WEB MINING APPLICATIONS AND RESEARCH ISSUES

Gowthami Kumari G S[1] , Saritha M V[2] , Suneetha K R[3]
Department of Computer Science and Engineering,
Bangalore Institute of Technology, VV Pura, Bangalore, India

Abstract

Web mining is an application of data mining techniques to extract knowledge from web data. The paper concentrates on related work on mining and highlights drawbacks of existing system. This work provides information for beginners and also helps in view to improve performance of web site design.

Keywords: Web Mining, Web Usage Mining, Web log, Web Structures.

.1. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining is supported by three technologies like massive data collection, powerful multiprocessor computers and data mining algorithms for business applications. Data mining derives its name from the similarities between searching for valuable business information in large databases of sufficient size and quality; data mining technology can generate new business opportunities by providing automated prediction of trends and behaviors, automated discovery of previously unknown patterns.

The most commonly used data mining techniques are:

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Web Data Mining

Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web (WWW).

Web mining is divided into three categories:

i. **Web Content Mining:** is used to search, collate and examine data by search engine algorithms Web data contents: text, Image, audio, video, metadata and hyperlinks.

ii. **Web Structure Mining:** is used to examine the structure of a particular website and collate and analyze related data. Depending upon the hyperlink, ‘Categorizing the Web pages and the related Information at inter domain level discovers the Web Page Structure.

iii. **Web Usage Mining:** mining is used to examine data related to the client end, such as the profiles of the visitors of the website, the browser used, the specific time and period that the site was being surfed, the specific areas of interests of the visitors to the website, and related data from the form data submitted during web transactions and feedback. Discovering user 'navigation patterns' from web data. Prediction of user behavior while the user interacts with the web, helps to improve large collection of resources.

Web usage mining is used in pattern discovery, Content processing(conversion of Web information like text, images, scripts and others into useful forms) and Structure processing(analysis of the structure of each page contained in a Web site).

Analysis of this usage data will provide the companies with the information needed to provide an effective presence to their customers.

The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.

Application of data mining techniques to automatically discover and extract information from Web data. Web is the single largest data source in the world due to heterogeneity and lack of structure of web data, mining is a challenging task multidisciplinary field: data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc.

The contents of the paper is ordered as follows, section 2 refers categorization of mining section 2 briefs related work, section 3 explains about applications of web usage mining and section 5 provides conclusion.

2. RELATED WORK

This section reviews a set of papers related to applications of Web Mining. In paper [1] contains symbolic classification rules using neural networks. The effectiveness of the proposed approach is clearly demonstrated by the experimental results, in order to achieve performance efficiency. With the proposed approach, concise symbolic rules with high accuracy can be extracted from a neural network. The network is first trained to achieve the required accuracy rate. Redundant connections of the network are then removed by the algorithm. The activation values of the hidden units in the network are analyzed, and classification rules are generated using the result of this analysis. In this paper a neural network based approach proposed to mine classification rules from given databases. The work involves Constructing and training a network to classify, Pruning the network while maintaining the classification accuracy and *extracting* symbolic rules from the pruned network. A set of experiments were conducted using a well-defined set of data. The work proposed to generate rules similar to that of decision trees by reducing the training time of neural networks.

The paper [2] discusses on the knowledge discovery process, data mining, various open source tools and improvements in the field of data mining from past to the present and explores the future trends. The paper presents the knowledge Discovery Process, advantages, disadvantages and challenges of data mining, various open source tools and trends from its beginning to the future. The paper helps to do the research by focusing on the various issues of data mining. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains (finance, marketing, banking, insurance, health care and retailing). Data mining is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs, enhance research.

The authors Anand, Vaibhav, Prithi et al [3] discussed about large amount of data stored in databases using new techniques and tools. The research field Knowledge Discovery in Databases (KDD) or Data Mining. Attract attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization. Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable and predictive models from large-scale data. The paper overviews different tasks include in

Data mining, anomaly detection, classification, regression, association rule learning, summarization and clustering.

The Wei Fan, Andreas L. Prodromidis, and et al [4] proposed methods for combining multiple learned fraud detectors under a “cost model” demonstrate useful; The empirical results demonstrated in this work significantly reduces loss due to fraud through distributed data mining of fraud models. This process is automated, but it is unavoidable because the desired distribution highly depends on the cost model and the learning algorithm. . Unlike a monolithic approach of learning one classifier using incremental learning, the proposed modular multiclassifier approach facilitates adaptation over time and removes out-of-date knowledge. The paper focuses on growing problem of intrusion detection in network- and host-based computer systems. Here sort of task as in the credit card fraud domain and build models to distinguish between bad (intrusions or attacks) and good (normal) connections or processes by applying feature-extraction algorithms, followed by the application of machine-learning algorithms (such as Ripper).

The goal of the paper [5] to discuss about all the possibilities of identifying the way post documents in users’ navigation paths and to propose an optimization of navigation structure of a web site based on users navigation patterns. Web sites may contain numerous documents. Using some web techniques, it’s possible to analyze users’ data about using resources, contents of documents and structure of web sites.

The paper [6] implements a web usage Mining Intelligent System to provide taxonomy on user information based on transactional data by applying data mining algorithm, and also offers a public service that is direct access of websites to the third party.

In paper[7] the authors James .Andrews, et al discussed and apply a lightweight formal method for checking test results .The method assumes that the software under test writes a text log file, and is then analyzed by a program to check for failures. They suggested a state-machine-based log file analyzer programs, describe a language and implementation based on that formalism. Report the application of log file analysis to test the random units. They described the results of experiments are done to comparing the performance and effectiveness of random unit testing and checking log file analysis to other units testing procedures. Experimental results shows LFA that use of analyzer is cost effective and using LFA for complex specifications requiring scores of state machines or hundreds of transactions could be accepted only on safety-critical projects.

In paper [8] the authors Yiyao Lu, et al presented an atomic annotation approach that first aligns the data units on a result page into different groups that is same group have the same semantic data. Then each group can annotate it from different aspects and different aggregate to different annotations to predict a final annotation label for it. An annotation wrapper for search site is automatically constructed and it is used to annotate new result pages from the same web database. The experimental results show that each of annotators is useful and also capable of generating high quality annotation.

In paper [9] the Thient Thient Shwe, et al discusses on the mining of web access logs, web usage data. The framework composes defining the purpose that is multipurpose analyzer ,defining the ontology mapping based on web sites, perform preprocessing step, Web usage based mining on frequent item set and proposed the algorithm and Naïve Bayesian classifier. The framework is applied to predict for further depending on the current analysis outcomes. The paper also provides details for Web site maintenance, Personalization systems, Pre-fetched systems, Resource systems, Recommender system and web site analysts etc.

In paper [10] the authors Patrice Buche, et al presented the design of ONDINE system which allows loading and querying of a data warehouse, it is proposed by an Ontological and Terminological Resource (OTR). ONDINE system has been implemented through the development of @web system and development of MIEL++ software. The ONDINE is only software which allows one to simultaneously annotate accurately a data table with an OTR and perform the query process, comparing preferences expressed by the end-user with fuzzy annotations. ONDINE has been tested on three ways of applications are microbial risk in food, chemical risk in food and aeronautics.

In paper [11] the authors Sheng Di, Member et al proposes a fully distributed, Virtual Machine (VM)-multiplexing resource allocation scheme to manage decentralized resources. The approach presented here not only achieved maximize resource utilization using the proportional share model (PSM), but also delivers provably and adaptively optimal execution efficiency. The self-organizing cloud (SOC) architecture presented in this paper acts as both producer and consumer resource. They showed the SOC with optimized algorithms can make an improvement by 15-60% n system throughput than a P2P grid model. The solution also exhibits high adaptability in a dynamic node-churning environment. The novel scheme proposed in this work for virtual resource allocation on a SOC, with three contributions such that Optimization of tasks resource allocation under users budget, Maximized resource utilization based on PSM and Lightweight resource query protocol with low contention.

From the literature survey, the existing work suffers from following issues:-

i. The issues of individual privacy.
ii. Data issues and is integrating conflicting or redundant data from different sources.
iii. Technical issues on database structure.
iv. Expensive nature of Data mining Technologies.

3. APPLICATIONS OF DATA MINING

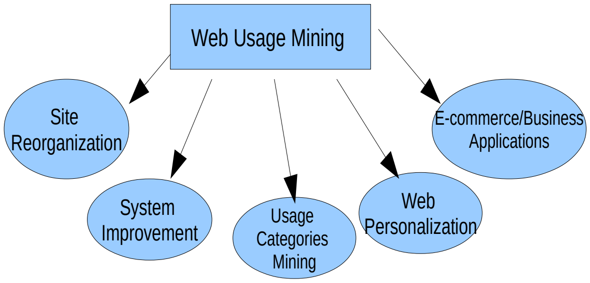

Figure1.Applications of web usage data mining

Applications of Data Mining is reflected in Figure 1.

a) Web personalization: It is defined as task of adopting services and information available on website and expectations available of a target user, active user. Web server log user used to cluster web users having similar interest.
b) System improvement: Web coaching, load balancing, network transmission are the common application areas of web mining for improving the system performance.
c) Site Reorganization: The link structure and content structure of any website are two significant factors on web site. The reorganization task is performed with respect to the frequent patterns extracted.
d) E-commerce/ business Applications: The uses of Web Usage Mining allow different organizations to understand and built customer profiles on the basis of customers perspective, their needs and interests so that companies can increase their profit.
e) Usage categorization: In this process the information stored in Web log servers is processed by applying various data mining techniques such as extracting statistical information and discovering interesting usage patterns, Clustering the users into groups according to their navigational behavior and determine possible links between web pages and user groups.

3.1 Web Usage Mining Techniques

Web usage mining is the “Applying data mining techniques to web data repositories to extract patterns”.

Data mining techniques that are commonly used includes Association rules, Sequential pattern, Clustering, and Classification.

a. Association rules are used to find the relationship between attributes from the item set. Rules are applied to discern pages which are often looked together, in order reveal associations between groups of users with specific interests.
b. Classification is a method that maps a data item one of several predefined classes. In web usages mining the users are in different classes according to their profiles.
c. Sequential pattern is used to discover navigational pattern for user session.
d. Clustering is a technique to group together items that have similar features. In Web usage domain, there are two clustering groups such as user and page clusters.
e. Web usage mining is also known as web log mining, web log file is log file automatically created and maintained by a web server. Every "hit" to the Web site, including each view of a HTML document, image or other object, is logged. The raw web log file format is essentially one line of

text for each hit to the web site. This contains information about who was visiting the site, where they came from, and exactly what they were doing on the web site.

3.2 Applications of Web Mining

With the rapid growth of World Wide Web, Web mining becomes a very hot and popular topic in Web research. Web mining plays an important role for E-commerce website and E-services to understand how their websites and services are used and to provide better services for their customers and users. It includes applications like E-commerce Customer Behavior Analysis, E-commerce Transaction Analysis, E-commerce Website Design, E-banking, M-commerce, Web Advertisement, Search Engine, Online Action etc.

4. CONCLUSIONS

As the web and its usage continues to grow, analysis of web data and extraction of knowledge plays important role in e-commerce. The paper reviews related work on mining and provides information for researchers in the area of web mining.

REFERENCES

1. Hongjun Lu, *Member, IEEE Computer Society*, Rudy Setiono, and Huan Liu, *Member, IEEE*, " Effective Data Mining Using Neural Networks*"*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 8, NO. 6, pp.957-961, DECEMBER 1996.
2. S.Hameetha Begum*," Data Mining Tools and Trends – An Overview", *International Journal of Emerging Research in Management &Technology ISSN: 2278-9359,pp.6-12*, February 2013.
3. Anand V. Saurkar, Vaibhav Bhujade, Priti Bhagat Amit Khaparde, " A Review Paper on Various Data Mining Techniques", Volume 4, Issue 4, ISSN: 2277 128X International Journal of Advanced Research in computer Science and Software Engineering,pp.98-101, April 2014.
4. Philip K. Chan, Florida Institute of Technology,Wei Fan, Andreas L. Prodromidis, and Salvatore J. Stolfo, Columbia University, " Distributed Data Mining in Credit Card Fraud Detection", 1094-7167/99/$10.00 © 1999 IEEE,pp.67-73.
5. Zeljko Eremic*, Dragical Radosav*, Branko Markoski*,"Mining user Logs to Optimize Navigational Structure of Adaptive Web Sites", 11th IEEE International Symposium on Computational Intelligence and Informatics,Budapest,Hungary, pp.271-275,18-20 November,2010.
6. B.Naveena Devi,Y.Rama Devi, B.Padmaja Rani, R.Rajeshwar rao,a, "Design and Implementation of Web Usage Mining Intelligence System in the field of e-commerce",International Conference on Communication technology and System design, pp.20-27,2012.
7. James H.Andrew, Member,IEE,and Yingjum Zhang, "General Test Result Checking with log File Analysis", IEEE TRANSACTION ON SOFTWARE ENGINEERING ,VOL 29, pp.634-648,July 2003.
8. Yiyo Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE , and Clement Yu, Senior member,IEEE. " Annotating Search Results from Web databases", IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, VOL.25, NO.3, pp 514-527, March 2013.
9. Thient Thient Shwe,Thient Myint,Thient Thient Aye, Su Su Htay, Swe Swe Nyein, Mie Mie Su Thwin, "FrameWork for Multi-Purpose Web log access Analyzer", IEEE 2nd International Conference on Computer Engineering and Technology,pp.v3.289-v3.293,2010.
10. Patrice buche,Julette Dibie-Barthelemy, Liliana Ibanescu , and lydie Soler, "Fuzzy Web data tables Integration Guided by an Ontological and Terminological resource " ,IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, VOL.25,NO.4, pp.805-814,April 2013.
11. Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE, and Clement Yu,Senior Member, IEEE, " Dynamic Optimization of Multiattribute Resource Allocation in Self-Organizing Clouds",IEEE TRANSACTION ON PARALLEL AND DISTRIBUTED SYSTEMS,VOL.24 ,NO.3 ,pp.464-476, March 2013.