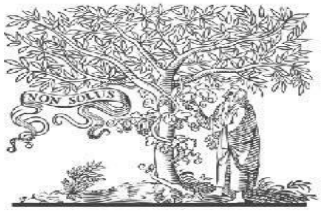
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 26th Nov 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 11

10.48047/IJIEMR/V11/ISSUE 11/57

TITLE: IMPACT OF DATA MINING TECHNIQUES TOWARDS INDUSTRIAL APPLICATION
Volume 11, ISSUE 11, Pages: 454-462

Paper Authors **SP. VENKATA RAMANA,Dr. Mukesh Kumar**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

IMPACT OF DATA MINING TECHNIQUES TOWARDS INDUSTRIAL APPLICATION

SP. VENKATA RAMANA
Research Scholar Monad University, Delhi Hapur Road Village & Post Kastla, Kasmabad, Pilkhuwa, Uttar Pradesh
Dr. Mukesh Kumar
Research Supervisor Monad University, Delhi Hapur Road Village & Post Kastla, Kasmabad, Pilkhuwa, Uttar Pradesh

ABSTRACT
Data mining has recently become one of the most progressive and promising fields for the extraction and manipulation of data to produce useful information. Thousands of businesses are using data mining applications every day in order to manipulate, identify, and extract useful information from the records stored in their databases, data repositories, and data warehouses. With this kind of information, companies have been able to improve their businesses by applying the patterns, relationships, and trends that have lain hidden or undiscovered within colossal amounts of data. For example, data mining has produced information that enables companies to create profiles of current and prospective customers to help in gaining and retaining their customers. Other uses of data mining include development of cross selling and marketing strategies, exposure of possible crimes or frauds, finding patterns in the access of users to their web sites, and process improvement. The power of data mining is yet to be fully exploited by industry. Manufacturing, for example, is one of the new fields in which data mining tools and techniques are beginning to be used successfully. Process optimization, job shop scheduling, quality control, and human factors are some of the areas in which data mining tools such as neural networks, genetic algorithms, decision trees, and data visualization can be implemented with great results.

KEYWORDS: Data Mining, Application, Industrial Engineering, data warehouses

INTRODUCTION

The presentation of the application of the proposed methodology for the analysis of users' RSS address file of the RSS reader website was showcased. We have developed an online, Real-Time recommendation expert system that can assist the web designer and administrator to improve the content, presentation and impressiveness of their website by recommending a unique set of objects that satisfies the need of active user based on the user's current click stream. Data mining techniques have various applications in e-learning. Data mining techniques can transform the learning processes as per the perception of the learner, their need; learning styles and learning approaches and can assist instructors in delivering a suitable content, so that the desired level of learning is achieved. Data mining is often described as the process of discovering correlations,

patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. The kinds of relationships that exist are believed to be sometimes unclear to information analysts because the amounts of information are too large or the kinds of relationships are too difficult to imagine. Humans, in that sense, are limited by information overload; thus, new tools and techniques are being developed to solve this problem through automation.

In the last few years, users began to realize the need for more tools and techniques in order to identify and find relationships in data so that the information obtained was more meaningful for their applications. Additionally, companies recognized that they had accumulated volumes of data; and, as a result, they needed new tools to sort through it all and meet their informational needs. Such tools enabled the system to search for possible hidden relationships in the data, without the direct intervention of the end users. Data mining tools were first developed to help scientists find meaningful relationships or patterns from huge amounts of data that, if done in a traditional way, would require much time and many resources to find. The next step is to exploit these tools for meaningful applications.

Data mining uses a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining must also be considered as an iterative process that requires goals and objectives to be specified. Once the intended goals are completely defined, it is necessary to determine what data is available or can be collected. Sometimes the data is available in data warehouses, but before it can be used, some filtering is performed to transform it into information.

Data mining also involves a methodology for implementation. The methodology, or structured approach, usually varies from vendor to vendor. SAS Institute, for example, promotes SEMMA (sample, explore, modify, model and assess). Another methodology is CRISP-DM by SPSS, Inc. Each methodology strives to help users obtain the best data to provide the most responsive information to address their needs. The recognition that effective decisions are based on appropriate information from accurate and current data is not new.

LITERATURE REVIEW

Siguenza-Guzman, L., Van Den Abbeele, A. (2020) It transforms the input into a high m-dimensional vector space then utilizes the algorithm to build the classifier finding the best linear separating hyper plane Latkowski and Osowski combined SVM with other algorithms in an ensemble to select the most significant genes in the expression microarray of autism. Presented an approach using SVM for predicting the impact of publishing individual posts on a social media network of company's page the predictive knowledge could support manager's decisions on whether to publish each post

Zahra Zamani Alavijeh, Isfahan, Iran, (2020) Sentiment analysis techniques have been well-studied in different domain, for

example stock prediction in Accounting/Finance, online product sales in marketing, and corporate reputation in corporate governance, etc. stated, sentiment analysis is a task of judging the opinion(positive or negative) which is transforming unstructured qualitative data into quantitative data that can be used for decision making, for example the reviews of customers about products and services(document, sentence, paragraph, etc.)

Esmaeil Fakhimi gheshlagh mohammad beig, (2019) many educational experts believe that the delivery of basic course concepts and materials through e -learning could potentially free up faculty's time so that they could provide mo re direct support to students. Additionally, some researchers also predict that students would learn more effectively due to the provision of instantaneous feedback and the self-paced learning style.

Smita Bhanap, Dr. Seema Kawthekar, (2019) An attempt to integrate IS model and TAM to evaluate E-learning success in developing countries ended up in a research framework linking the drives of E-Learning to its outcomes. The model proposed that education system quality, service quality, technology system quality, information quality preserved ease of use and perceive usefulness as the drives of Elearning that affect satisfaction and intention. Learning assistance and actual use are the E-learning outcomes. The empirical validation proves most of the hypothesis to be true as the path relationships holds good as in most of the other E-leaning studies. One important implication of this study is that enhancing user awareness of the system capabilities would increase the acceptance of educational technologies in the developing countries.

Pooja Rohilla, Ochin Sharma, (2020) has proposed a cancer prediction system based on data mining techniques. This system estimates the risk of the breast cancer in the earlier stage. The system is validated by comparing its predicted results with patient's prior medical information. The main aim of this model is to provide the earlier warning to the users and it is also cost efficient to the user. A prediction system is developed to analyze risk levels which help in prognosis. This research helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming.

Hilal Ahmad Khanday, Dr. Rana Hashmy, (2018) suggested a feature selection method and two feature extraction methods for K-Means clustering. The first feature extraction method is based on random projections. The second feature extraction method is based on fast approximate SVD factorizations. Both the feature extraction and feature selection methods provided better results and that too in less time as compared to the original K Means algorithm.

DATA MINING TECHNIQUES

There are a number of techniques used in data mining, but not all of them can be applied to all types of data. Neural network algorithms, for example, can be used to quantify data (numerical data), but they cannot qualify data precisely (categorical

data); therefore, categorical data is usually broken up into multiple dichotomous variables, each of them with values of 1 (“yes”)or 0 (“no”). For that reason, one single technique cannot be used to perform a complete data mining study and each technique has its own scope of applications. Some of the techniques applied in data mining include traditional statistics, induction, neural networks, and data visualization. These are described in the following sections.

- **Traditional Statistics**

Some of the traditional statistical methods that can be used for data mining are the following

- Cluster analysis, also called segmentation.
- Discriminate analysis.
- Logistic regression.
- Time series forecasting.

Cluster analysis (or segmentation) is one of the most frequently used data mining techniques; it involves separating sets of data into groups that include a series of consistent patterns. After the data reveals a consistent pattern, it is then sorted into subsets that are easier to analyze. This information is also used to identify subgroups of a population for supplementary studies, as well as to generate profiles for target marketing. Kohonen feature maps and K-means are some of the most important algorithms applied for cluster analysis.

Discriminant analysis is one of the oldest classification techniques. It finds hyper planes that separate classes so that users can then apply them to determine the side of the hyper plane in which to catalogue the data. Discriminate analysis has limitations, however. It assumes that all predictor variables are normally distributed--but this is not always true. Moreover, unordered categorical values cannot be classified, and boundaries are restricted to linear forms. New versions of discriminant analysis are being developed to handle these limitations by using quadratic boundaries, estimates of real distributions, and bins defined by the categorical variables.

STEPS IN PERFORMING WEB USAGE DATA MINING TASK

Data mining task can be categorized into different stages based on the objective of the individual analyzing the data.

The overview of the task for each step is presented in detail in four subsections as follows:

Data acquisition, preprocessing and data mart development

Data acquisition:

This refers to the collection of data for mining purpose, and this is usually the first task in web mining application. The said data can be collected from three main sources which includes

i) web server
ii) proxy server and
iii) Web client.

In this study, the web server source was chosen for the fact that it is the richest and most common data source; more so, it can be used to collect large amount of information from the log files and databases they represent. The user profile information, the access and navigation pattern or model are extracted from the historical access data

recorded in the RSS reader site, users' address database. The data are so voluminous as it contains so many detailed information such as date, time in which activities occur, saver's name, IP address, user name, password, dailies name, required feed, news headlines, and contents, as recorded in the database file.

Data pre-processing:

In the original database file extracted, not all the information are valid for web usage data mining, we only need entries that contain relevant information. The original file is usually made up of text files that contain large volume of information concerning queries made to the web server in which in most instance contains irrelevant, incomplete and misleading information for mining purpose. Researcher, described data preprocessing as the cleansing, formatting and grouping of web log files into meaningful session for the sole aim of utilizing it for web usage mining.

Data cleansing:

Data cleansing is the stage in which irrelevant/noisy entries are eliminated from the log file. For this work the following operations were carried out:

(i) Removal of entries with "Error" or "Failure" status.

(ii) Removal of requests executed by automated programs such as some access records that are automatically generated by the search engine agent from access log file and proxies.

(iii) Identification and removal of request for picture files associated with request for a page and request include Java scripts (.js), and style sheet file

(iv) Removal of entries with unsuccessful HTTP status code, etc.

Data mart development:

Two crown corporation explained that data mart is a logical subset of data warehouse. If the data warehouse DBMS can support more resources, that will be required of the data mining operation, otherwise a separate data mining database will be required. Since the raw log file is usually not a good starting point for data mining operation, the development of a data mart of log data is required for the data mining operation. In this work a separate data mart of users' RSS address URL was developed using relational database Management software MySQL.

Transaction identification

There is need for a mechanism to distinguish different users so as to analyze user's access behavior. Transaction identification is meant to create meaningful clusters of references for each user. Researcher, stated that a user navigation behavior can be represented as a series of click operations by the user in time sequence, usually call click stream, which can further be divided into units of click descriptions usually referred to as session or visit.

- **Session identification:**

According to him a session can be described as a group of activities carried out by a user from the user's entrance into the web site up to the time the user left the site. It is a collection of user clicks to a single web server. Session identification is the process of partitioning the log entries into sessions

after data cleansing operation. In order to achieve this researcher, suggested the use of cookies to identify individual users, so as to get a series of clicks within a time interval for an identified user. One session can be made up of two clicks, if the time interval between them is less than a specific period.

Pattern discovery

Pattern discovery is the key process of web mining which includes grouping of users based on similarities in their profile and search behavior. There are different web usage data mining techniques and algorithms that can be adopted for pattern discovery and recommendation, which includes path analysis, clustering, and associate rule. In our work, we have experimented with the K-Nearest Neighbor classification technique as described in Section in order to observe and analyze user behavior pattern and click stream from the pre-process to web log stage and to recommend a unique set of object that satisfies the need of an active user, based on the users' current click stream.

Pattern analysis

Pattern analysis is the final stage in web usage mining which is aimed at extracting interesting rules, pattern or statistics from the result of pattern discovery phase, by eliminating irrelevant rules or statistics. The pattern analysis stage provides the tool for the transformation of information into knowledge. We have incorporated an SQL language to develop a data mart using MySQL DBMS software specifically created for web usage mining purpose in order to store the result of our work. The data mart is populated from raw users RSS address URL file of the RSS reader's site that contains some basic fields needed; our experiment result is presented in Section.

Our approach

The problem at hand is a classification problem; therefore the K-Nearest Neighbor method of data mining is ideal. The objective of the system is to create a mapping, a model or hypothesis between a given set of documents and class label. This mapping was later to be used to determine the class of a given Test (unknown or unlabeled) documents. The K-Nearest Neighbor model is the simplest and most straightforward for class prediction, it is the most popular similarity or distance based text and web usage classification and recommendation model.

THE WORKING OF K-NEAREST NEIGHBOR CLASSIFIER

The K-Nearest Neighbor classifier usually applies either the Euclidean distance or the cosine similarity between the training tuples and the test tuple but, for the purpose of this research work, the Euclidean distance approach will be applied in implementing the K-NN model for our recommendation system.

In our experiment, suppose our data tuples are restricted to a user or visitor/client described by the attribute Daily Name, Daily Type and News category and that X is a client with Dayo as username and Dy123 as password.

The Euclidean distance between a training tuple and a test tuple can be derived as follows:

Let Xi be an input tuple with p features (xi1, xi2, …, xip)

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Let n be the total number of input tuples (i = 1, 2, …, n)

Let p be the total number of features (j = 1, 2, …, p)

The Euclidean distance between Tuple Xi and Xt(t = 1, 2, …, n) can be defined as

$$d(x_i,x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})}$$

(3.1)

In general term, The Euclidean distance between two Tuples for instance X1 = (x11, x12, …, x1n) and X2 = (x21, x22, …, x2n) will be,

$$\text{dist}(x_1,x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

(3.2)

Eq. (3.2) is applicable to numeric attribute, in which we take the difference between each corresponding values of attributes tuple x1 and x2, square the result and add them all together then get the square root of the accumulated result this gives us the distance between the two points x_1 and x_2 . In order to prevent attributes with initially large ranges from outweighing attributes with initial smaller ranges, there is a need to normalize values of each attributes before applying Eq. (3.2).

The min–max normalization can be applied to transform for instance value V of a numeric attribute A to V1 in the range [0,1] by using the expression

$$V^1 = \frac{V - \min A}{\max A - \min A} \quad (3.3)$$

minA and maxA are attribute A, minimum and maximum values .

In K-NN, classification, all neighboring points that are nearest to the test tuple are encapsulated and recommendation is made based on the closest distance to the test tuple, this can be defined as follows:

Let C be the predicted class

$$C_i = \{x \in C_p;\, d(x,x_i) \leq d(x,x_m), i \# m\}$$

(3.4)

The nearest tuple is determined by the closest distance to the test tuple. The K-NN rule is to assign to a test tuple the majority category label of its K-Nearest training tuple.

DATA PREPROCESSING

Input: Dataset is collected from webserver log file

Output : Pre-processed Data set

Steps

- Upload the dataset into server.
- Check the missing values.
- If the missing value occurred for Search class then replace the value as ‖general‖.
- Eliminate the records that have http status code greater than 400 and less than 200

Grouping Similar Users

Input: Pre-processed data set

Output: clustered result data set

Steps

- The pre-processed dataset is taken as input.
- Search category clusters will be formed based on count.
- If the count value is one, then the record will be eliminated

Steps for Improving KNN

- Determine Parameter K, where K is the number of nearest neighbours
- Calculate the distance between the query and all the training examples

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

- Sort the distance and determine nearest neighbour based on the k-th minimum distance
- Gather the category Y of the nearest neighbours
- Use simple majority of the category of nearest neighbours as the Prediction value of the query distance is calculated as in equation (3.5)

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(\omega_i^2)(a_i(x)-a_i(y))^2} \quad \textbf{(3.5)}$$

where

ϕ-weight of the attribute

Justification for using KNN algorithm over other existing algorithm

The K-Nearest Neighbor (K-NN) algorithm is one of the simplest methods for solving classification problems; it often yields competitive results and has significant advantages over several other data mining methods. Our work is therefore based on the need to establish a flexible, transparent, consistent straightforward, simple to understand and easy to implement approach. This is achieved through the application of K-Nearest Neighbor technique, which we have tested and proved to be able to overcome some of the problems associated with other available algorithms. It is able to achieve these by the following:

- Overcoming scalability problem common to many existing data mining methods such as decision tree technique, through its capability in handling training data that are too large to fit in memory.
- The use of simple Euclidean distance to measure the similarities between training tuples and the test tuples in the absence of prior knowledge about distribution of data, therefore makes its implementation easy
- Reducing error rate caused by inaccuracy in assumptions made for usage of other technique such as the Naïve Bayesian classification technique, such as class conditional independency and the lack of available probability data which is usually not the case when using KNN method.
- Providing a faster and more accurate recommendation to the client with desirable qualities as a result of straightforward application of similarity or distance for the purpose of classification.

CONCLUSION

The system performs classification of users on the simulated active sessions extracted from testing sessions by collecting active users' click stream and matches this with similar class in the data mart, so as to generate a set of recommendations to the client in a Real-Time basis. The result of our experiment shows that an automatic Real-Time recommendation engine powered by K-NN classification model implemented with Euclidean distance method is capable of producing useful and a quite good and accurate classifications and recommendations to the client at any time based on his immediate requirement rather

than information based on his previous visit to the site.

REFERENCES

Siguenza-Guzman, L., Van Den Abbeele, A., Vandewalle, J., Verhaaren, H., & Cattrysse, D. (2020). A holistic approach to supporting academic libraries in resource allocation processes. The Library Quarterly: Information, Community, Policy, 85 (3).

Zahra Zamani Alavijeh, Isfahan, Iran, (2020) "The Application of Link Mining in Social Network Analysis", ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 3, No.15

Esmaeil Fakhimi gheshlagh mohammad beig, (2019) "Data Mining Techniques for Web Mining: A Review", Applied mathematics in Engineering, Management and Technology 3(5):81-90.

Smita Bhanap, Dr. Seema Kawthekar, (2019) "Data Mining for Business Intelligence in Social Network: A survey", International Advanced Research Journal in Science, Engineering and Technology Vol. 2, Issue 12

Pooja Rohilla, Ochin Sharma, (2020) "Web Content Mining: An Implementation on Social Websites", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 7.

Hilal Ahmad Khanday, Dr. Rana Hashmy, (2018) "Exploring Different Aspects of Social Network Analysis Using Web Mining Techniques", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 4, Issue 2.

Faris Kateb,Jugal Kalita, (2018) "Classifying Short Text in Social Media: Twitter as Case Study", International Journal of Computer Applications (0975 8887) Volume 111 - No. 9.

Ritu,Ajit Singh,Akash Srivastava, (2019) "Opinion Mining Techniques on Social Media Data", International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 6.

Remya R S, Smitha E S, (2020) "Text Categorization using Data Mining Technique on Social Media Data", International Journal of Advanced Research in Education & Technology (IJARET), Vol. 2, Issue 4

Kuldeep Singh Rathore, Sanjiv Sharma, (2020) "A Review on Web Usage Mining For Web Personalization Using Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue.