Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/10193
DOI URL: http://dx.doi.org/10.21474/IJAR01/10193

RESEARCH ARTICLE

COMPARISON OF WEB - MINING TECHNIQUES - A SURVEY PAPER

Surender Singh[1] and Rajinder Singh[2]

1. Maharaja Surajmal Institute of Technology, IT Department, New Delhi - 110058.
2. HMR Institute of Technology and Management, CSE Department, New Delhi - 110036.

Manuscript Info

Manuscript History
Received: 08 October 2019
Final Accepted: 10 November 2019
Published: December 2019

Key words:-
Data Mining, Web Content Mining,Web Structure Mining

Abstract

Mining is a technique of extracting actionable information from the big pile of data. This paper surveys and compares the various web mining techniques which is an emerging field of data mining.

Introduction:-

Data collection and storage strategies have made it viable for different organizations to escalate large amount of data which is later used to find meaning full information or patterns from the useful or actionable data. In [1], the subsequent definition is given: Data mining (DM) is the method of exploration and analysis, by way of computerized or semiautomatic means, of massive quantities of facts (data) in order to find out significant patterns and rules. Data mining is a subfield of computer science which involves computational method of large data (information) sets patterns discovery. The techniques used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about solving problems by analyzing data already present in databases [2]. Recently, interest has risen in data mining because it finds useful knowledge hidden in a large amount of accumulated documents. However, it is difficult to find suitable tools for examining raw web log data to retrieve significant and useful information. Data mining tasks can be categorized in two categories-descriptive and predictive. Descriptive mining tasks distinguish the general properties of the data in database. Predictive mining tasks perform inference on the current data in order to make predictions.

Mining Algorithms:

Mining strategies (techniques) are used to discover the data available online after which extract the relevant information from the Internet. Searching on the web is a complicated process that requires exclusive algorithms for which we use one-of-a-kind techniques that are available::

Support Vector Machine (SVM):

SVM is a kernel-based learning algorithm. It was firstly implemented on classification problems and later applied for regression task. Primarily, vector machine employs the kernel trick for projection of non- linear separable training facts or data onto better dimensional feature (attribute) space by preserving dimensions of relatedness in the data. In a classification scenario it then obtains the maximum-margin hyperplane as the decision boundary pushed against by those support vectors and thus become capable of extracting the global optimal solutions regardless of the scarcity of the training data and less overfitted to it.

Corresponding Author:- Surender Singh
Address:- Maharaja Surajmal Institute of Technology, IT Department, New Delhi-110058.

Page Rank:
Page Rank is used is used when we want to rank pages in order of relevance. Some are page rank are based on structure and other based on content. Page rank algorithm gives different scores to different pages and sort in order of their pertinence. The web pages having higher page ranks are listed in the top and thus help the user in collecting required and important information in the least possible time.

KNN:
The KNN is the important classification technique when there is little or no prior knowledge about the distribution of the data. This rule simply keeps the entire training set during learning and assigns to each query a class represented by the majority label of its k-nearest neighbors' in the training set [4].

Apriori:
Apriori Algorithm is an algorithm based on association rules, which recite all of frequent item sets. This algorithm increases the efficiency of different organizations because it uses past knowledge of frequent item set properties which are used to predict different add on products with the product we want to purchase.

Neural Networks:
Neural Network is a machine that is designed to clone the way in which brain performs. In this numbers of processes are connected in a manner suggestive of the neurons working in human brain. The concept of neural networks is a sub part of machine learning i.e. the processes in neural network are able to learn and find errors by itself like a human.

Types of Neural Network:
1. Feed forward Neural Network : In this neural networks data flows in a single direction. The data passes from input nodes to output nodes. These neural networks may or may not have hidden layers. E.g.: Speech Recognition
2. Multilayer Neural Network : Multilayer networks solve the classification problem by using hidden layers, whose neurons are not directly connected to the output. Eg : Mechanical fault diagnosis
3. Kohonen Self Organizing Neural Network : This network is based on unsupervised learning to produce an input space of training samples. It assigns different weights to the neurons; one with the closest gets the higher preference. E.g.: Organizing massive data in real time
4. Recurrent Neural Network : The Recurrent Neural Network works on principle of saving the time for output layer be confirming and predicting the accurate result and feeding back to input layer which help in predicting the outcome of the layer. E.g.: Robotics

Introduction To Web Mining:
The application of data mining methods or techniques to World Wide Web is referred to as web mining [5]. Web mining techniques provides a set of strategies which provide solutions to dissimilar problems. However information retrieval (IR) and natural language processing (NLP) can also be used to handle these problems. When we see web mining in terms of data mining it have three interest of operations say clustering, associations and sequential analysis.

Web content mining:
The extraction of valuable information and web knowledge from web sources or web contents such as text, image, audio, video, and structured records is referred as web content mining [8]. Web content mining can be further categorized as web page content mining and search result mining [9]. Web content may be unstructured (plain text), semi- structure (HTML documents), or structured (extracted from databases into dynamic web pages). A research area closely related to content mining is text mining.

Fig. 1:- Categories of Web Mining

Web structure mining:
It refers the hyperlink's toppology within the web (inter document structure). It classifies the web pages and generates the facts such as the resemblance and association between them, taking the advantage of their hyperlink structure. Personalization is a classic application of Web Mining, which can be used to improve web site usage by modifying the contents of a web site with respect to the user or visitors necessity [9].

Web usage mining:
Web usage mining is the process of recognizing browsing patterns by analyzing the visitor's navigational behavior. It main focus is to extract useful and remarkable patterns from usage data such as server logs, client browser logs, proxy server logs, cookies, user sessions, registration data, mouse clicks, user queries, bookmarks etc. and any other data as the results of user interactions [13]. Web usage mining may be used to support dynamic structural modifications of an internet website so as to suit the active user, and to form recommendations. [10]

Application Of Web Mining:
Personalization :
It is type of information filtering method that pursues to predict the 'ratings' or preferences' that a user would give to an items, they have not been considered, using a model contructed from the characteristics of an item (content-based methods or collaborative filtering methods) [11]. These structures closely examine the individual characteristics and habits, extract useful patterns and constructs computerized responses to justify individual needs without expecting much input from visitor or user. This mined information is used to promote business, understanding market dynamics, new promotions, personalized ads etc [12]. "Web Personalizer" [14] is a powerful frame for mining web log files to find the beneficial information for the purpose of recommendations based on the browsing resemblances of present user to previous user.

Marketing intelligence:
Web mining has several benefits for companies. Firstly, to increase of profits by sale of more products or services and by reducing the costs. For doing this, marketing intelligence is necessary. It can concemntrate on marketing strategies and competitive analyses or on the association with the customers. Web Mining cab be used to categorization and clustering techniques to construct comprehensive customer profiles. It helps organizations in two ways, (i) to hold current customers by providing them more personalized services and (ii) contributes in the lookh for potential customers.

Comparison:-

S.NO.	PAPER TITLE	AUTHORS	ADVANTAGES	DISADVA NTAGES
1	Database Intrusion Detection using Weighted Sequence Mining [11]	Abhinav Srivastava, Shamik Sural and A.K. Majumdar	As per experiments performed, detection rate was better than non-weight IDS.	The proposed database intrusion detection system generates more rules as compared to non-weighted approach which makes it less efficient.
2	A Real-Time Intrusion Detection System using Data Mining Technique [14]	Fang-Yie Leu and Kai-Wei Hu	This paper introduced Intrusion Detection and Identfication System (IDIS), which builds a profile for each user in an intranet to keep track his/her usage habits as forensic features with which IDIS can identify the underlying user in the intranet.	Difficult to implement and requires large set of data for accuracy.
3	Multi-class Enhanced Image Mining of Heteroge-neous Textual Images Using	S.Chitrakala, P.Shaminim	Using the J48 Classifier and a decision tree classifier based on C4.5, testing the region features with a 10-fold cross validation, results came up with 92% of average accuracy	The intensity histogram features of mean, variance and skewness could efficiently classify DOC but were useless for other types of images i.e. ST (Scene Text Image) and CT (Caption Text

	Multiple Image Features [15]			Image).
4	Stock Market Prediction Based on Public Attentions: a Social Web Mining Approach [3]	Ailun Yi	Models built on more complex features such as Loose n-gram and topic filtering and wrapper method bring cross related concepts and are effective in prediction	The simplest counting method failed to correlate directly with stock prices and such methods cannot lead to stable perfor-mance in prediction.
5	Mining the web for generating Thematic Metadata from Textual Data [12]	Chien chung Huang, Shui-Lung Chuang, Lee-Feng Chien	Experimental results confirm the potential and wide adaptability of approach of using feature extraction, HCQF and text instance categorization to generate semantically- deep meta data.	Quality of meta data is not guaranteed and differe depending upon the websites being mined.
6	Intrusion detection Using data mining along fuzzy logic and genetic algorithms [13]	Y.Dhanal-akshmi, Dr.I. Ramesh Babu	The fuzzy logic system based genetic algorithms potentially detects correct timing for most of IDS attacks.	Large data sets are needed, otherwise genetic algorithms will select almost all rules
7	ADAM: A Testbed for Exploring the Use of Data Mining in Intrusion Detection [19]	Daniel Barbard, Julia Couto, Sushil Jajodia, Ningning Wu	Accurately prevents against a variety of attacks such as DOS attack.	This method allows us to avoid the dependency on training data for attacks, ADAM still requires some training data to build the profile of normal activity.
8	Data Mining Techniques for (network) Intrusion Detection Systems [20]	Theodoros Lappas and Konstanti- nos Pelechrinis	Here, a survey of the various data mining techniques that have been proposed towards the enhancement of IDSs and has shown the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated by researchers	The paper provides a potential idea and is not tried and tested.
9	Customer Information System for Product and Service Manage-ment: Towards Knowledge Extraction from Textual and Mixed-Format Data [16]	Si Jie Phua , Wee Keong Ng , Haifeng Liu , Xiang Li , Bin Song	This paper proposes to develop a Customer Information System to efficiently extract customer knowledge from structured and textual data using computa- tional intelligence approaches. The system assists companies to identify profitable market and discover customer needs pattern.	It takes time to extract meaning-fulfindings from data but enables innovations that areessential for product and service manage-ment.

10	Text Mining and Expert Systems applied in Labor Laws [17]	Antonio Alexandre Mello Ticom, Beatriz de Souza Leite, P. de Lima	The paper showed that Text Mining is a promising research area to ap- plications in the judicial field which presents immense volumes of documents and un- structured data that need to be processed.	The application is capable of identifying more parameters from the judicial sentences. Presently, use less parameters.
11	Creation and Deployment of Data Mining-Based Intrusion Detection Systems in Oracle Database 10g [18]	Marcos M. Campos, Boriana L. Milenova	Database-centric IDSs offer many advantages over alternative systems. These include tight integration of individual compo- nents, security, scalability, and high availability.	Building an IDS is a complex task of knowledge engineering that requires an elaborate infrastruc- ture. An effective contempo- rary production-quality IDS needs an array of diverse compo- nents and features.

Conlusions:-

The paper focuses on surveying the concepts of web mining and its possible applications in everyday life. Several real-life applications such as Personalization, marketing intelligence etc. were studied. As a result of the survey, we were also able to compare various techniques in terms of advantages and disadvantages. The survey paper is a stepping stone in implementing either of the existing applications in the field or to generate a solution for an existing problem.

Refrences:-

1. Xingquan Zhu, Ian Davidson, Knowledge Discovery and Data Mining: Challenges and Realities, ISBN 978-1-59904-252, Hershey, New York, 2007.
2. Joseph, Zernik, Data Mining as a Civic Duty Online Public Prisoners Registration Systems, International Journal on Social Media: Monitor- ing, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.
3. Ailun Yi, Stock Market Prediction Based on Public Attentions: a Social Web Mining Approach-Univ of Edinburgh (2009).
4. Sadegh Bafandeh Imandoust And Mohammad Bolandraftar, Applica- tion of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background- S B Imandoust et al. Int. Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep- Oct 2013, pp.605-610
5. Er.Romil.V.Patel, Dheeraj Kumar Singh, Mr.Ankur.N.Shah, Introduc- tion to Integrating Web Mining With Neural Network , IRACST - In- ternational Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.6, December 2012
6. Vaishali A.Zilpe, Dr. Mohammad Atique, WEB USAGE MINING USING NEURAL NETWORK APPROACH: A CRITICAL REVIEW , Vaishali A.Zilpe et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012, 3073 3077
7. S.Jagan, Dr.S.P.Rajagopalan, A Survey on Web Personalization of Web Usage Mining, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 02 Issue: 01 — March-2015 ISSN: 2395-0072
8. Raymond Kosala and Hendrik Blockeel, Web mining research: A survey, SIGKDD Explorations, pages 95-104, July 2000.
9. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pangning Tan, Web usage mining: Discovery and applicationsof usage patterns from web data, ACM SIGKDD Explorations, 01(03):187-192, January 2000.
10. Mobasher B., Cooley R., and Srivastava J, Automatic personalization based on web usage mining, ACM Communication, 43(08):142-151, August 2000.
11. Srivastava, A., Sural, S. and Majumdar, A.K., 2006. Database intrusion detection using weighted sequence mining. Journal of Computers, 1(4), pp.8-17.
12. Huang, C.C., Chuang, S.L. and Chien, L.F., 2004, April. Mining the Web for generating thematic metadata from textual data. In Proceedings. 20th International Conference on Data Engineering (p. 834). IEEE.

13. Y.Dhanalakshmi, Dr.I. Ramesh Babu, Intrusion Detection Using Data Mining Along Fuzzy Logic and Genetic Algorithms, IJCSNS Inter- national Journal of Computer Science and Network Security, VOL.8 No.2, February 2008.
14. Fang-Yie Leu, Kai-Wei Hu, A Real-Time Intrusion Detection System using Data Mining Technique
15. S.Chitrakala , P.Shamini, Dr.D.Manjula, Multi-class Enhanced Image Mining of Heterogeneous Textual Images Using Multiple Image Features , 2009 IEEE International dvance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.
16. Si Jie Phua, Wee Keong Ng, Haifeng Liu, Xiang Li, Bin Song, Customer Information System for Product and Service Management: Towards Knowledge Extraction from Textual and Mixed-Format Data
17. Antonio Alexandre Mello Ticom, Beatriz de Souza Leite P. de Lima, Text Mining and Expert Systems applied in Labor Laws, Seventh Inter- national Conference on Intelligent Systems Design and Applications.
18. Campos, M.M. and Milenova, B.L., 2005, December. Creation and deployment of data mining-based intrusion detection systems in oracle database l0g. In Fourth International Conference on Machine Learning and Applications (ICMLA'05) (pp. 8-pp). IEEE.
19. Barbará, D., Couto, J., Jajodia, S. and Wu, N., 2001. ADAM: a testbed for exploring the use of data mining in intrusion detection. ACM Sigmod Record, 30(4), pp.15-24.
20. Lappas, T. and Pelechrinis, K., 2007. Data mining techniques for (network) intrusion detection systems. Department of Computer Science and Engineering UC Riverside, Riverside CA, 92521.