Pattern classification using Web Mining

Krunal Jaha[1] Krishna Patel[2] Krishna Punwar[3] Mr. Romil Patel[4]
[1,2,3]Student [4]Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]Sigma Institute of Engineering, Vadodara, India

Abstract— It is the application of data mining techniques on the web data to solve the problems to extracting useful information. As the information in the internet increases, the search engines lack the efficiency of providing relevant and required information. This project proposes an approach for the web content mining using the algorithm. The aim of our project is the pattern classification of dataset and analysis the data on E-commerce business or organization websites.

Key words: Pattern Classification, Web Mining, C 5.0 Algorithm

I. Introduction

Web mining is type of the data mining techniques to discover some interesting patterns from the web data. Classify interesting pattern from using some classification algorithm and techniques. [5] In this paper we read some papers which are related to the pattern classification from the web data. There are three types of web mining. Web content mining, web structure mining, and web usage mining.

Classification algorithms are used for the complex and real web lo data.

In web usage mining, pattern discovery is difficult because there are little bit information in data like IP addresses, user id, buy id, pin code are available. [1] In web usage mining there are several types of it which are web server logs, application level logs and application level logs. From that data we discover the pattern which is useful for the e-commerce websites or e-commerce companies like Amazon, flipkart. The use of these types of web mining helps to gather important information from customer buying from the e-commerce site. [1]

It helps the e-commerce companies for productivity flow, e-business depends on the information and the data which we conclude to take right decision.

Our system will improve the business of any e-commerce websites or e-commerce companies.

II. Pattern Classification Using Web Mining

After identifying the dataset there are various fields like IP addresses, use session, pin code, product id, and buy id and payment method. There are various kinds of pattern techniques to display the result. When dataset is cleaned than after apply an algorithm which for classification and then after we found particular pattern which we needed from the data. We used classification analysis, data items are classified according to predefined categories. [1]

In our work there are web log data which we divided in particular session and divided by the zip code or pin code. We classify the IP addresses and divided by the payment method either COD (cash on Delivery) or payment by card (credit card, debit card).

III. Proposed System

In proposed methodology for classification of web data in order some predefine our criteria. In this mode we present the steps of the system.

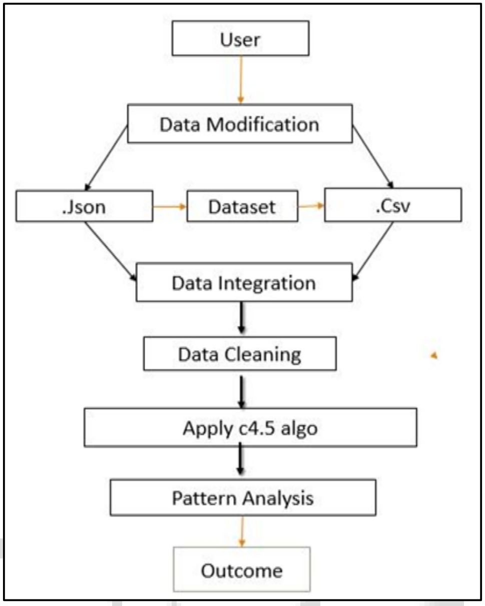

Fig. 1: Proposed System

IV. Data Cleaning

Data cleaning is the process to remove error data and inconsistent data. There are some missing values which can be removed from the data cleaning process and then after it reduces the size of the dataset. [1]

V. System Brief

In our system there is a step which is our proposed system. There is user which is work on the data. Data set are in many formats that first we have to modified the data which we have. Then converting the dataset. It is the major part of the system that dataset in any type of format like .json, .csv we have dataset in .json format first we have to convert data set .json to .csv.

After data conversion than data integration process made where dataset will be merged in one dataset.

Data cleaning is the major process of the system that in data set there are noisy data which is one type of error that in this process inconsistent data and noisy data will be removed from these process.

After data cleaning process there are data which we needed for the system than we apply the algorithm which is the main part of the system. And we produce the final result.

And result can be displayed in any type of format.

VI. Algorithm

Input: training dataset T; attributes S.
Output: decision tree

```
if T is NULL then
return failure
end if
if S is NULL then
return Tree as a single node S
end if
if |S| =1
return Tree as a single node S
end if
set Tree
for a € S do
set Info (a, T) = 0, and Split Info (a, T) =0
compute Entropy (a)
for v € values (a, T) do
set Ta, v as the subset of T with attribute a= v
Info a, T) + = - |Ta, v| / |Ta| Entropy (av)
Split Info (a, T) + = - |Ta, v| / |Ta| log |Ta, v| / |Ta|
end for
Gain (a, T) = Entropy (a) – Info (a, T)
Gain Ratio (a, T) = Gain (a, T) / Split Info (a, T)
end for
set abest = argmax {Gain Ratio (a, T)}
attach abest into Tree
for v € values (abest, T) do
call C4.5 (Ta, v)
end for
return Tree.
```

This algorithm used for the classification. It generates the tree at the end of it. [4]

Here Entropy calculation,

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where, pi – proportion of S, [2]

Information Gain Calculation,

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where, values (A) – set of all possible values of A
Sv – subset of s. [2]

VII. Future Work

In future there is the classification of the pattern from the above algorithm which includes the result of the dataset. In future there are large amount of data set and use of this algorithm reduces the error and give the more accurate output. This algorithm is faster than the other algorithm to improve the result and it saves the time.

References

[1] Er. Romil V Patel and Dheeraj Kumar Singh, “Pattern classification based on Web Usage Mining Using Neural Network Technique”, IJCA, vol.71-No.21, June 2013.

[2] Anurag Upadhyay, suneet shukla and sudsanshu kumar, “Empirical comparison by data mining classification algorithm (C 4.5 & C 5.0) for thyroid cancer data set”, International journal of computer science & communication network, vol.3(1), 64-68.

[3] Rutvija Pandya and Jayati Pandya, “C 5.0 algorithm to improved decision tree feature selection and reduced error pruning”, IJCA, vol. 117 – NO.16, May 2015.

[4] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia, “Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning”, IJCSIT, vol.3(2), 2012,3427-3431

[5] Monika Yadav and Mr. Pradeep Mittal, “Web Mining – An Introduction”, IJARCSSE, vol. 3, 3 March 2013.