INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2089 - 2094
RECEIVED : 19-03-2017. PUBLISHED: 20-04-2017.

April 20, 2017

SHOPPING PACKAGE SOFTWARE USING ENHANCED TOP K-ITEM SET ALGORITHM THROUGH DATA MINING

[1]T.Indhu, [2]Dr.D.Gayathridevi,

[1]Research Scholar, Department of Computer Science, Sri Ramakrishna college of Arts and Science for women, Coimbatore, India,
[2]Assistant Professor, Department of Computer Science, Sri Ramakrishna college of Arts and Science for women, Coimbatore, India.

ABSTRACT

The frequent item set and maximum threshold signature of the shopping package software item set. The frequent item set deals with the whole database of the shopping application. It contains various transactions like sales data, purchase data, customer data, item data and etc. Here enhanced TOP K – Item set has been implemented (TKI) which gives more accuracy and performance than TKU (mining Top-K Utility item sets) and TKO (mining Top-K utility item sets in One phase), which are implemented in the exiting methods for mining such item sets without the consideration of entire database. This may cause inaccurate result and improper output. These methods may use of assumption purpose only. TOP K-Item set Algorithm is used to analyze the items which are sold frequently that are reported by the client. The decisions can be made on the result of analysis, so that the item can be identified. Normally an input given by the client to the sale the item in project is taken as it is and service is provided without analyzing the input. This leads to wastage of time in decision making and also the delay in finding the frequently sold items. If the frequently sold items in project are analyzed, then it is easy to find out the relationship or association among the items. So, that the reason for sold item and how frequently sold item on each other can be found in a project. A new concept called data engineering is used in the system to find associations or relationships among the frequently sold items. So that both data mining and networking concepts are implemented. Data mining refers to detecting of patterns and hidden information from database. Several data mining techniques are available to mine the data and the results or the new or hidden information. The system provides the information about the associations among the frequently sold item in an item level. The system has used two data mining techniques namely association rules and its algorithms to finding the frequently sold items. Some of the results are displayed in a graphical manner also. The results of these techniques would be helpful in decision making, so that the client needs can be satisfied in a faster way.

Keywords : TKU (mining Top-K Utility item sets), TKO (mining Top-K utility item sets in One phase), Decision Making, Frequent Item mining.

1. INTRODUCTION

Generally, any software development organization receives a report sold the items in projects from its user and provides services to its user based on their input without analyzing them. This leads to

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2089 - 2094
RECEIVED : 19-03-2017. PUBLISHED: 20-04-2017.
April 20, 2017

wastage of time in decision making and also delay in finding the frequently sold items. It is because they do not know the how much items sold. The results of the above queries are unknown and they can be answered only by using data mining techniques. Hence, a system is needed with data mining techniques to analyze the sold items in package in order to find the relationship among the frequently sold items. The dependencies among modules and programs can also be found using data mining techniques. The frequently sold items can also be done quickly by using information generated by data mining techniques. The hidden or unknown information is useful in decision making and the decisions quickly.

2. RELATED WORKS

The subgroup discovery algorithm CN2-SD, based on a separate and conquer strategy, has to face the scaling problem which appears in the evaluation of large size data sets. To avoid this problem, in this paper we propose the use of instance selection algorithms for scaling down the data sets before the subgroup discovery task. The results show that CN2-SD can be executed on large data set sizes pre-processed, maintaining and improving the quality of the subgroups discovered [1]. We present an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. We also present results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm [2]. If the cost parameters are not known at training time, Receiver Operating Characteristic (ROC) analysis can be applied (Provost & Fawcett 1997; Swets, Dawes & Monahan 2000). ROC analysis provides tools to distinguish classifiers that are optimal under some class and cost distributions from classifiers that are always sub-optimal, and to select the optimal classifier once the cost parameters are known[3]. ROC analysis for two classes is based on plotting the true-positive rate (TPR) on the y-axis and the false-positive rate (FPR) on the x-axis. This gives a point for each classifier[4].

3. IMPLEMENTATION

K - Mine Algorithm Pseudo code

```
procedure kmine (T, min Support)
{
//T is the database and min Support is the minimum support
 L1= {frequent items};
 for (k= 2; Lk-1 !=Ø; k++)
{
 Ck= candidates generated from Lk-1
//that is cartesian product Lk-1 x Lk-1 and
 Eliminating any k-1 size itemset that is not
```

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2089 - 2094
RECEIVED : 19-03-2017. PUBLISHED: 20-04-2017.
April 20, 2017

```
//frequent

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 min Support

}

//end for each

}

//end for return ∪;

}
```

Implementation Process

As is common in association rule mining, given a set of itemsets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets. K - Mine uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Sample usage of K - Mine algorithm A large supermarket tracks sales data by Stock-keeping unit (SKU) for each item, and thus is able to know what items are typically purchased together. K - Mine is a moderately efficient way to build a list of frequent purchased item pairs from this data. Let the database of transactions consist of the sets {1,2,3,4}, {1,2,3,4,5}, {2,3,4}, {2,3,5}, {1,2,4}, {1,3,4}, {2,3,4,5}, {1,3,4,5}, {3,4,5}, {1,2,3,5}.

Each number corresponds to a product such as "butter" or "water". The first step of K - Mine is to count up the frequencies, called the supports, of each member item separately:

Item Support 1 6 2 7 3 9 4 8 5 6 We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let min support = 4. Therefore, all are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, K - Mine prunes the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent . Item Support {1,2} 4 {1,3} 5 {1,4} 5 {1,5} 3 {2,3} 6 {2,4} 5 {2,5} 4 {3,4} 7 {3,5} 6 {4,5} 4 We generate the list of all 3-triples of the frequent items (by connecting frequent pair with frequent single item). Item Support {1,3,4} 4 {2,3,4} 4 {2,3,5} 4 {3,4,5} 4 The algorithm will end here because the pair {2,3,4,5} generated at the next step does not have the desired support. We will now apply the same algorithm on the same set of data considering that the min support is 5. We get the following results: Step 1: Item Support 1 6 2 7 3 9 4 8 5 6 Step 2: Item

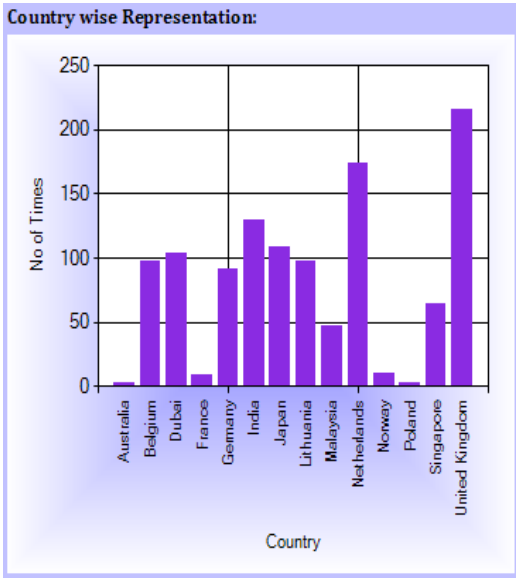
INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2089 - 2094
RECEIVED : 19-03-2017. PUBLISHED: 20-04-2017.
April 20, 2017

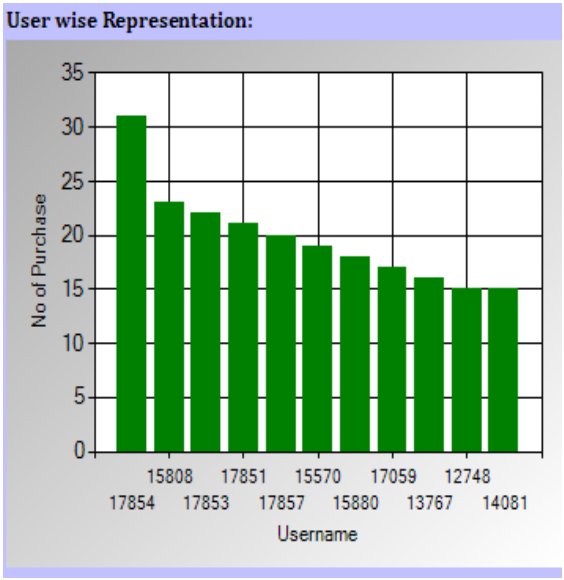
Support {1,2} 4 {1,3} 5 {1,4} 5 {1,5} 3 {2,3} 6 {2,4} 5 {2,5} 4 {3,4} 7 {3,5} 6 {4,5} 4 The algorithm ends here because none of the 3-triples generated at Step 3 have de desired support.

4. RESULT AND ANALYSIS

Interpreting and Comparing Results

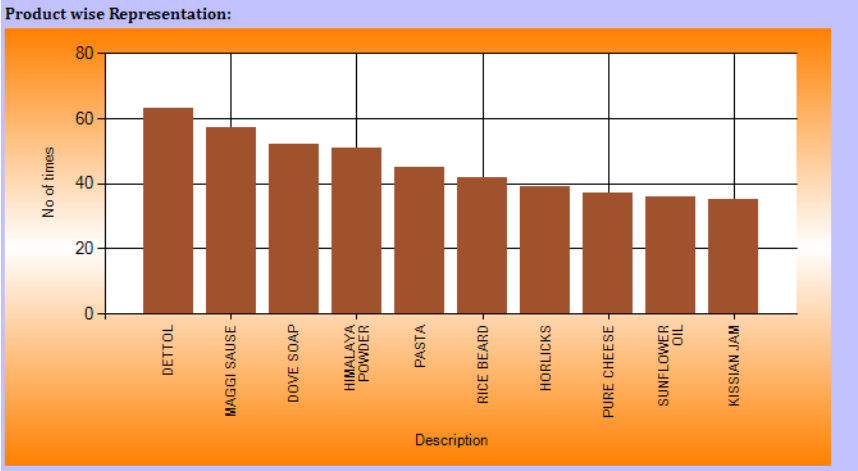
When comparing the results of applying association rules to those from simple frequency or cross-tabulation tables, you may notice that in some cases very high-frequency codes or text values (items) are not part of any association rule. This can sometimes be perplexing.





To illustrate how this pattern of findings can occur, consider this example: Suppose you analyzed data from a survey of insurance rates for different makes of automobiles in America. Simple tabulation would very likely show that many people drive automobiles manufactured by Ford, GM, and Chrysler; however, none of these makes may be associated with particular patterns in insurance rates, i.e., none of these brands may be involved in high-confidence, high-correlation association rules linking them to particular categories of insurance rates. However, when applying association rules methods, automobile makes which occur in the sample with relatively low frequency (e.g., Porsche) may be found to be associated with high insurance rates (allowing you to infer, for example, a rule that if Car=Porsche then Insurance=High). If you only reviewed a simple cross-tabulation table (make of car by insurance rate) this high-confidence association rule may well have gone unnoticed.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2089 - 2094
RECEIVED : 19-03-2017. PUBLISHED: 20-04-2017.
April 20, 2017



Product Wise Representation

CONCLUSION

The output has been verified as per the committed abstract. K mine has been implemented successfully as per (1) Constructing the Decision-Tree, (2) Populating potential k – mine for high utility item sets (PKHUIs) (3) Identifying top-k and UT s from the set of PKHUIs. A huge data set of 5,40,000 of data has been executed successfully. The output has been verified with three different data types in various conditions. Output has been verified, according to the given input. The obtained result is prompt according to the given commitments. The K-Mine can make a major impact in the data mining concept. Also it may usefully for sales business like retail, whole sale and online business. Clustering of the data has been implemented successfully in the pre processing stage itself. More over many categorization processes has been done for generating various outputs from single dataset. For graphical representation, various charts has been developed. User can clear the database for fresh use of these methods. So that this project has been implemented successfully and result has been verified.

REFERENCE

1. The CN2 induction algorithm", Machine Learning.Clark, P. and Niblett, T., "The CN2 induction algorithm", Machine Learning, Vol. 3(4), pp. 261–283, 1999.
2. Mining Association Rules between Sets of Items in Large Databases. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large database.In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'93), pp:207-216, Washington, DC, May 1993.
3. Learning Decision Trees Using the Area under the ROC Curve Cesar Ferri-Ram´ ırez, Peter A. Flach, and Jose Hernandez-Orallo. Learning decision trees using the area under the roc curve. In Proceedings of the Nineteenth International Conference on Machine Learning, pages 139–146, Morgan Kaufmann, 2002.

4. Discovering business intelligence from online product reviews: A rule-induction framework. W. Chung and H. Chen. Web-Based Business Intelligence Systems: A Review and Case Studies. In G. Adomavicius and A. Gupta, editors, Business Computing, volume 3, chapter 14, pages 373–396. Emerald Group Publishing, 2009.
5. Logical Design of Data Warehouses from XML. M. Banek, Z. Skocir, and B. Vrdoljak.Logical Design of Data Warehouses from XML . In ConTEL '05: Proceedings of the 8th international conference on Telecommunications, volume 1, pages 289–295, 2005.
6. A multisession-based multidimensional model. M. Body, M. Miquel, Y. B´edard, and A. Tchounikine. A multidimensional and multi version structure for OLAP applications. In DOLAP '02: Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, pages 1–6, New York, NY, USA,2002. ACM.
7. Transaction Management for a Main-Memory Database. P. Burte, B. Aleman-meza, D. B. Weatherly, R. Wu, S. Professor, and J. A. Miller. Transaction Management for a Main-Memory Database. The 38th Annual South eastern ACM Conference, Athens, Georgia, pages 263–268, January 2001.
8. Discovering business intelligence from online product reviews: A rule-induction framework. W. Chung and H. Chen. Web-Based Business Intelligence Systems: A Review and Case Studies. In G. Adomavicius and A. Gupta, editors, Business Computing, volume 3, chapter 14, pages 373–396. Emerald Group Publishing, 2009.
9. Crowd sourcing Predictors of Behavioural Outcomes. Josh C. Bongard, Member, IEEE, Paul D. H. Hines, Member, IEEE, Dylan Conger, Peter Hurd, and Zhenyu Lu. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING YEAR 2013.
10. Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data. Kashif Javed, Haroon A. Babri, and Mehreen Saeed. Ieee transactions on knowledge and data engineering, vol. 24, no. 3, march 2012.
11. Techniques, Process, and Enterprise Solutions of Business Intelligence. Li Zeng, Lida Xu, Zhongzhi Shi, Maoguang Wang, and Wenjuan Wu. 2006 IEEE Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.
12. Support vector machine with adaptive parameters in financial time series forecasting. Cao, L.J. ; Dept. of Mech. Eng., Nat. Univ. of Singapore, Singapore ; Tay, F.E.H. EEE Transactions on, On page(s): 1167 - 1178 Volume: 19, Issue: 7, July 2008