

PREDICTION OF USERS' PERFORMANCE ON A MARKETING WEBSITE

Ketan Badhe¹, Sanyukta Garje², Rutuja Jagdale³, Sameer Herkal⁴, Vaishali Malpe⁵

^{1,2,3,4} Student, BE-Computer Engineering, Dept. of Computer Engineering, Terna Engineering college, Maharashtra, India.

⁵ Professor, Dept. of Computer Engineering, Terna Engineering college, Maharashtra, India.

Abstract - Online shopping is becoming more and more common in our daily lives. Understanding users' interests and behavior is essential to adapt e-commerce websites to customers' requirements. The information about users' behavior is stored in the web logs. The analysis of such information has focused on applying data mining techniques, where a rather static characterization is used to model users' behavior, and the sequence of the actions performed by them is not usually considered. Therefore, incorporating a view of the process followed by users during a session can be of great interest to identify more complex behavioral patterns. To address this issue, we are implementing data mining algorithms and perform analysis of structured e-commerce web logs. Then identify different behavioral patterns that consider the different actions performed by a user will help to improve the website as whole.

Key Words: Data mining, online shopping, sequence pattern, behavioural pattern.

1. INTRODUCTION

The increase in popularity of the Internet and the rapid development of E-commerce, Internet-based businesses' websites are facing increasing competition. E-commerce sites generate large amounts of data daily, and these data include potential consumer-related information that is valuable for market analysis and prediction. E-commerce business analysts require to know and understand consumers' behaviour when those navigate through the website, as well as trying to identify the reasons that motivated them to purchase, or not, a product[1]. Therefore, the most important challenge of E-commerce is to elucidate customers' wants, love, and value orientation as much as possible to ensure competitiveness in the E-commerce era.

2. LITERATURE SURVEY

Data mining (DM) is used to attain knowledge from available information in order to help companies make weighted decisions. Data mining is field that focuses on access of information useful for high level decisions and to help online shopping stores to identify online customer behaviour to recommend for him the appropriate products he/she is interesting to them [3].

Various algorithms are used to find the pattern or frequent sequence in the data. Thus, the most efficient algorithms are extracted and used for our system. Moreover, an organization needs to invest only on the group of products which are

frequently purchased by its customers as well as price them appropriately in order to attain maximum customer satisfaction [8].

Hidden relationships in sales data can be discovered from the application of data mining techniques [1]. Thus, techniques for identifying the areas of improvement or the area which is liked or disliked are needed by data analyst or owner of ecommerce website.

2. PROPOSED SYSTEM

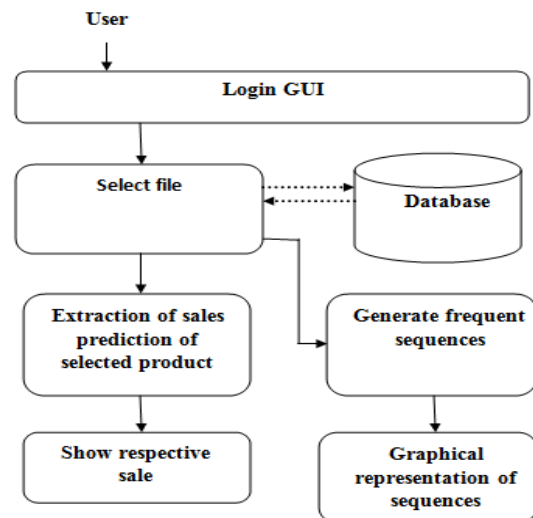


Fig . Architecture of proposed system

The user will have to login to the system. The selection of database file is done by the user.

On the selected dataset the SPADE and ID3 algorithms are processed applied.

The SPADE algorithm will do the generation of frequent sequences.

After that its minimum support will be calculated. Graphical view shows that which item or combinations of items were selling frequently. Result gives proper combinations to the user.

The ID3 algorithm will do the extraction of sales prediction of selected product.

The result shows the predicted sell of an item according to the views, ratings given by the user.

4. RELATED WORK

4.1 ID3 (ITERATIVE DICHOTOMISER 3)

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node[1]. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric---information gain.

Algorithm;

ID3 (Learning Sets S, Attributes Sets A, Attributes values V)
Return Decision Tree.

Begin Load learning sets first, create decision tree root node 'rootNode', add learning set S into root node as its subset.

For rootNode, we compute Entropy(rootNode.subset) first

If Entropy(rootNode.subset)==0, then rootNode.subset consists of records all with the same value for the categorical attribute, return a leaf node with decision attribute:attribute value;

If Entropy(rootNode.subset)!=0, then compute information gain for each attribute left(have not been used in splitting), find attribute A with Maximum(Gain(S,A)). Create child nodes of this rootNode and add to rootNode in the decision tree.

For each child of the rootNode, apply ID3(S, A, V) recursively until reach node that has entropy=0 or reach leaf node.

End ID3.

Attribute Selection

How does ID3 decide which attribute is the best? A statistical property, called information gain, is used. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. In order to define gain, we first borrow an idea from information theory called entropy. Entropy measures the amount of information in an attribute.

Given a collection S of c outcomes

$$\text{Entropy}(S) = S - \sum p(I) \log_2 p(I)$$

where p(I) is the proportion of S belonging to class I. S is over c. Log2 is log base 2.

Note that S is not an attribute but the entire sample set.

4.2 SPADE (Sequential Pattern Discovery using Equivalence Class)

SPADE algorithm was introduced in 2001 by M.J.Zaki. It makes use of Apriori vertical formatting approach [4]. The original sequence database is transformed into vertical id-list data format, in which each id-list associates with the corresponding items (SID) and time stamp (TID). The aim of this algorithm is to find frequent sequences using efficient lattice search techniques and simple join operations. It requires only three database scans to discover all the sequences.

SPADE algorithm working is as follows:

Step 1: Scan the database once and discover frequent sequences of length one by using Apriori property.

Step 2: Generation of candidate sequences set of length two by joining all pairs of frequent items.

a) Check if the two items have the same SID and is in sequential order of events.

b) A list of frequent sequences of length two is discovered and finalized.

Step 3: Traversing the lattice for support count and enumeration of frequent sequences.

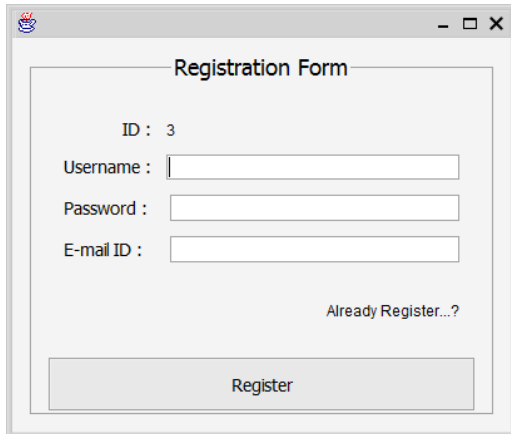
a) Lattice is traversed in either breadth first search or depth first search. It is quite large to be filled in main memory. So decomposition of lattice into equivalence classes by the prefix of sequence. Sequences that are in same class have a common prefix.

5. COMPARISON OF METHODS

Table -1: Comparison

PARAMETERS	ID3	SPADE
Algorithm	Classification algorithm	Sequential pattern mining
Method	Decision Tree	Time series analysis
Advantages	Easy to understand and implement; Classification results with strong interpretability; Does not need complex data pre-processing	It can find the potential connection in the sequence with certain order.
Disadvantages	It can be difficult to control the size of the tree; In some complex cases, splitting data into classes might not be helpful;	Often there are ways around getting a model that is time-series based where the predictions are almost as good and is faster to implement.

6. RESULTS



Registration Form

ID : 3

Username :

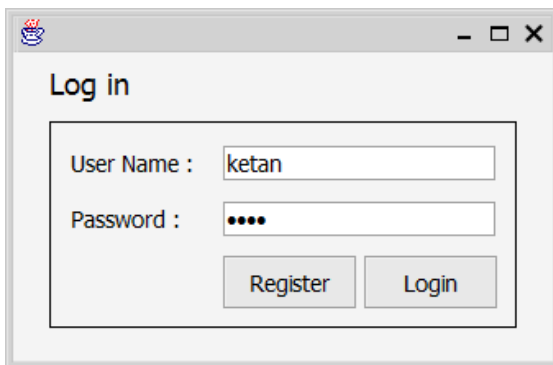
Password :

E-mail ID :

Already Register...?

Register

Fig - 1: Registration form



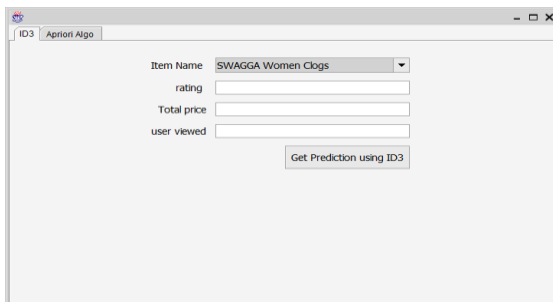
Log in

User Name :

Password :

Register Login

Fig - 2: Login form



Item Name: SWAGGA Women Clogs

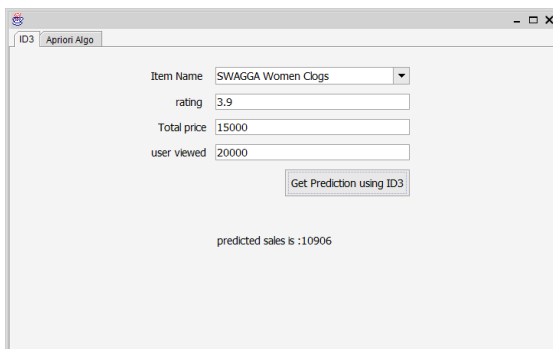
rating:

Total price:

user viewed:

Get Prediction using ID3

Fig - 3: ID3 page(1)



Item Name: SWAGGA Women Clogs

rating: 3.9

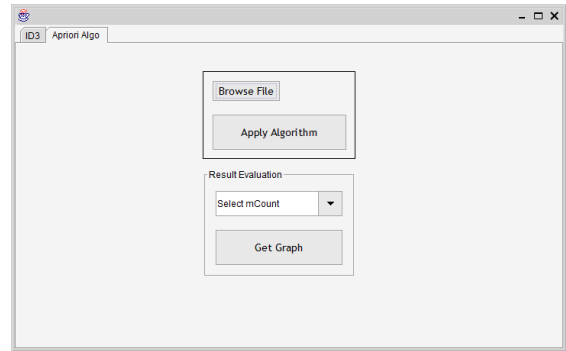
Total price: 15000

user viewed: 20000

Get Prediction using ID3

predicted sales is :10906

Fig - 4: ID3 page(2)



Browse File

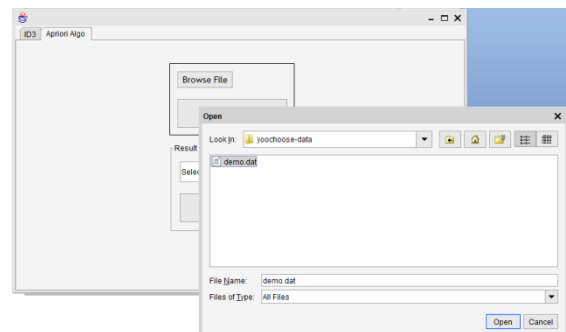
Apply Algorithm

Result Evaluation

Selected mCount

Get Graph

Fig - 5: Spade(1)



Open

Look In: yoochoose-data

demo.dat

File Name: demo.dat

Files of Type: All Files

Open Cancel

Fig - 6: Spade(2)

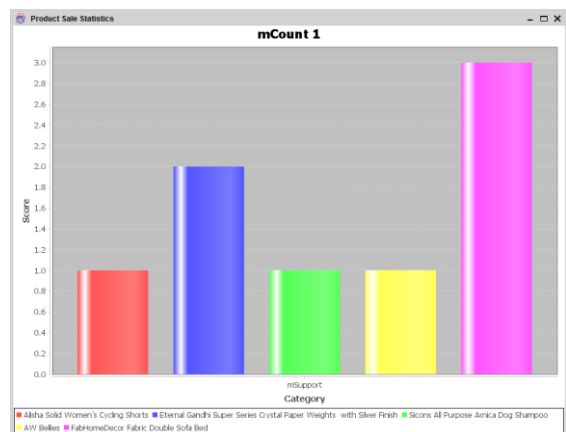


Fig - 7: minimum support is 1

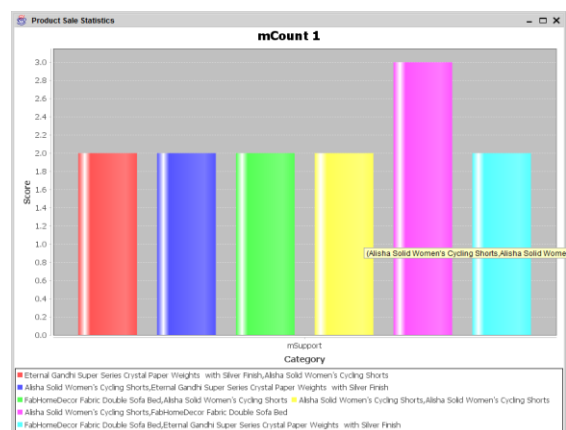


Fig - 8: minimum support is 2

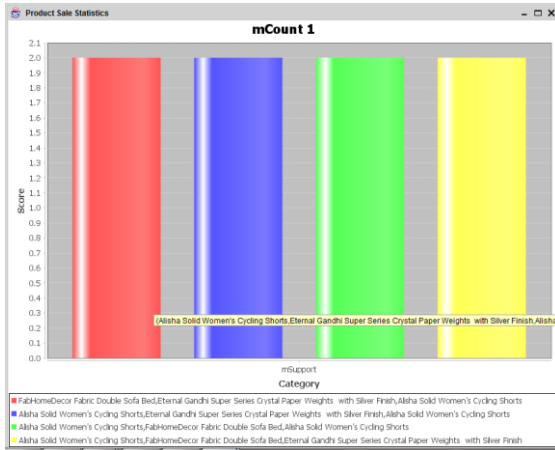


Fig – 9: minimum support is 3

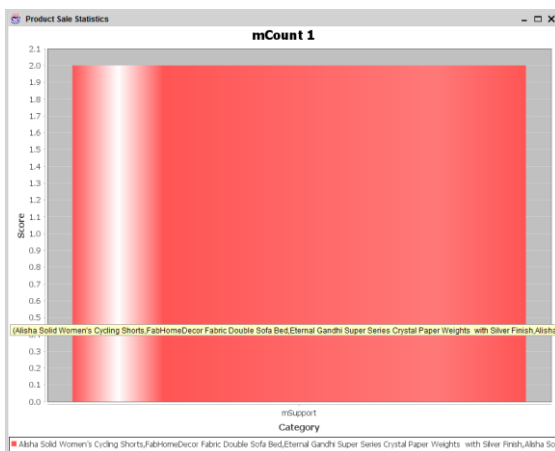


Fig – 10: minimum support is 4

[4] Zaki, M.J.: SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning* 42(1), 31–60 (2001).

[5] Ayres, J., Flannick, J., Gehrke, J., Yiu, T.: Sequential pattern mining using a bitmap representation. In: *Proc. 8th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, pp. 429–435. ACM (2002).

[6] Yang, Zhenglu, Yitong Wang, and Masaru Kitsuregawa. "LAPIN: effective sequential pattern mining algorithms by last position induction for dense databases." *Advances in Databases: Concepts, Systems and Applications*. Springer Berlin Heidelberg, 2007. 1020- 1023.

[7] Fournier-Viger, Philippe, et al. "Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information." *Advances in Knowledge Discovery and Data Mining*. Springer International Publishing, 2014. 40-52.

[8] Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining: concepts and techniques" Morgan kaufmann, 2006.

7. CONCLUSIONS

In this paper, a detailed study based on data mining techniques was conducted in order to extract knowledge in a data set with information about user’s history associated to an e-commerce website[8].

The set of descriptive data mining techniques are applied that allows data analyst working at ecommerce companies make strategic decisions to boost their sales as well as provide effective customer service.

REFERENCES

[1] Anurag Bejju ‘Sales Analysis of E-Commerce Websites using Data Mining Techniques’, IEEE 2017

[2] Chetna Kaushal, Harpreet Singh ‘Comparative Study of Recent Sequential Pattern Mining Algorithms on Web Clickstream Data’, IEEE 2015.

[3] Rana Alaa Eleen Ahmeda , M.Elemam.Shehaba , Shereen Morsya , Nermeen Mekawiea , ‘Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining’, IEEE 2015.