Web Personalization of Indian e-Commerce Websites using Classification Methodologies

Agarwal Devendera[1], Tripathi S.P[2] and Singh J.B.[3]

**[1] School of Computer Science & Information Technology, Shobhit University, Research Scholar
Meerut, U.P. 250110, India**

**[2] Department of Computer Science & Engineering, Gautam Buddh Technical University, IET
Lucknow, U.P. 226021, India**

**[3] School of Computer Science & Information Technology, Shobhit University
Meerut, U.P. 250110, India**

Abstract

The paper highlights the classification methodologies using Bayesian Rule for Indian e-commerce websites. It deals with generating cluster of users having fraudulent intentions. Secondly, it also focuses on Bayesian Ontology Requirement for efficient Possibilistic Outcomes.

Keywords: *E-commerce, Possibilistic Outcome, Bayesian Rule, Ontology.*

1. Introduction

Electronic Commerce is fast emerging as most popular method of purchasing, let it be a small pen drive or bulky LED TV. Recent survey [3] has estimated that around 3-5% of Indians have transacted or are well versed with working of online shopping websites. The strategy which is being followed until now related to the various policy initiatives like:

- ***Consumer Proportion:*** This model is being propagated by the government based on certain guidelines for the protection of consumers.
- ***Legality:*** It deals with formal recognition of electronic signatures; In India digital signatures are necessary for e-Tendering.
- ***Security***: Central Government has issued its policy relating to cryptography techniques to ensure secure electronic commerce in third party transfer.

In order to deal with security and web personalization [2] issues we develop two basic classification methods: Naïve Bayes and K-nearest neighbor.

2. Our Model

In order to make our model more illustrative we are taking example of **"Predicting Fraudulent Transaction".**

An Indian e-commerce company has a very large customer base; each customer has to submit his personal information before making a transaction. In this way each company is acting as a record and the response of internet is given as

$$Z = \{Fraudulent, Trustworthy\} \qquad (1)$$

these are the classification in which we can categorize a customer. By analyzing from a sample e-commerce site we are able to find out that in case of ***Fraudulent*** the customer-id should be reposted to the e-fraud cell. Two set of data are taken to check the consistency of data.

Table 1: Report of Customer on e-commerce site.

	Reporting to e-fraud cell	***No Reporting Required***	***Total***
fraudulent	20	80	100
trustworthy	100	300	400
Total	**120**	**380**	**500**

2.1 Naïve Bayes

In order to classify record into **'m'** classes by ignoring all predictor information $\boldsymbol{X_1}$, $\boldsymbol{X_2}$,….., $\boldsymbol{X_p}$ is to classify the record as a member of majority class. For example in our case naïve rule would classify all the customers to be

"**Trustworthy**", because 90% of the companies were found to be ***Truthful***.

Naïve Bayes classifier [1] is an advanced version of Naïve rule. The logic to introduce Bayes is to integrate the information given in the set of predictors into the naïve rule to obtain more accurate classifications. The methodology suggests in finding out the probability of record belonging to a certain class is evaluated on the prevalence of that class along with additional information that is being given on that record in terms of ***X*** information.

Since our dataset is very large we prefer Naïve Bayes method. In a classification task our goal is to estimate the probability of membership to each class given a certain set of predictor variables. This type of probability is called a conditional probability. In our example we are interested in ***P (Fraudulent | Reporting to e-fraud cell)***. In general, for a response of '**m**' classes $C_1, C_2, \ldots, C_m$ and the predictors $X_1, X_2, \ldots, X_p$ we compute as:

$$P(C_i \mid X_1, \ldots, X_p) \text{ where } i = 1, 2, \ldots, m. \qquad \textbf{(2)}$$

When the predictors are all categorical we can use a pivot to estimate the confidential probabilities of class membership. Consider its application in our example we compute the probabilities divided into two classes as: For ***P (Fraudulent | Reporting to e-fraudulent cell) = 20/120*** and

P (Trustworthy | Reporting to e-fraudulent charges) = 100/120.

The above statement indicates that although the firm is still more likely to be ***Trustworthy*** than ***Not Trustworthy***, the probability of its being ***Truthful*** is much lower than the naïve rule.

However, the method usually gives good result partly because what is important is not the exact probability estimate but the ranking for that case in comparison to others.

In order to convert the desired probabilities into class probability we use Bayes Theorem. The Bayes Theorem gives us the following formula to compute the probability that the record belongs to class C_i:

$$P(C_i \mid X_1, \ldots, X_P) = \frac{P(X_1, \ldots, X_P \mid C_1)P(C_1)}{P(X_1, \ldots, X_P \mid C_1)P(C_1) + \ldots + P(X_1, \ldots, X_P \mid C_m)P(C_m)} \qquad \textbf{(3)}$$

C_i: To compute the numerator we filter two pieces of information

i) The proportion of each class in the population $[P(C_1) \ldots\ldots P(C_m)]$

ii) The probability of occurrence of the predictor vales $X_1, X_2, \ldots, X_p$ within each class from the training set.

We develop another table of the User which is categorized as "**Frequent Buyers**" and "**Occasional Buyers**", for each of these two categories of Buyers we have information on whether or not reporting has been done, and whether it turned out to be ***Fraudulent*** or ***Trustworthy***.

Table 2: Sample of 10 users.

Reporting to e-fraud cell	***User-Type***	***Status***
Yes	Occasional Buyer	Fraudulent
No	Occasional Buyer	Trustworthy
No	Frequent Buyer	Fraudulent
No	Frequent Buyer	Trustworthy
No	Occasional Buyer	Trustworthy
No	Occasional Buyer	Trustworthy
No	Frequent Buyer	Trustworthy
Yes	Occasional Buyer	Fraudulent
Yes	Frequent Buyer	Fraudulent
No	Frequent Buyer	Fraudulent

The probability of fraud can be defined by four possible states **{Yes, Occasional Buyer}**, **{Yes, Frequent Buyer}**, **{No, Occasional Buyer}**, **{No, Frequent Buyer}.**

i) P(Fraudulent | Reporting = Yes, Customer Type = Occasional Buyer) = 1/2 = 0.5

ii) P(Fraudulent | Reporting = Yes, Customer Type = Frequent Buyer) = 2/2 = 1

iii) P(Fraudulent | Reporting = No, Customer Type = Occasional Buyer) = 0/3 = 0

iv) P(Fraudulent | Reporting = No, Customer Type = Frequent Buyer) = 1/3 = 0.33

We can extend this for Naïve Bayes probabilities, for analyzing the conditional probabilities of fraudulent behavior **"Reporting to e-fraud cell" = Yes**, and "**User Type" = Occasional Buyer**, the numerator is a proportion of "**Reporting to e-fraud cell**". Instances amongst the type of Buyers, times the proportion of Fraudulent Customers

$$= (3/4)\ (1/4)\ (4/10) = 0.075$$

To get the actual probability we calculate the numerator for the conditional probability of truth given

Reporting to e-Fraudulent Cell = Yes;
Type of Customer = Occasional Buyer;

The denominator is then the sum of two conditional probabilities

$$= (0.075 + 0.067) = 0.14$$

Therefore the conditional probability of fraudulent behaviors is given by

$$P_{NB}(\text{Fraudulent} \mid \text{Reporting to e-Fraudulent cell} = \text{Yes; Buyer Type} = \text{Occasional})$$

$$= \frac{(3/4)(1/4)(4/10)}{(3/4)(1/4)(4/10)+(1/6)(4/6)(6/10)}$$

= 0.075/0.14 = 0.53

P_{NB}***(Fraudulent | Reporting to e-Fraudulent cell = Yes; Buyer Type = Frequent) = 0.087***

P_{NB}***(Fraudulent | Reporting to e-Fraudulent cell = Yes; Buyer Type = Occasional) = 0.031***

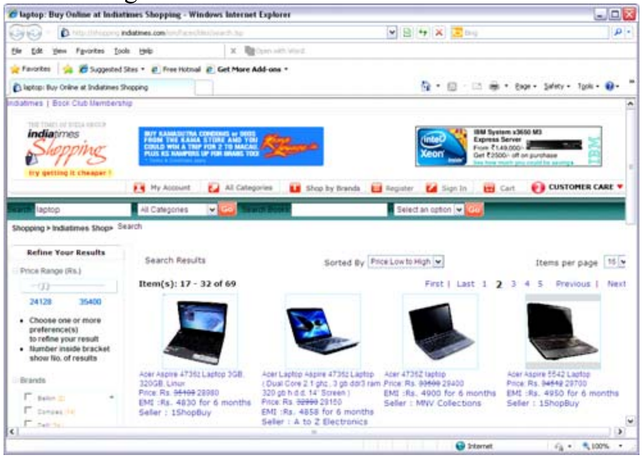
Rank Ordering of probabilities are even closer to exact Bayes method than are the probabilities themselves, to further analyze we can use classification matrix.

2.1 Advantages & Disadvantages of Naïve Bayes Classifier

The logic of using Naïve Bayes Classification Technique [7] is to attain computational efficiency and good performance.

2.2 Fuzzy Information Classification and Retrieval Model

The above section deals with a classification technique [6] by which we can categorize the customer visiting our site based on their transaction history. In this section we have highlighted the problem which our customer face while selecting the best possible combinations of product, the problem is because of the uncertainty in Semantic Web Taxonomies [8]. Consider Indiatimes shopping portal shown in fig. 1.


laptop: Buy Online at Indiatimes Shopping - Windows Internet Explorer
indiatimes.com
File Edit View Favorites Tools Help
Favorites
Suggested Sites
Free Hotmail
Get More Add-ons
laptop: Buy Online at Indiatimes Shopping
Page
Safety
Tools
Indiatimes | Book Club Membership
indiatimes Shopping
try getting it cheaper !
IBM System x3650 M3 Express Server
From ₹1,49,000/-
Get ₹2500/- off on purchase
My Account
All Categories
Shop by Brands
Register
Sign In
Cart
CUSTOMER CARE
laptop
All Categories
Go
Select an option
Go
Shopping > Indiatimes Shop> Search
Internet
100%

Refine Your Results

Price Range (Rs.)

24128 35400

- Choose one or more preference(s) to refine your result
- Number inside bracket show No. of results

Brands

Search Results

Sorted By Price Low to High

Items per page 16

Item(s): 17 - 32 of 69

First | Last 1 2 3 4 5 Previous | Next

Acer Aspire 4736z Laptop 3GB, 320GB, Linux
Price: Rs. ~~35499~~ 28980
EMI :Rs. 4830 for 6 months
Seller : 1ShopBuy

Acer Laptop Aspire 4736z Laptop (Dual Core 2.1 ghz , 3 gb ddr3 ram 320 gb h.d.d. 14' Screen)
Price: Rs. ~~32999~~ 29150
EMI :Rs. 4858 for 6 months
Seller : A to Z Electronics

Acer 4736Z laptop
Price: Rs. ~~33699~~ 29400
EMI :Rs. 4900 for 6 months
Seller : MNV Collections

Acer Aspire 5542 Laptop
Price: Rs. ~~34642~~ 29700
EMI :Rs. 4950 for 6 months
Seller : 1ShopBuy

Fig. 1: Indiatimes Shopping Portal.

If a buyer wants a laptop in the range of Rs.25000 < x < Rs.35000, and with features F = {f1, f2, f3} in brands B = {b1, b2}, then he must be shown the best possibilistic outcome of the above query.

The above problem looks very simple but it is not so, there exists an uncertainty in the query, what if, if there is no laptop with all the features of ‘F’ present in Brand ‘B’. Here comes a probabilistic method to overcome such situation.

In our method, degrees of subsumption will be covered by Bayesian Network based Ontology’s [4]. The Venn diagram shown in figure 2

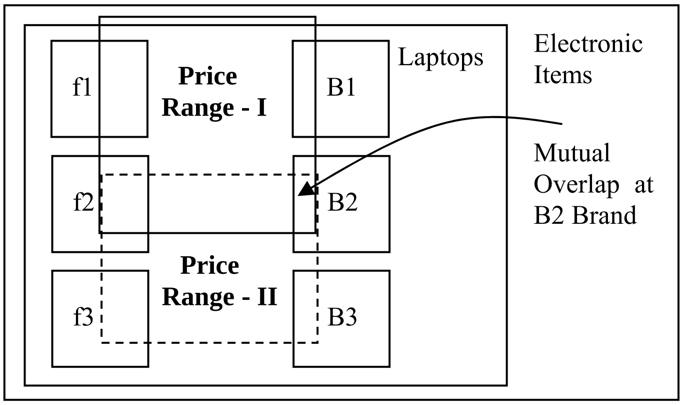


Fig. 2: Venn Diagram Illustrating Electronic Items with Laptops as one of their Categories & their Overlap.

Our method enables the representation of overlap between a selected concept and every other is referred taxonomy. The **Price Range-I** represent the prices at the start of the price band while **Price Range-II** represent the higher side of the price band.

The overlap is logic term expressed as

$$\text{Overlap} = \frac{|\,\text{Selected} \bigcap \text{Referred}\,|}{|\,\text{Referred}\,|} \in [0,1] \qquad \textbf{(4)}$$

The overlap region represents the value ***0*** for disjoint concepts and ***1***, if the referred concept is subsumed by the selected one. This overlap value can be used in information retrieval tasks. The match with the query is generalized by the probabilistic sense and the hit list can be sorted into the order of relevance accordingly.

If ‘***F***’ and ‘***B***’ are sets; then ‘***F***’ must be in one of the following relationships to ‘***B***’.

i) ‘***F***’ is a subset of ‘***B***’ ie $\mathrm{F} \subseteq \mathrm{B}$.

ii) ‘***F***’ partially overlaps ‘***B***’ ie $\exists x, y : (x \in F \wedge x \in B) \wedge (y \in F \wedge y \in B)$

iii) ‘***F***’ is disjoint from ‘***B***’ ie $F \cap B = \varphi$

Based on these relations we develop a simple transformation algorithm. The algorithm processes the overlap graph ***G*** in a **Breadth First** manner starting from root concept defined as ‘***CON***’. Each processed concept ‘***CON***’ is written as the part of **Solid Path Structure (SPS)**.

The overlap values ‘***O***’ for a elected concept ‘***F***’ and a referred concept ‘***B***’

```
if      F subsumes B then
|       O := 1
else
|       $C = F_S \cap B_S$
|       if $C = \varphi$ then
|       |       O : = 0
|       else
|       |       $O := \frac{\sum_{c \in C} m(C)}{m(B)}$
|       end
end
```

Fig. 3: Computing the Overlap.

If ***F*** is the selected concept and ***B*** is referred one, then the overlap value ***0*** can be interpreted as the conditional probability

$$P(B' = \text{true} \mid F' = \text{true}) = \frac{S(F) \cap s(B)}{s \mid (B) \mid} = 0$$

where ***S(F)*** and ***S(B)*** are taken is and interpreted as a probability space, and the elements of the sets are not interpreted as elementary outcomes of some random phenomenon.

The implementation stages of the probabilistic search starts with the Input of Ontology Rule which are refined in "**Refinement Stage**". It is than passed to the "**Quantifier**" which develops a set of Association Rules. It is then fed to the further preprocessing by the "**Naïve Bayesian Transformation**" module which finally generates the best possible overlapping outcome as shown in figure 4.

3. Conclusions & Future Scope

The model in both the cases uses interactive query refinement mechanism to help to find the most appropriate query terms. The Ontology is organized according to narrower term relations. We have developed an algorithm in which taxonomies can be constructed without virtually any knowledge of Probability and Bayesian network.

The future extension could be to expand it using Fuzzy Regression [7] with Bayesian Network.

Acknowledgments

Agarwal, Devendera thanks Darbari, Manuj without his essential guidance this research paper would not have been possible and also to management of Goel Group of Institutions for their financial support in his maiden endeavor.

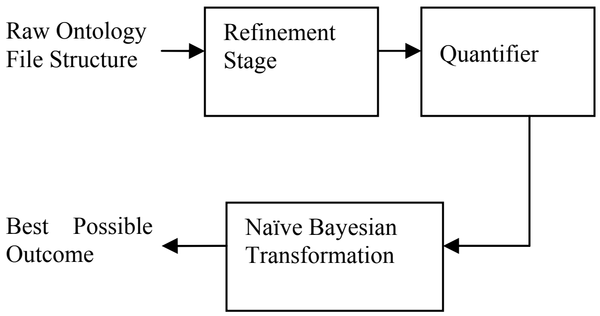


Fig. 4: Implementation Framework.

References

[1] Bergenti, F., "Improving UML Designs using Automatic Design Pattern Detection", in 12th International Conference on Software Engineering and Knowledge Engineering (SEKE), 2000.

[2] Chen, Ming-Chung., "Mining User Progressive User Behavior for E-Commerce using Virtual Reality Technique", M.S. Thesis, Faculty of Graduate School, University of Missouri-Columbia, 2007.

[3] IAMAI, "I-CUBE 2009-10", Report by IMRB International, India, 2010.

[4] Ding, Z., "A Probabilistic Extension to Ontology Language OWL", in 12th Hawaii International Conference on Systems Science, 2004.

[5] GuT, "A Bayesian Approach for Dealing with Uncertain Concepts", in Conference Advances in Pervasive Computing, Austria, 2004.

[6] Schafer, J.Ben., et. al., "E-Commerce Recommendation Applications", as Grouplens Research Project, University of Minnesota, 2008.

[7] Wang, Lipo., "Fuzzy Systems & Knowledge Discovery", Springer, 2006.

[8] Zongmin, Ma., "Soft Computing in Ontologies and and Semantic Web", Springer, 2006.

Agarwal, Devendera is currently working as Prof. & Director (Academics) at Goel Institute of Technology & Management, Lucknow. He has over 12 years of teaching & 5 years of industrial experience. Having done his B.Tech in Computer Science from Mangalore University in 1993, M.Tech from U.P.Technical University, Lucknow in 2006, he is pursuing his Ph.D. from Shobhit University, Meerut.

Tripathi, S.P. (Dr.) is currently working as Assistant Professor in Department of Computer Science & Engineering at I.E.T. Lucknow. He has over 28 years of experience. He has published numbers of papers in referred National Journals.

Singh, J.B. (Dr.) is currently working as Dean Students Welfare at Shobhit University, Meerut. He has 38 years of teaching experience and has published number of papers in referred National Journals.