USERS' BEHAVIOR PREDICTION ON E-COMMERCE WEBSITE

Ketan Badhe[1], Sanyukta Garje[2], Rutuja Jagdale[3], Sameer Herkal [4], Vaishali Malpe[5]

[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 indentify 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.

3. METHODS

3.1 K-MEANS CLUSTERING ALGORITHM

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups)[3]. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

Algorithm:

Initialize the centre of the clusters
μi= some value ,i=1,...,k
2. Attribute the closest cluster to each data point
$ci=\{j:d(xj,\mu i)\leq d(xj,\mu l),l\neq i,j=1,...,n\}$
3. Set the position of each cluster to the mean of all data points belonging to that cluster $\mu i=1|ci|\sum j\in cixj,\forall i$
4. Repeat steps 2-3 until convergence

3.2 Apriori Algorithm

The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules[7]. Key Concepts:

- Frequent Item sets: The sets of item which has minimum support (denoted by Li for ith-Itemset).

- Apriori Property: Any subset of frequent item set must be frequent.

- Join Operation: To find L k, a set of candidate k-item sets is generated by joining Lk-1 with itself.

Apriori Algorithm Pseudo code:

```
Procedure Apriori (T, minSupport)
{ L1= {frequent items};
for (k= 2; Lk-1 !=∅; k++) {
Ck= candidates generated from Lk-1
for each transaction t in database do {
#increment the count of all candidates in Ck that are contained in t
Lk = candidates in Ck with minSupport
}}
UkLk ;   }
```

3.3 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.

3.4 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.

3.5 SPAM (Sequential Pattern Mining)

SPAM algorithm that uses a depth first search strategy to traverse the lexicographic sequence tree [4].In the lexicographic sequence tree, each sequence is considered either sequence–extended sequence or item set extended sequence. The generation of sequences by traversing the tree can either generate sequence-extended children (called the S-step or S-extension) or item set extended children (called the I-step or I-extension). The sequence tree is traversed in depth first manner and the support of each extension is checked against the minimum support. At each S-step and I-step pruning based on apriori is applied to reduce the number of children nodes and ensure that all the nodes of frequent sequences are traversed. While traversing the database, a vertical bitmap is created for each item. Each bitmap has a bit corresponding to each item set in the database. The bit corresponding to the item set of the bitmap for the item is set to zero if an item does not appear in the item set otherwise it is set to one.

3.6 LAPIN (Last Position Induction)

LAPIN sequential pattern mining algorithm based upon the thought that the last position of an item is the key used to find if a frequent sequence pattern of length (k) can be extended to be a frequent pattern of length (k+1) by appending it with last position item of pattern of length(k)[6]. By searching the last position of every item in each sequence there is a need to only search a small portion of the database. If the last position of item is smaller or equal than the position of the last item of pattern of length (k) then that cannot be appended to it. According to the last position list element we can construct the S-step item-last-position list and I-step last-position list in ascending order. The lists of items can be merged to construct a frequent sequence pattern of greater length.

4. 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.

[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.