productivity because it is able to quickly and efficiently gather information that accurately describes the organization. It facilitates customer relationship management because it provides a consistent view of customers and items across all lines of business, all departments and all markets. The creation of a data warehouse requires significant data transformations from an operational system, sometimes called On-Line Transaction Processing (OLTP). 80% of time is spent in data transformation. The process of data warehouse requires data cleaning, data integration, data selection and data transformation. This is the process of knowledge discovery from the database [13].

1. Data Cleansing

The data cleansing task is to remove fields or attributes or variables which are irrelevant. In transactional system, customer record duplication problem arises. Mapping between customers and accounts is a many-to-many relationship (one customer may have multiple accounts or one account may be shared amongst multiple customers). Dirty data can cause confusion for the mining procedure, resulting in unreliable output.

2. Data Integration

Multiple heterogeneous data sources are combined in a common source [13]. This would involve integrating multiple databases or files. For example, the attribute for customer identification may be referred to as customer_id in one data store and cust_id in another. Large amount of redundant data may slow down or confuse the knowledge discovery process. In addition to data cleaning, steps must be taken to help avoid redundancies during data integration.

3. Data Selection

Relevant data is selected from data collection. We need to eliminate noise of initial data to produce efficient data.

4. Data Transformation

Data are transformed or consolidated into forms appropriate for mining. Data transformation can involve smoothing, aggregation, generalization, normalization, attribute construction.

4.3 Data Mining

Data mining process ideally consists of a set of functional modules for task such as clustering, classification, and association analysis. It is an essential process where intelligent methods are applied in order to extract data patterns. In steps 1 to 4, data are prepared for mining. Data mining system may interact with the user or knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

1. Clustering

Clustering is the method where classes are defined, cluster of objects are created those are somehow similar in characteristics. The objects of similar properties are placed in one class of objects and a single access to the disk make the entire class available. For example in library, various books are available. The books which are of some kind of similarities among them are placed in one cluster and then cupboards are labeled with the relative name. Clustering techniques can be used in many applications for example biological, financial and many more. Clustering is the methodology for more effective search and retrieval functions pertaining to data sets [8]. Once clusters are formed, we can match data records in the same data region accordingly. In e-commerce, the customers who have similar purchasing behavior form a cluster.

2. Classification

Classification is the process of assigning a web page to one or more predefined category labels. Classification methods are applied to web mining in order to construct efficient trees that classify web pages depending on there features [10]. For example users are divided into 5 kinds by their different network consumption behaviors: high-quality users, potential users, ordinary users and no visiting users. We can easily get classification rules, so we can guide and help different types of users to carry out consumptions. The customers can be classified based on purchase history and propensity to purchase. Multiple monthly campaigns are run based on this segmentation. Each campaign results in an insert for the monthly mailing sent to the customer.

3. Association Rule Mining

Finding frequent patterns, associations, correlations among sets of items or objects in transaction databases, relational databases and other information repositories. For example basket data analysis, if set of transactions are given, it finds rules that will predict the occurrence of an item based on the occurrence of other.

Table 2. Product Item Occurrence

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Examples

{Bread} => {Milk}

{Soda} => {Chips}

{Bread} => {Jam}

Product recommender model comes up with thousands or even millions of association rules. It is impossible for a human to manually go through even a small number of these rules.

4.4 Analysis Evaluation

Analysis evaluation presents technical lessons after analysis of data from clients.

This step includes -
Understanding and enriching the data,
Building models and identifying insights,
Deploying models,
Empowering business users to conduct their own analysis.

For example, visualization tools represented in graphical format are very useful in identifying interesting trends and patterns in the data. Customers are the center of many data analysis projects in retail e-commerce. Customer analysis process involves customer registration information, customer web visits including referrers, areas of the site visited, products viewed and purchased, average session length, visit frequency etc.

4.5 Knowledge Formulation

Visualization and knowledge representation techniques are used to present web mined knowledge to user.

The goal of web data mining exercise in e-commerce is to improve processes which help to deliver quality and value to target customers [12]. Attracting new customers, retaining existing customers and analyzing buyer behavior will definitely make progress in producing quality products, services and obviously gaining of profits.

Consider an example of online store like http://www.dell.com or http://www.compaq.com where if customer wants to configure a PC then she can configure of her choice, customer can place an order for the same and also pay for the product and services. The information available in web log files can predict what purpose customers are seeking from a site, are they really shopping or just doing window shopping or browsing? Are customers buying products in which they are familiar with or know litter about? From which location customers are shopping like from hotel, home or office? Once a customer has purchased a certain number of computers, they want to further purchase other peripherals like backup devices, internet connections etc. Association Rule mining can be used to propose such alternatives to the customers. Recommender system in e-commerce [3] automatically informs about important events of interest to them. The event prediction system is based on association rule mining and clustering algorithms. For example PENS system [3]. WDM in web personalization gives comprehensive overview of the personalization process based on web usage mining. Web usage mining requires preprocessing and integration of data from multiple sources. In e-commerce, 80% of WDM effort is in data filtering which rely on the web logs those are generated by the HTTP protocol. Problem Analysis in E-Commerce web data mining

1. To generate log files for millions of transactions is a costly exercise.
2. Many web data mining algorithms are unable to process the huge amount of data gathered at web sites in specified time.
3. Data models are very complicated for analysis.
4. Certain mechanism is essential to avoid noisy data automatically in order to apply data mining algorithm.

5. CONCLUSION

The Web Data Mining is the popular technology in the world. There are many research works going on in this field. In current study we have done survey on various applications where web data mining techniques are used. We have analyzed problems in each application study of web data mining. We discussed web data mining with respect to clustering, classification, association rule and how it works in e-commerce applications. The domain experts plays important role in web data mining. Implementation of data mining techniques is an important task in any domain. We have extended the study to design web data mining model on specified problem domain and soon it will be in practice.

6. REFERENCES

[1] *ezDataMiner and the Strategic Advantages of Data Mining.* Moheb A Kasem, Chris Bassell, Dean Amo, Andrew Jambor.

Central Connecticut State University, USA

[2] Web Data Mining Research: A Survey. Brijendra Singh, Hemant Kumar Singh. Lucknow, India

[3] A Web Data Mining Framework for E-commerce Recommender Systems. Jinhua Sun, Yanqi Xie, China

[4] A Short Introduction to Data Mining and Its Applications. Zhang Haiyang, China

[5] Web Data Mining for Web Intelligence. *Jiawei HanKevinChen-ChuanChang,* University of Illinois at Urbana-Champaign

[6] Web Mining : Research & Practice

[7] Design of Healthy Eating System based on Web Data Mining. 2010 WASE International Conference on Information Engineering, Xiaocheng Li, Xinliu, Zengjie Zhang, Yongming Xia, Songrong Qian, Shanghai, China

[8] Advanced Data Clustering Methods of Mining Web Documents. *Samuel Sambasivam-USA, Nick Theodosopoulo-UK,* Volume 3, 2006

[9] Data Extraction for Deep Web Using WordNet. Jer Lang Hong, PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 6, NOVEMBER 2011

[10] Data Mining Techniques for Web Page Classification. Gabriel Fiol-Roig, Margaret Miró-Julià, Eduardo Herraiz, spain

[11] Research on Web Data Mining and its Application in Electronic Commerce. GUANGCAN YU, CHUANLONG XIA, XINGYUE GUO, China.

[12] Data mining in e-commerce: A survey. N. R. Shrinivasa Raghavan. Department of Management Studies, Indian Institute of Science, Bangalore, Vol. 30, Parts 2 & 3, April/June 2005.

[13] Data mining system and applications: A review. Mr. S. P. Deshpande, Dr. V. M. Thakare, Department of MCA, Amravati, India, International Journal of Distributed and Parallel systems (IJDPS) Vol. 1, No. 1, September 2010.

[14] A research on web data mining and its application in electronic commerce. Guangcan Yu, Chauanlong Xia, Xingyue Guo. International school of software Wuhan University.

[15] Application of the data mining in the personalized information service. Xinying Gu, Zhimin Li. Hebei University of science and technology, ShiJiaZhuang, China.

[16] Application of data mining on students' quality evaluation. He Yongqiang, Zhang Shunli. Henan Institute of Engineering, Zhengzhouk, China.