Log-mining using machine learning algorithms And Visualization.

Mari Kirthima[1], Praveen Karkada[2] and Gopinath Ramanna[3]
1.Assistant Professor ,Department of CSE,BMSIT&M ,Bangalore,E-mail-krithi.a@bmsit.in.
2. M.Tech(PG),Dept. of CSE, BMSIT&M,Bangalore.
3. M.Tech(PG),Dept. of CSE, BMSIT&M,Bangalore.

Abstract—Data is considered to be an asset to any organization, the world has now become more data centric as terabytes of data gets generated every day and moreover the data is highly unstructured and contains all the information, but all of that data is not important for us mining required information from the whole lot is called data mining; similarly some product based companies manufacture handheld devices which runs on different operating system, any application running on the device puts up messages for every event that occurs in the device and those messages are called logs, the log contains all the information such as timestamp, host, type of application, messages etc., mining the required stuff from those logs would help different teams in analyzing and rectifying the problem so that quality of service can be improved, In this paper we propose log mining using the machine learning techniques to perform predictive analysis, correlation analysis, trend analysis on the data set so that a conclusion can be drawn out of the unstructured, voluminous data; But to understand the analytics needs some domain expertise, whereas the summary has to be reported to top level management hence we propose a visualization tool which gives a visual interface to report the summary.

I. INTRODUCTION

Data mining is a technique where a large chunk of data is mined to produce meaning full insights, for any organization data is an asset. A large chunk of meaningless data is of no use, with this data decisions cannot be taken hence, "data to decision" is a process in which data is finely pruned and mining rues are applied so that meaningful insights are produced. Log-mining is part of data mining here the handheld device logs are considered as data and the process of mining is applied here; for example a device puts up a log event, that has relevant and irrelevant information the intension is to get meaningful insights about how a test engineer or quality person can handle any problems related to device in future and how they can conclude it to top level management. To do this we follow few simple steps like data pruning, frequent pattern analysis, applying mining rules and building a knowledge base.

The steps are as follows:

1. **Data pruning**- this process takes multiple iterations, as to make the data more comfortable.
2. **Frequent pattern analysis**- this process is for finding any frequently occurred pattern inside the data.
3. **Applying mining rules**- the solution lies in this step, the mining rules are clustering, classification and association where-in we get to know how close issues are associated and how big the clusters are based on issues.
4. **Building knowledgebase**- the final step would be building 'KB. This KB contains meaningful, structured data so that one can bring out any conclusion out of it.

As said above the entire process takes four steps, the intension is to make decisions form data. How it is done is explained in the chapters below.

The data collected by a particular team is in unstructured format, the data is pruned using perl and python scripts. Then it is analyzed for strong correlation and association. It is similar to web mining or market basket analysis where the customer searching for some products and analyzed and products related to that are grouped and recommended to customer, for instance the customer buying bread and butter will be recommended to buy egg and butter.

The following are key contributions of this paper:

- This paper recommends advanced mining concepts wherein the teams can easily rectify the issue and find simple solutions to overcome the same.
- It also assures quality by minimizing the problems and improves user experience.

- The top level management can take effective decisions in minutes by looking at the demographic images, I,e “visualization”.

II. MOTIVATION

A. *Web Mining*

Now a day’s the people cannot leave without IoT. The E-commerce era has changed the lifestyle of human being such that everything is in one touch of a finger and the product is shipped to door steps. This evolution has left the data mining branch to further classify as “web mining”. The activities of web user is collected as data and are mined, for example the user activities are analyzed so that his age group, likings, the products which he searches for and other related information is collected and any related products are recommended. The recommendations are more over advertisements keeping the similar age group in mind. For instance anyone buying a mobile handset gets recommendations related to mobile accessories such as headset, case-cover, screen guard etc.

B. *Market Basket Analysis*

The MBA is a type of analysis used in data science arena the consumer data in a market place is analyzed such that the analysis gives the shopkeeper some insights of how he can improve sales of particular product or the entire business. As the name says the items purchased by any consumer is carefully watched and correlating item is recommended. For example any customer buying onions, butter, egg and bread is supposed to buy chees, pizza, burger buns etc. such recommendations are made to customer as well as shopkeeper So that customer gets product and owner gets business.
The association rule mining, clustering and correlation mining is used in MBA.

III. OVERVIEW OF THE PROPOSED SYSTEM

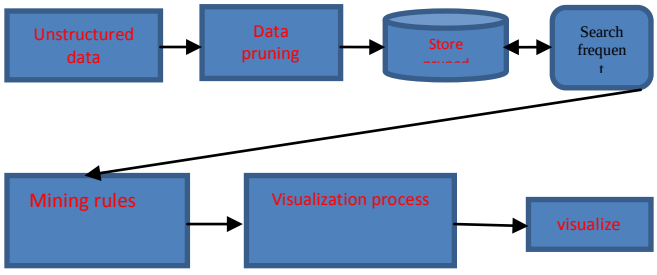

Figure: 3.1 overview of recommendation system.

As shown in figure 3.1 it consists of unstructured data store, data pruning frame work, pruned data store, and a frequent pattern database. The frequent pattern is then mined for meaningful insights and the mined data is then visualized using visualization process. A visualization engine is built upon java script library such that browser based tool handles the visualization process and plots demo graphs.

The process involves four to five steps as mentioned above the initial step is to collect data from source and then the unstructured data is pruned in many iterations the pruned data is then stored in data base such that it is accessed by an engine called frequent pattern engine and those frequent patterns are mined using mining rules like correlation, association, classification and clustering.

Correlation: Is a process wherein the frequent item are grouped and relation between them are studied and strong correlation are marked the whole idea behind this is to find cause and effect scenario.

Classification: Is a process where the data/ frequent sets are classified according to classification algorithms. These classifications are furthered studied for better insights.

Association: A process where the closely associated frequent items are grouped and analyzed the one to one association or similar concepts are studied, the apriori algorithm is used for association rule mining.

A. *VISUALIZATION*

There is a saying “A picture can convey thousand words”.
Yes, when it comes to data mining and visualization it’s true.
The data mined can be visualized so that anyone looking at the data for the first time or without domain knowledge can make out of the data and could be able to make decisions quickly.

The visualization engine is built on top of D3.js and other java script library so that it takes data and plots required demo graphs.

B. DATA PRUNING

Data pruning also called data massaging is a process where the data is pruned or cleaned and brought to fine structure, this process takes lot of iterations and we need to bring down to frequent pattern phase. The pruned data is then stored in a database from there the frequent item sets are queried.

IV. DATA MINING using MACHINE LEARNING ALGORITHMS

There are three algorithms used in entire mining process as mentioned below:

1. Classification
2. Clustering

3. Association mining and correlation

I. *Classification*: is a data mining (machine learning) technique used to predict group membership for data instances.
For example, we may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy.
Popular classification techniques include decision trees and neural networks (j48 decision tree and Bayesian networks).
Classification methods:
Learning from examples, concept learning
Step 1: Using a learning algorithm to extract rules from (create a model of) the training data.
Step 2: Evaluate the rules on test data. Usually split known data into training sample (2/3) and test sample (1/3).
Step 3: Apply the rules to (classify) new data.

Goals of classification:
Instance-based (lazy) learning: predict class label of new examples using training data directly.
Popular approaches are nearest neighbor, Bayesian classification and neural networks.

Decision tree algorithm:

ID3(D, Attributes, Target) //used in actual analysis.
t = createNode()
IF ∀hx, c(x)i ∈ D : c(x) = 1 THEN label(t) = '+', return(t) ENDIF
IF ∀hx, c(x)i ∈ D : c(x) = 0 THEN label(t) = '-', return(t) ENDIF
label(t) = mostCommonClass(D, Target)
IF Attributes = ∅ THEN return(t) ENDIF
A∗ = argmaxA∈Attributes(informationGain(D, A))
FOREACH a ∈ A∗ DO
Da = {(x, c(x)) ∈ D : x|A∗ = a}
IF Da = ∅ THEN
t0 = createNode()
label(t0) = mostCommonClass(D, Target)
createEdge(t, a, t0)
ELSE
createEdge(t, a, ID3(Da,Attributes \ {A∗}, Target))
ENDIF
ENDDO
return(t)

II. *Clustering:*

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.
A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".
A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Clustering algorithms and methods

The four most used clustering algorithms are K-means, Fuzzy C-means, Hierarchical clustering, Mixture of Gaussians.

k-means algorithm or The Lloyd's algorithm is as follows

step1: Initialize the center of the clusters
2. Attribute the closest cluster to each data point
3. Set the position of each cluster to the mean of all data points belonging to that cluster
step4. Repeat steps 2-3 until convergence.

III. *Association and correlation rule mining:*

A Frequent pattern is a pattern that occurs frequently in a data set.
Association rule mining uses apriori and filtered apriori in our analysis

Apriori:

1. Find all frequent itemsets:
Get frequent items:
Items whose occurrence in database is greater than or equal to the min.support threshold.
Get frequent itemsets:
Generate candidates from frequent items.
Prune the results to find the frequent itemsets.
2. Generate strong association rules from frequent itemsets
Rules which satisfy the min.support and min.confidence threshold.

Association and correlation are used to find a strong pattern and the relation between them in the given data set.
The rules obtained by association and correlation are used to make meaningful decision.

V. CONCLUSION AND FUTURE WORK

Any data is always an asset for the organization but data without any sense is of no use, hence making data more meaningful is better and useful to make decision. Data to decision is what we do using the above mentioned methods, and finally the data is plotted in the form of graphs, info-graphic models to make better sense and to make it more meaningful to person even without knowledge about the same. Our work can help in making right decision and improve quality the work mentioned above can be taken to next level

where more machine learning techniques can be used and the whole idea can be automated. The visualization techniques used here built on top java script library but further it can be built using JSP and the entire process can be automated to a one click solution.

REFERENCES

1. A Study on Market Basket Analysis Using a Data Mining Algorithm. ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 6, June 2013 Phani Prasad J1, Murlidher Mourya2 1,2Vardhaman Engineering College, Hyderabad, India.

2. Market Basket Analysis with Data Mining methods, Trnka, A. ; Dept. of Appl. Inf., Univ. of SS. Cyril & Methodius, Trnava, Slovakia Networking and Information Technology (ICNIT), 2010 International Conference on 11-12 June 2010.

3. P. Giudici, S. Figini, "Applied Data Mining for Business and Industry. Second Edition". John Wiley & Sons Ltd; 2009. ISBN 978-0-470-05886-2.

4. https://en.wikipedia.org/wiki/Clustering.

5. https://en.wikipedia.org/wiki/Association_rule_learning

6. https://en.wikipedia.org/wiki/Statistical_classification