IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Web Mining – Data Mining Concepts, Applications, and Research Directions

Mrs. S. R. Kalaiselvi[1], S. Maheshwari[2], V. Shobana[3]

Assistant Professor, Department of Computer Science, Dr.N.G.P. Arts and Science College, Coimbatore[1,2,3]

Abstract: The prolific growth of web-based applications and the enormous amount of data involved therein led to the development of techniques for identifying patterns in the web data. Web mining refers to the application of data mining techniques to the World Wide Web. Web usage mining is the process of extracting useful information from web server logs based on the browsing and access patterns of the users. The information is especially valuable for business sites in order to achieve improved customer satisfaction. Based on the user's needs, Web Usage Mining discovers interesting usage patterns from web data in order to understand and better serve the needs of the web based application. Web Usage Mining is used to discover hidden patterns from weblogs. It consists of three phases like Preprocessing, pattern discovery and Pattern analysis. In this paper, we present each phase in detail, the process of extracting useful information from server log files and some of application areas of Web Usage Mining such as Education, Health, Human-computer interaction, and Social media.

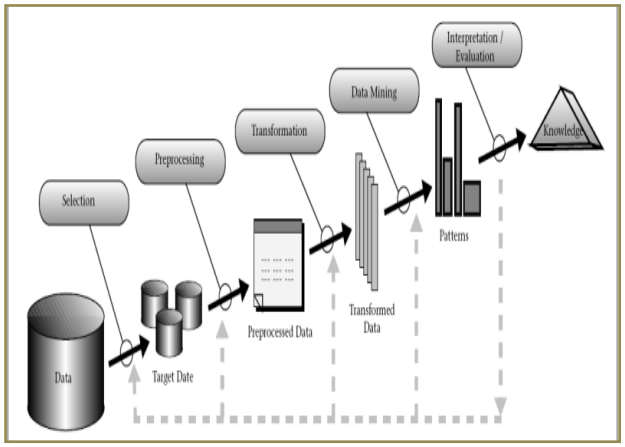
Keywords: Web Mining, Data Mining, World Wide Web, server log files.

1. INTRODUCTION

In customer relationship management (CRM), Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. (Mining means extracting something useful or valuable from a baser substance, such as mining gold from the earth.) Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign.

1.1 Overview of Data Mining

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.



Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are
- Exploration
- Pattern identification
- Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

Patterns Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired outcome.

1.2 Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, and Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

1.3 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:
- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

1.4 Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods
- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

1.5 Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods
- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

1.6 Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule
- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

1.7 Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of neural networks
- Back Propagation

2. DATA MINING APPLICATIONS

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Here is overview of business problems and solutions found using data mining technology.

3. FBTO DUTCH INSURANCE COMPANY

Challenges
- To reduce direct mail costs.
- Increase efficiency of marketing campaigns.
- Increase cross-selling to existing customers, using inbound channels such as the company's sell center and the internet a one year test of the solution's effectiveness.

DOI 10.17148/IJARCCE.2015.41188

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Results
- Provided the marketing team with the ability to predict the effectiveness of its campaigns.
- Increased the efficiency of marketing campaign creation, optimization, and execution.
- Decreased mailing costs by 35 percent.
- Increased conversion rates by 40 percent.

3.1 ECtel Ltd., Israel

Challenges
- Fraudulent activity in telecommunication services.

Results
- Significantly reduced telecommunications fraud for more than 150 telecommunication companies worldwide.
- Saved money by enabling real-time fraud detection.

3.2 Provident Financier's Home credit Division, United Kingdom

Challenges
- No system to detect and prevent fraud.

Results
- Reduced frequency and magnitude of agent and customer fraud.
- Saved money through early fraud detection.
- Saved investigator's time and increased prosecution rate.

3.3 Standard Life Mutual Financial Services Companies

Challenges
- Identify the key attributes of clients attracted to their mortgage offer.
- Cross sell Standard Life Bank products to the clients of other Standard Life companies.
- Develop a remortgage model which could be deployed on the group Web site to examine the profitability of the mortgage business being accepted by Standard Life Bank.

Results
- Built a propensity model for the Standard Life Bank mortgage offer identifying key customer types that can be applied across the whole group prospect pool.
- Discovered the key drivers for purchasing a remortgage product.
- Achieved, with the model, a nine times greater response than that achieved by the control group.
- Secured £33million (approx. $47 million) worth of mortgage application revenue

3.4 Shenandoah Life insurance company United States

Challenges
- Policy approval process was paper based and cumbersome.
- Routing of these paper copies to various departments, there was delays in approval.

Results
- Empowered management with current information on pending policies.
- Reduced the time required to issue certain policies by 20 percent.
- Improved underwriting and employee performance review processes.

3.5 Soft map Company Ltd., Tokyo

Challenges
- Customers had difficulty making hardware and software purchasing decisions, which was hindering online sales.

Results
- Page views increased 67 percent per month after the recommendation engine went live.
- Profits tripled in 2001, as sales increased 18 percent versus the same period in the previous year.

4. WEB MINING TECHNIQUES

Web Content Mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Graph base web Mining: The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Utilisation in web Mining: Web Utilised Mining is the application of database mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web based applications. Text Mining**:** Due to the continuous growth of the volumes of text data, automatic extraction of implicit previously unknown and potentially useful information becomes more necessary to properly utilize this vast source of knowledge. Text mining, therefore, corresponds to extension of the data mining approach to textual data and its concerned with various tasks, such as extraction of information implicitly contained in collection of documents or similarity- based structuring.

5. WEB USAGE MINING

5.1 Concept of web usage mining

5.1.1 Data accumulation:

Data accumulation is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality.

5.1.2 Data preprocessing:

Some databases are insufficient, inconsistent. The data pre-treatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

5.1.3 Data Cleaning

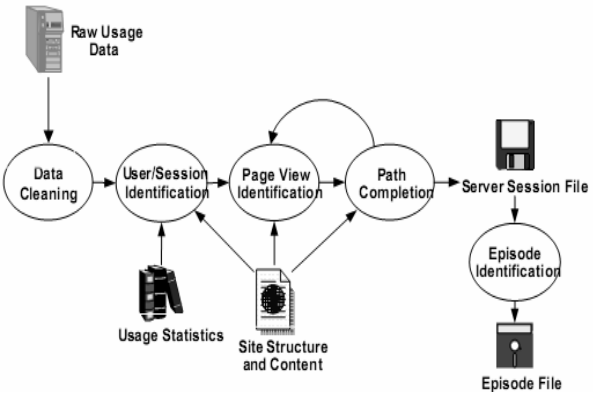
The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. 1. The records of graphics, videos and the format information. The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record. 2. The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or fewer than 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.



5.1.4 User and Session Identification

The task of user and session identification is find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions

5.1.4 Path completion

Another critical step in data pre-processing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators(URL) recorded in log may be less than the real one. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time.

Knowledge Discovery Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery.

Pattern analysis Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

6. PERSONALIZATON ON THE WEB MINING

Web personalization is a strategy of marketing tool, and largely art not a science. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. It is important to detail what it is you hope to do and, from that knowledge, develop an understanding of how you get from an idea to implementation. Web personalization can be seen as an interdisciplinary field that includes several research domains from user modeling, social networks, web data mining, human-machine interactions to Web usage mining; Web usage mining is an example of approach to extract log files containing information on user navigation in order to classify users. Other techniques of information retrieval are based on documents categories' selection Contextual information extraction on the user and/or materials (for adaptation systems) is a technique fairly used also include, in addition to user contextual information, contextual information of real-time interactions with the Web. Proposed a multi-agent system based on three layers: a user layer containing users' profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents. Propose reformulation query by adding implicit user information. Requests can also be enriched with predefined terms derived from user's profile develop a similar approach based on user categories and profiles inference. User profiles can be also used to enrich queries and to sort results at the user interface level. Other approaches also consider social-based filtering and collaborative filtering. For example, user queries can be enriched by adding new properties from the available domain ontology. User modeling by ontology can be coupled with dynamic update of user profile using results of information-filtering and Web usage mining techniques.

6.1 Results

In this paper in classification on web mining and concatenation with the web personalization and in that various method. Through web usage mining, performs six major task data gathering, data preparation, navigation pattern discovery, patter analysis, pattern visualization and pattern application. In the web mining there are we classified in terms web content mining, graph base web mining, utilization in web mining and text mining.

7. CONCLUSION

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

This article have outlined three different modes of web mining, namely web content mining, web structure mining and web usage mining Web usage mining model is a kind of mining to server logs. Web Usage Mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. We have presented in this paper the significance of introducing the web mining techniques in the area of web personalization. Content, usage, and structure data from different sources promise to lead to the next generation of intelligent Web applications. Use of artificial intelligence in these techniques should be next topic in this area of research.

REFERENCES

1. Agrawal R. and Srikant R. (2000). Privacy preserving data mining.Maier T. (2004).
2. A Formal Model of the ETL Process for OLAP-Based Web Usage Analysis. . Eirinaki M., Vazirgiannis M. (2003).
3. Web mining for web personalization. ACM Transactions On Internet Technology .Jiawei Han and Micheline Kamber (2006),
4. Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed..Dr. Gary Parker, vol 7, 2004,
5. Data Mining: Modules in emerging fields, CD-ROM.
6. Crisp-DM 1.0 Step by step Data Mining guide.