International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249–8958, Volume-1 Issue-3, February 2012

Web Data Mining & Applications

Sita Gupta, Vinod Todwal

Abstract: ***With an enormous amount of data stored in databases and data warehouses, it is increasingly important to develop powerful tools for analysis of such data and mining interesting knowledge from it. Data mining is a process of inferring knowledge from such huge data. The main problem related to the retrieval of information from the World Wide Web is the enormous number of unstructured documents and resources, i.e., the difficulty of locating and tracking appropriate sources. In this article, a survey of the research in the area of web mining and suggest web mining categories and techniques. Furthermore, a presentation of a web mining environment generator that allows naive users to generate a web mining environment specific to a given domain by providing a set of specifications.***
Application of data mining techniques to the World Wide Web, referred to as Web mining, has been the focus of several recent research projects and papers. However, there is no established vocabulary, leading to confusion when comparing research efforts. The term Web mining has been used in two distinct ways. The first, called Web content mining in this paper,is the process of information discovery from sources across the World Wide Web.The second, called Web usage mining,is the process of mining for user browsing and access patterns. In this paper we define Web mining and present an overview of the various research is sues, techniques, and development efforts. We briefly describe WEBMINER, a system for Web usage mining, and conclude this paper by listing research issues.

Index Terms**: Data, Mining, Warehouse Web.**

I. INTRODUCTION

Data Mining:
Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses . Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions . The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because

it lies outside their expectations. Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Web Mining
The web as we all know is the SINGLE largest source of data available. Web mining aims to extract and mine useful knowledge from the web. It is used to understand the customer behavior, evaluate the effectiveness of a website and also to help quantify the success of a marketing campaign. Due to the large availability of data the world wide web, it has become very important for users to use automated tools to find the desired information resources.
For example a user uses Google or Yahoo search for finding information.These factors thus give rise to the necessity of creating server and client side intelligent systems which can effectively mine for knowledge.The information gathered through the Web is further evaluated by using traditional data mining techniques such as clustering, classification and association.

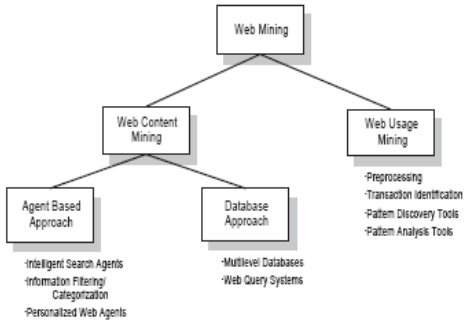

Figure 1: Taxonomy of Web Mining

II. DATA MINING TASKS

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

1. Exploratory Data Analysis: It is simply exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.

Manuscript published on 28 February 2012.
***** Correspondence Author (s)
First Sita Gupta*, Department of MCA., Rajasthan Technical University RCEW, Jaipur (Rajasthan), India, 9785404340, (e-mail: sita.todwal@gmail.com).
Second Vinod Todwal, Department of IT, Rajasthan Technical University, RCEW, Jaipur (Rajasthan), India, 9414607919, (e-mail: vinodtodwal@gmail.com).

2. Descriptive Modeling: It describe all the data, It includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.
3. Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.
4. Discovering Patterns and Rules: It concern with pattern detection, the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest.
5. Retrieval by Content: It is finding pattern similar to the pattern of interest in the data set. This task is most commonly used for text and image data sets.

III. TYPES OF DATA MINING SYSTEM

Data mining systems can be categorized according to various criteria the classification is as follows:

- Classification of data mining systems according to the type of data source mined: This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

- Classification of data mining systems according to the data model: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.

- Classification of data mining systems according to the kind of knowledge discovered: This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

- Classification of data mining systems according to mining techniques used: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

IV. DATA MINING LIFE CYCLE

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

1. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2 Data Understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. Data Preparation: It covers all activities to construct the final dataset from the initial raw data.

4. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5. Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

6. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

V. DATA MINING TECHNIQUES

There are several major ***data mining techniques*** have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine those data mining techniques with example to have a good overview of them.

ASSOCIATION

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

CLASSIFICATION

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group.

CLUSTERING

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

PREDICTION

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

SEQUENTIAL PATTERNS

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

VI. DATA MINING APPLICATION

The data mining applications can be generic or domain specific. The generic application is required to be an intelligent system that by its own can takes certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. Some generic data mining applications cannot take its own these decisions but guide users for selection of data,Selection of data mining method and for the interpretation of the results. The multi agent based data mining application has capability of automatic selection of data mining technique to be applied. The Multi Agent System used at different levels: First, at the level of concept hierarchy definition then at the result level to present the best adapted decision to the user. This decision is stored in knowledge Base to use in a later decision-making. Multi Agent System Tool used for generic data mining system development uses different agents to perform different tasks.

A multi-tier data mining system is proposed to enhance the performance of the data mining process. It has basic components like user interface, data mining services, data access services and the data. There are three different architectures presented for the data mining system namely One-tire, Two-tire and Three-tire architecture. Generic system required to integrates as many learning algorithms as possible and decides the most appropriate algorithm to use. CORBA (Common Object Request Broker Architecture) has features like: Integration of different applications coded in any programming language considerably easy. It allows reusability in a feasible way and finally it makes possible to build large and scalable system. The data mining system architecture based on CORBA is given by Object

Management Group has all characteristics to accomplish a distributed and object oriented computation. A datacentric focus and automated methodologies makes data mining accessible to nonexperts. The use of high-level interfaces can implement the automated methodologies that hide the data mining concepts away from the users. A data-centric design hides away all the

details of mining methodology and exposes them through high-level tasks that are goal-oriented.

These goal-oriented tasks are implemented using data-centric APIs. This design makes data mining task like other types of queries that users perform on the data.

In data mining better results could be obtained if large data is available. It leads to the merging and linking of local databases. A new data-mining architecture based on Internet technology addressed this problem.

The context factor plays vital role in the success of data mining. The importance and meaning of same data in the different context is different. A data in one context is very important may not be much important in other context. A context-aware data-mining framework filters useful and interesting context factors, and can produce accurate and precise prediction using those factors.The domain specific applications are focused to use the domain specific data and data mining algorithm that targeted for specific objective. The applications studied in this context are aimed to generate the specific knowledge. In the different domains the data generating sources generate different type of data. Data can be from a simple text, numbers to more complex audio-video data. To mine the patterns and thus knowledge from this data, different types of data mining algorithms are used. The collection and selection of context specific data and applying the data mining algorithm to generate the context specific knowledge is thus a skillful job. In manydomain specific data mining applications the domain experts plays vital role to mine useful knowledge.

In the identification of foreign-accented French the audio files were used and the best 20 data mining algorithms were applied the Logistic Regression model found the most robust algorithm than other algorithm.

In language research and language engineering many time extra linguistic information is needed about a text. A linguistic profile that contains large number of linguistic features can be generated from text file automatically using data mining.This technique found quite effective for authorship verification

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

and recognition. A profiling system using combination of lexical and syntactic features shows 97% accuracy in selecting correct author for the text. The linguistic profiling of text effectively used to control the quality of language and for the automatic language verification. This method verifies automatically the text is of native quality. The results show that language verification is indeed possible.

In medical science there is large scope for application of data mining. Diagnosis of diesis, health care, patient profiling and history generation etc. are the few examples. Mammography is the method used in breast cancer detection. Radiologists face lot of difficulties in detection of tumors. Computer-aided methods could assist medical staff and improve the accuracy of detection. The neural networks with back-propagation and association rule mining used for tumor classification in mammograms. The data mining effectively used in the diagnosis of lung abnormality that may be cancerous or benign. The data mining algorithms significantly reduce patient's risks and diagnosis costs. Using the prediction algorithms the observed prediction accuracy was 100% for 91.3% cases. The use of data mining in health care is the widely used application of data mining. The medical data is complex and difficult to analyze. A REMIND (Reliable Extraction and Meaningful Inference from Non-structured Data) system integrates the structured and unstructured clinical data in patient records to automatically create high quality structured clinical data. The high quality of structuring allows existing patient records to be mined to support guidelines compliance and to improve patient care.

Data mining in distance learning automatically generate useful information to enhance the learning process based on the vast amount of data generated by the tutors and student's interactions with web based distance-learning environment.The Data Mining Applications transfers the data into information and feedback to the e-learning environment. This solution transforms large amounts of useless data into an intelligent monitoring and recommendation system applied to the learning process.

In Web-based Education the data mining methods are used to improve courseware. The relationships are discovered among the usage data picked up during students' sessions. This knowledge is very useful for the teacher or the author of the course, who could decide what modifications will be the most appropriate to improve the effectiveness of the course.

Sports are ideal for application of data mining tools and techniques. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. Data mining can be used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. Data mining can be used for scouting, prediction of performance, selection of players, coaching and training and for the strategy planning. The data mining techniques are used to determine the best or the most optimal squad to represent a team in a team sport in a season, tour or game.

VII. CHALLENGES IN WEB MINING

Finding useful and relevant information. Creating knowledge from available information. As the coverage of information is very wide and diverse, personalization of the information is a tedious process. Learning customer and individual user patterns. Much of the web information is redundant, as the same piece of information or its variant appears in many pages. The web is noisy i.e. a page typically contains a mixture of many kinds of information like, main content, advertisements, copyright notice, navigation panels.

VIII. USES OF WEB MINING

This technology has enabled ecommerce to do personalized marketing, which eventually results in higher trade volumes. The predicting capability of the mining application can benefit the society by identifying criminal activities. The companies can establish better customer relationship by giving them exactly what they need.

Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers, they can save on production costs by utilizing the acquired insight of customer requirements.

They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer.

IX. WEB MINING APPLICATION

Personalization, Recommendation engines Web-commerce applications Intelligent web search Hypertext classification and Categorization Information/trend monitoring Analysis of online communities Improving the relationship between the website and the user Recommendations to modify the web site structure and content Web personalization.

Web mining applications in E-commerce and E-services is a new research direction in the area of web mining. Among all of the possible applications in web research, e-commerce and e-services have been identified as important domains for Web-mining techniques. Web-mining techniques also play an important role in e-commerce and eservices, proving to be useful tools for understanding how ecommerce and e-service Web sites and services are used. This book therefore collects new developments and high quality researches for the readers of this book to understand the topics of web mining applications in e-commerce and e-services as well as the state-of-the-arts in this area. The chapters in this book include web usage mining and user browsing behavior analysis, semantic web mining, web performance mining, web mining for users' need understanding, web mining for social network analysis and web mining for P2P services.

X. CONCUSION

The web offers prospecting and user relationship management opportunities that are limited only by the imagination. Data mining is a tool that can extract predictive information from large quantities of data, and is data driven. It uses mathematical and statistical calculations to uncover trends and correlations among the large quantities of data stored in a database. It is a blend of artificial intelligence technology, statistics, data warehousing, and machine learning. This data mining technology is becoming more and more popular, and is one of

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

the fastest growing technologies in information systems today. The future of data mining is wide open, and it will be exciting to see how far this technology will go.
The term Web mining has been used to refer to techniques that encompass a broad range of issues. However , while meaningful and attractive, this very broadness has caused Web mining to mean different things to different people , and there is a need to develop a common vocabulary. Towards this goal we proposed a definition of Web mining, and developed a taxonomy of the various ongoing efforts related to it.
Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge.

REFERENCES

2. Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, *Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.*

3. Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", *ISBN 0-471-66657-2*

4. *John Wiley & Sons, Inc, 2005*Han, J., Kamber, M. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann. Jain, A.K., Murty, M.N., Flynn, P.J. (1999) Data Clustering: A Review, *ACM Computing Surveys*, **31**,3:264-323.

5. Salton, G. (1989) *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading.

6. Salton, G., Wong, A., Yang C.S.A. (1975) Vector Space Model for Automatic Indexing,
7. *Communications of the ACM* , **18**: 613-620.

8. Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", *Pearson Education,New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.*