CLASSIFICATION OF E-COMMERCE DATA USING DATA MINING

Aditi Todi[1], Anahita Agrawal[2], Ankit Taparia[3], Nikhlesh Lakhmani[4],Dr. Rajashree Shettar[5]

[1]*Student, Computer Science Engineering, R.V.College of Engineering,Bangalore,India,**aditi.todi89@gmail.com***
[2]*Student, Computer Science Engineering, R.V.College of Engineering,Bangalore,India,**anahita.17@gmail.com***
[3]*Student, Computer Science Engineering, R.V.College of Engineering,Bangalore,India,**ankit.taparia89@gmail.com***
[4]*Student, Computer Science Engineering, R.V.College of Engineering,Bangalore,India,**nikhillakhmani@gmail.com***
[5]*Professor, Computer Science Engineering,R.V.College of Engineering,Bangalore,India,**rajashreeshettar@rvce.edu.in***

Abstract

This paper discuss the development of an application for both consumers and companies, which would make use of information available on E-commerce websites to classify the data in accordance with the quality of the products available.Electronic commerce or E-commerce refers to the buying and selling of products over electronic systems such as the Internet. The amount of trade conducted electronically has grown extraordinarily with widespread Internet usage. In today's world the major players such as Amazon, Flipkart and eBay have a large database of products and a number of consumers that use these services. Since, there is such a large amount of data available with no one to collate and understand it,this paper attempts to classify the data to benefit both consumers and companies. The classification is done using two popular algorithms – Naïve Bayes and Decision Tree which are supervised learning methods. A comparison of the classification is done to understand which algorithm produces better results.The result which we have obtained using the mobile phone datasets shows that Decision Tree performs better than Naive Bayes.Once such an application is built competitors can understand how their competitors are priced and consumers understand what quality of products are available. This is mutually beneficial making the E-commerce world a flatter playing ground.

Index Terms: *XML, DOM Parser, ARFF, Naive Bayes, Decision Tree.*

-- *** --

1. INTRODUCTION

The World Wide Web has a vast amount of data present today. One of the most popular standards to express the data is XML. The World Wide Web has given rise to a flourishing E-commerce industry today. Electronic commerce, commonly known as e-commerce or e-comm, refers to the buying and selling of products or services over electronic systems such as the Internet and other computer networks.

Electronic commerce draws on such technologies as electronic funds transfer,supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web at least at one point in the transaction's life cycle, although it may encompass a wider range of technologies such as e-mail, mobile devices and telephones as well.

The amount of trade conducted electronically has grown extraordinarily with widespread Internet usage. In today's world the major players such as Amazon, Flipkart and eBay have a large database of products and a number of consumers that use these services. The product, pricing and policy differ from country to country. Mining the e-commerce data involves understanding the products available on different websites and the prices that the products are offered at. It is useful for consumers for whom choice is expanded and companies who offer the products at different pricing levels.The purpose of classification of e-commerce data is for consumers and companies to understand the different products available and the prices they are offered at. It gives an idea of which products are popular in the market and what consumers are willing to pay for certain products. Automation of the classification when it comes to product genre and pricing reduces cost and timing for both consumers and companies. Armed with this data, consumers can get the best deals for their products whereas companies can focus on offering better services and deals to increase their consumer base [1].

With the e-commerce industry growing rapidly in recent times, it is imperative to understand the consumer base that a company is appealing to. This not only helps the company that is trying to enter the market but also offers a consumer choice. In this work data is mined in such a way that it presents the company with coherent data to understand the consumer base and tailor made policies. It also provides the consumer an opportunity to view different products and their differential pricing to make a more informed choice.

The traditional way of doing market survey is a very time consuming and tedious process. It has now been simplified with the e-commerce industry coming into the competitive markets. However, consumers are unwilling to do surveys on the Internet and usually stick to sites they have liked at their first attempt. This causes a new company, which is entering the market, a great amount of grief and revenue loss. It becomes difficult for competitors to lure consumers away as data and information is not available on consumers' choices and preferences. Consumers stand to lose, as they are not in a position to make an informed decision as there is too much data available but none of it is in an understandable form. Our application helps to bridge the gap between the consumer and the company by giving the company information on consumers' choices and preferences. It benefits the consumers by giving them the price differentials present and the product genres available.

2. METHODOLOGY

The methodology adopted in this project is presented diagrammatically in Fig 1. The data sets are collected and processed after which it is converted into a suitable format. Post this step, classification is done and results are displayed in an understandable format.[2]

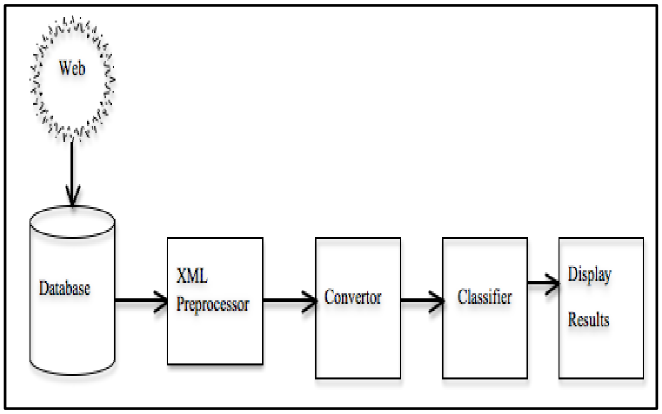

Fig-1: Methodology for Implementation of Application

From Fig 1, the web is the starting point of this application as represented . The web contains a large amount of data. However, this large amount of data is unprocessed and unusable for end users without being processed beforehand. The standard that the Web is moving towards is XML.[5]

The second component in Fig 1 stores the E-commerce data that is to be classified. The E-commerce data has been collected from popular E-commerce sites.Popular datasets are used for the classification of devices such as mobiles, pen drives and laptops.

The third component present in Fig 1 deals with the representation of the XML document. Preprocessing is done using the Document Object Model (DOM) Parser. The DOM parser is preferred over the Simple API for XML (SAX) parser as changes are not required in the XML data. Also, referencing back to the data is not required making DOM a viable option.[8]

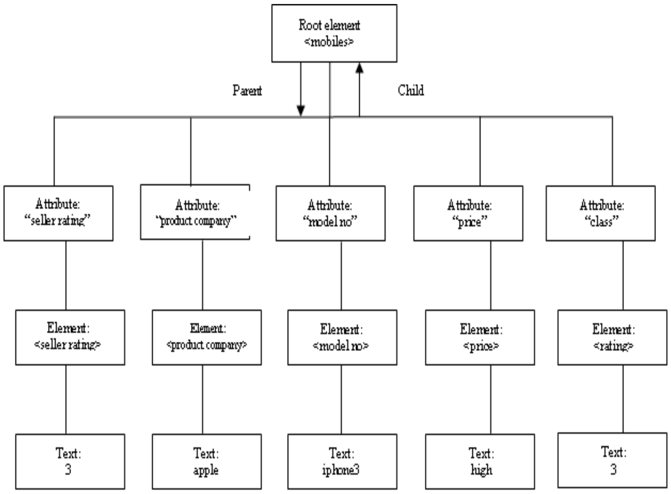

Fig-2: XML DOM Tree Example

Fig 2 represents the output obtained after applying the DOM parser to an XML document. The DOM parser gives the output as an XML DOM tree.

The attributes present in the the dataset are seller rating, product company, model no and price. The DOM parser is constructed manually to parse these attributes.

The next component in Fig 1 deals with converting the XML data to an Attribute-Relation File Format (ARFF). ARFF files have two distinct sections. The first section is the Header information, which is followed the Data information.The ARFF is then fed to the Waikato Environment for Knowledge Analysis (WEKA) tool for classification. The number of data points provided for training are 70 whereas the test set has 30 data points.

3. CLASSIFIER

The classification algorithm that is used is Naïve Bayes and Decision Tree. The methodology of classification for each algorithm is different and hence, gives us different results.

(a) Naïve Bayes

Naïve Bayes classifier is the simplest instance of a probabilistic classifier. The output *Pr* (*C*|*d*) of a probabilistic classifier is the probability that a document *d* belongs to a class *C*.[6]

Let Y be a random variable where

$$Y_i = \begin{cases} 1 \text{ if a event occurred,} \\ 0 \text{ otherwise} \end{cases} \quad \textbf{-Eq (1)}$$

and X_1 , X_2 , …. X_p be a set of predictor variables. If the predictors are conditionally independent given Y, the conditional joint probabilities can be written as

$$P(X_1 , X_2 , \ldots. X_p|Y) = \prod_{j=1}^{p} P(Xj \mid Y) \textbf{ --Eq (2)}$$

Combining this with Bayes Theorem leads to the Naive Bayes Classifier [7]

$$\log \frac{P(Y = 1 \mid X1,\ldots,Xp)}{P(Y = 0 \mid X1,\ldots,Xp)} = \log \frac{P(Y = 1)}{P(Y = 0)} + \sum_{j=1}^{p} \log \frac{f(Xj \mid Y) = 1)}{f(Xj \mid Y) = 0)}$$

Wheref (x_j|Y) is the conditional density of X_j.**-Eq (3)**

For Example: Fictional data set that describes the different rating for buying some specified mobile phones

Seller rating	Product company	Model no	Price	Rating	Seller rating	Product company	Model no	Price	Rating
3	nokia	n8	high	2	1	samsung	wave	low	1
2	nokia	n8	low	1	1	apple	iphone3	low	1
3	nokia	n8	mid	2	1	nokia	n8	mid	1
2	apple	iphone3	low	2	2	samsung	wave	mid	2
3	samsung	wave	high	3	3	apple	iphone3	mid	3
3	samsung	wave	mid	3	1	apple	iphone3	low	1
3	apple	iphone3	high	3	2	samsung	wave	low	2

Table-1: Mobile Phones Dataset

	Seller rating				Product company				Modelno				Price			Rating		
	3	2	1		3	2	1		3	2	1		3	2	1	3	2	1
3	4	2	0	nokia	0	2	2	n8	0	2	2	high	2	1	0	4	5	5
2	0	3	1	apple	2	1	2	iphone3	2	1	2	mid	2	2	1			
1	0	0	4	samsung	2	2	1	wave	2	2	1	low	0	2	4			
	3	2	1		3	2	1		3	2	1		3	2	1	3	2	1
3	4/4	2/5	0/5	nokia	0/4	2/5	2/5	n8	0/4	2/5	2/5	high	2/4	1/5	0/5	4/14	5/14	5/14
2	0/4	3/5	1/5	apple	2/4	1/5	2/5	iphone3	2/4	1/5	2/5	mid	2/4	2/5	1/5			
1	0/4	0/5	4/5	samsung	2/4	2/5	1/5	wave	2/4	2/5	1/5	low	0/4	2/5	4/5			

Table-2: Frequencies and probabilities for the mobile phones data

(b) Decision tree

Decision trees are the most commonly used because of its ease of implementation and ease in understanding as compared to other classification algorithms. Decision Tree classification algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm. [3]

Decision tree algorithm is a data mining induction techniques that recursively partitions a data set of records using depth-first greedy approach or breadth-first approach until all the data items belong to a particular class. [4] A decision tree structure is made of root, internal and leaf nodes. The tree structure is used in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measures. The tree leaves is made up of the class labels in which the data items have been grouped.

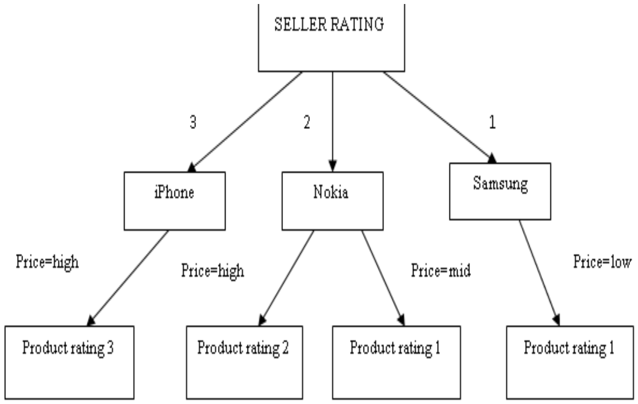

Fig-3: Decision Tree

Figure 3 shows us the decision tree which is made using the table 1 of mobile phones dataset.

There are numerous decision tree algorithms available today. The one, which we will be using, is the C4.5 Algorithm. In C4.5, Pruning takes place by replacing the internal node with a leaf node thereby reducing the error rate. It has an enhanced method of tree pruning that reduces misclassification errors due to noise or too-much detail in the training data set.

4. RESULTS

This component represented in Fig 1 deals with displaying the results in a human readable format. The rating of the products is done in accordance with the rules. The rating of the product has 3 classes viz. 3,2, 1. Class 3 indicates the best class of products while Class 1 indicates the lowest class of products.

The evaluation of the classifier is done on the following parameters.

$$\text{Recall} = \frac{\text{number of documents retrieved that are relevant}}{\text{total number of documents that are relevant}} \textbf{-Eq(4)}$$

$$\text{Precision} = \frac{\text{number of documents retrieved that are relevant}}{\text{total number of documents that are retrieved}} \textbf{-Eq(5)}$$

$$\text{F} - \text{Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \textbf{-Eq(6)}$$

Types of Classifier	**Analysis**									
	Accuracy (%)	**Precision**			**Recall**			**F-Measure**		
		3	**2**	**1**	**3**	**2**	**1**	**3**	**2**	**1**
Naïve Bayes	67	1	.68	.51	1	.60	.59	1	.63	.55
Decision Tree	99	1	.98	1	1	1	.97	1	.99	.98

Table-3: Analysis of Mobile Phones Dataset

In Table 3, it shows that Decision Tree has a better accuracy as compared to Naive Bayes for the mobiles phones dataset.

5. CONCLUSION

This papershows how we can overcome the traditional way of survey for customer feedback and use analysis to aggregate and summarize the feedback from the customer which is available online as product reviews. By using the analysis system, the cost and time needed to analyze how a product is doing in the market can be reduce drastically. This provides the opportunity to react faster to the customer complaints and needs. Using analysis one can easily monitor their product online in real time by collecting data from Amazon, E-bay and Flipkart.

The experimental result of analysis using training set shows that, the Decision classifier has performed better than Naïve Bayes classifier with better precision, recall and F-Measure. For mobile phones dataset, the classifier has achieved an accuracy level of 99% by Decision Tree classifier and 67% by Naïve Bayes classifier.

ACKNOWLEDGEMENT

We would like to thank our **HOD**, **Dr. N.K.Srinath** for his constant support extended towards us during the course of the paper. His help and advice has beenentirely responsible for instilling the drive in us to complete the paper. We would also like to thank our principal **Dr. B**. **S.Satyanarayan** who has always been a great source of inspiration.

REFERENCES

[1] Zhang Na, Zhang Dongzhan, Yu Ye and DuanJiangjiao "An improved method for classifying XML documents based on structure and content" in Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCSCT '10) Jiaozuo, P. R. China, 14-15, August 2010, pp. 426-430

[2] Wu Haitao, Tang Zhenmin "Automatic Classification Method for XML Documents" in International Journal of Digital Content Technology and its Applications (JDCTA) Volume5, Number12, December 2011

[3] Matthew N. Anyanwu&Sajjan G. Shiva "Comparative Analysis of Serial Decision Tree Classification" in Algorithms International Journal of Computer Science and Security, (IJCSS) Volume (3): Issue (3) 230

[4] Diana Gorea, Sabin CorneliuBuraga "Towards Integrating Decision Tree with XML Technologies" in 8th International Conference on Development and Application Systems, Suceava, Romania, May 25 -27, 2006

[5] Zailani Abdullah, Muhammad SuzuriHitam "Features

Extraction Algorithm from SGML for Classification" in Journal of Theoretical and Applied Information Technology, 2007

[6] S.L. Ting, W.H. Ip, Albert H.C. Tsang "Is Naïve Bayes a Good Classifier for Document Classification?" in International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011

[7] Kim Larsen "Generalized Naive Bayes Classifiers", in SIGKDD Explorations. Volume 7, Issue 1 in page 76-81

[8] Hosam F.EL-Sofany, Samir A. El-Seoud, Fayed F.M. Ghaleb, Jihad M. Al Ja'am, Sameh S. Daoud, and Ahmad M. Hasnah "A DOM-Based Approach of Stroage and Retrieval of XML Documents Using Relational Databases" in International Journal of Computing & Information Sciences, Vol.5, No. 2, August 2007

BIOGRAPHIES

Aditi Todi. She is presently pursuing Computer Science Engineering at RV College of Engineering. Her area of interest is Database Management Systems.

Anahita Agrawal She is presently pursuing Computer Science Engineering at RV College of Engineering. Her area of interest is Networking.

Nikhlesh Lakhmani He is currently pursuing Computer Science Engineering at RV College of Engineering. His area of interest is Operating Systems.

Ankit Taparia He is currently pursuing Computer Science Engineering at RV College of Engineering. His area of interest is Operation Research.

Rajashree Shettar She is a professor at Computer Science Engineering at RV College of Engineering (BITS, PILANI),PhD. She has 14 years of teaching experience and has a number of international journal and international conference publications to her name.