Development of Top-K Association Rules for Mining E-Commerce Datasets and Their Applications in Healthcare Consumer Behavior Analytics

Prof. Giulia Conti[1]
[1] University of Toronto, Department of Data Science and Biomedical Informatics, Toronto, Canada

ABSTRACT

Data mining an interdisciplinary research area spanning several disciplines such as machine learning, database system, expert system, intelligent information systems and statistic. Data mining has gradually became an active and important area of research because of previously unknown and interesting knowledge from very large real-world database. Many ideas of data mining have been investigated in several related fields. A unique but important aspect or idea of the problem lies in the significance of needs to extend their studies to include the nature of the contents of the real world database. In this paper we developed the data mining system using Top K Rules for the association rule mining on e-commerce dataset.

Keywords: *Artificial Neural Network, Data Mining, CHARM algorithm, K rule mining, CM SPAM Algorithm*

I. INTRODUCTION

In present days human beings are used in the different technologies to adequate in there society. Every day the human beings are using the vast data and these data are in the different fields .It may be in the form of documents, may be graphical formats, may be the video, may be records (varying array). As the data are available in the different formats, it is needed to take proper action for better utilization of the available data. Whenever the customer will require the data should be retrieved from the database and make the better decision.

This technique is actually we called as a data mining or Knowledge Hub or simply KDD (Knowledge Discovery Process).The important reason that attracted a great deal of attention in the information technology is the discovery of useful information from large collections of data industry towards field of “Data mining” is due to the perception of “we are data rich but information poor”. A very huge amount of data is available today, but we hardly able to turn them in to useful information and knowledge for managerial decision making in different fields. To produce information it requires very huge database. It may be available in different formats; as like audio/video, numbers, text, figures, and Hypertext formats. To take complete advantage of data; the data retrieval alone is simply not enough, it requires a tool for extraction of the essence of information stored, automatic summarization of data and the discovery of patterns in raw data.

With the extremely large amount of data stored in databases, files, and other repositories, it is very important to develop powerful software or tool for analysis and interpretation of such type of data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all of this is ‘Data Mining’. Data mining is the method of extracting related or hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses [1][2][3][4]. Data mining tools predict behaviors and future trends and help organizations and firms to make proactive knowledge-driven decisions [2]. The automated and possible analysis which is offered by data mining tools, now moves beyond the analysis of past events which is use typically by decision support systems. Data mining tools can now answer the questions that traditionally were too much time consuming to resolve. These tools created databases for finding predictive information, finding hidden patterns that experts may miss because it lies outside their expectations.

One of the data mining problems is classification. Various different classification algorithms have been designed by various researchers to tackle the problem in different fields such as mathematical programming, machine learning, and statistics. Recently, there is a gush of data mining research in the database community. The classification problem is re-examined by researchers in the context of large databases. Database researchers are giving more attention to the issues related to the volume of data. They are also very much concerned with the effective use of the available database techniques, such as efficient data retrieval mechanisms. With all these concerns, most algorithms proposed are basically based on decision trees.

II. RELATED WORK

Author Liu Jian [5] presents a new algorithm of web mining, which is aimed at enriching the interpretation of the draft results of association rule mining. The proposed algorithm in this paper was based on Apriori algorithm which was proposed by Agrawal and Srikant, and contains the first support, the second support and Rel-confidence, and has been extensively used recently for the web mining. The algorithm is developed to excavate user preferred navigation patterns. Finally, they give the example to elaborate the new methodology. The research presented in this paper makes a contribution to web mining. But there is still space for improvement.

A procedure for mining association rules from a database for on-line recommendation is developed by Changchien and Lu [6]. The implemented system depends on three different technologies: a start schema database, a SOM, and RST (Rough Set Theory). In particular, SOM is just used to cluster the transformed and normalized transaction records. Since the authors determined that SOM cannot explain the resulting clusters itself, they proposed association rules to explain their meaning. Hence, RST is employed to derive rules that explain the characteristics of each cluster and the attribute relationships among the different clusters. This work does neither generate frequent itemsets nor calculate itemset support; instead, it uses a confidence metric based on RST to form rules. It does not discuss the accuracy of the final rules at all, but based on the explanation of the work, it can be classified as a soft-mining solution. Moreover, a strong dependency between the SOM and the database is presented and exploited for the performance of the entire system.

Gupta et al. [7] undertook the problem of pruning or grouping final association rules for inspection and analysis. Even if the work focused on proposing a new distance metric to group association rules, a SOM took part in the proposed grouping methodology. The idea developed was initially to calculate the distance values among the rules through their metric. Then, MDS (Multi-Dimensional Scaling) was employed to form a vectorial representation of the distances which served as inputs to the SOM which clustered such vector space in order to visualize the rules. Due to the orientation of the work, it can be categorized as part of the techniques developed for the management and visualization of rules rather than their discovery, in which a neural network was used to cluster the rule space.

Author Chai Wenguang [8] in his paper put in an intelligent intrusion detection system based on web data mining, which is, compared to other traditional intrusion detection systems, safer and more efficient.

III. MINING TOP K ASSOCIATION RULE

Association rule mining [9] consists of discovering associations between items in transportation. It is the most important data mining tasks and has been integrated in many commercial data mining software and has wide applications in several domains.

The idea of mining top-k association rules presented in this paper is analogous to the idea of mining top-k itemsets [10] and top-k sequential patterns [11] [12] [13] in the field of frequent pattern mining in database. Note that even though many authors have previously used the term “top-k association rules”, they did not use the actual standard definition of an association rule. KORD [14] [15] only finds rules with a single item in the consequent, whereas the algorithm of You et al. [16] consists of mining association rules from a stream instead of a transaction database.

To achieve this objective, a question is how to combine the concept of top-k pattern mining with association rules? Two thresholds are used for association rule mining. But, it is observed that, in practice minsup is much more difficult to set than minconf because minsup depends on database characteristics that are unknown to most users. Whereas minconf represents the minimal confidence that is needed by the user in rules and is generally easy to determine. Because of this reason, we define “top-k” on the support rather than the confidence.
The algorithm’s main idea is the following. Top K Rules first sets an internal variable that is minsup to 0. Then, the algorithm starts searching for rules. As and when a rule is found, it is added to a list of rules L ordered by the support. The list is used to keep track of the top-k rules found until now. Once k valid rules are found, the internal minsup variable is raised to the support of the rule with the lowest support in L. Raising the minsup value is used to reduce the search space when searching for more rules. Thereafter, every time a valid rule is found, the rule is inserted in L. The rules in L not respecting minsup anymore are removed from L, and minsup is raised to the value

of the least interesting rule in L. The algorithm continues for searching more number of rules until no rule is found, which means that it has found the top-k rules.

To search for rules, Top K Rules does not rely on the classical two steps approach to generate rules because it would not be efficient as a top-k algorithm (as explained in the introduction). The strategy that is used by Top K Rules alternatively consists of generating rules containing a single item in the antecedent and a single item in the consequent. Then, each rule is recursively grow by adding items to the antecedent or consequent. To select the items that are added to a rule to grow it, it scans the transactions containing the rule to find single items that could expand its right or left part. The name of the two processes for expanding rules in Top K Rules is right expansion and left expansion. These processes can be applied recursively to explore the search space of association rules.

Another idea incorporated in Top K Rules is to try for generating the most promising rules first. This is because if rules with high support are found earlier, Top K Rules can raise its internal minsup variable faster to reduce the search space. To perform this, Top K Rules uses an internal variable R which stores all the rules that can be expanded to have a chance of finding more valid rules. Top K Rules uses this set of rules R to determine the rules that are the most likely to produce valid rules with a high support to raise minsup more quickly and reduce a larger part of the search space.
The Top K Rule algorithm is as follows:

TOPKRULES(T, k, minconf) R := Ø. L := Ø. minsup := 0.
1. Scan the database T once to record the tidset of each item.
2. FOR each pairs of items i, j in the dataset, such that |tids(i)| ×|T| ≥ minsup and |tids(j)| ×|T| ≥ minsup:
3. Set sup({i}→{j}) := |tids(i) ∩ tids(j)| / |T |.
4. Set sup({j}→{i}) := |tids(i) ∩ tids(j)| / |T|.
5. Set conf({i}→{j}) := |tids(i) ∩ tids(j)| / |tids(i)|.
6. Set conf({j}→{i}) := |tids(i) ∩ tids(j)| / |tids(j)|.
7. IF sup({i}→{j}) ≥ minsup THEN
8. IF conf({i} → {j}) ≥ minconf THEN
SAVE({i} →{j}, L, k, minsup).
9. IF conf({j}→{i}) ≥ minconf THEN
SAVE({j}→{i}, L, k, minsup).
10. Set flag expandLR of {i}→{j} to true.
11. Set flag expandLR of {j}→{i} to true.
12. R := R∪{{i}→{j}, {j}→{i}}.
13. END IF
14. END FOR
15. WHILE ∃r ∈ R AND sup(r) ≥ minsup DO
16. Select the rule rule having the highest support in R
17. IF rule.expandLR = true THEN
18. EXPAND-L(rule, L, R, k, minsup, minconf).
19. EXPAND-R(rule, L, R, k, minsup, minconf).
20. ELSE EXPAND-R(rule, L, R, k, minsup, minconf).
21. REMOVE rule from R.
22. REMOVE from R all rules r ∈ R | sup(r) < minsup.
23. END WHILE

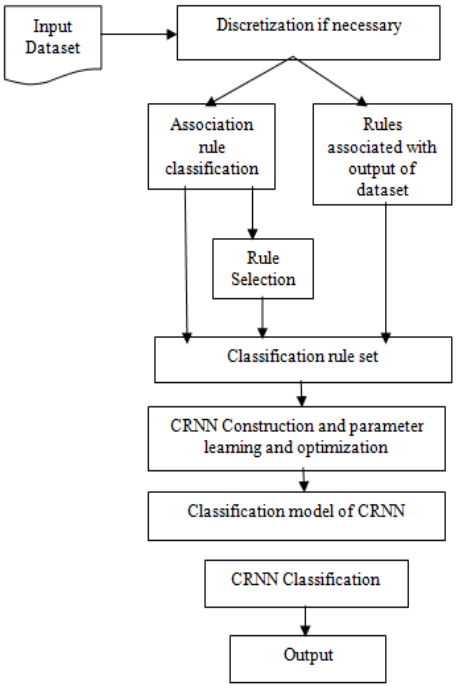
The main procedure of Top K Rules is shown above. The algorithm firstly scans the database once to calculate tids({c}) for each single item c in the database (line 1). Then, it generates all valid rules of size 1*1 by considering each pair of items i, j, where i and j each have at least minsup×|T| tids (if this condition is not satisfied, clearly, no rule having at least the minimum support can be created with i, j) (line 2). The supports of the rules {i}→{j} and {j}→{i} are easily obtained by dividing |tids(i→ j)| by |T| and |tids(j→ i)| by |T| (line 3 and 4). The confidence of the rules {i}→{j} and {j}→{i} is easily obtained by dividing |tids(i→ j)| by |tids(i)| and |tids(j→ i)| by | tids(j)| (line 5 and 6). Then, for each rule {i}→{j} or {j}→{i} which is valid, the procedure SAVE is called with the rule and L as parameters so that the rule is recorded in the set L of the current top-k rules found (line 7 to 9). Also, each rule {i}→{j} or {j}→{i} which is frequent gets added to the set R, to be later considered for expansion and a special flag named expandLR is set to true for each of such rule (line 10 to 12).

After that, a loop is performed to recursively select the rule r with the highest support in R such that sup(r) ≥ minsup and expand it (line 15 to 23). The idea is to always expand the rule that has the highest support because it is more likely to generate rules having a high support and thus to allow to raise minsup more quickly for reducing the search space. When there is no more rules in R with a support higher than minsup the loop terminated. For each of the rule, a flag variable expandLR indicates if the rule should be left and right expanded by calling the procedure EXPAND-L and EXPAND-R or just left expanded by calling EXPAND-L. For all the rules of size 1*1, this flag is set to true.

IV. PROPOSED METHOD

In this paper we are going to implement Top K Rule Mining algorithm. The flowchart of the proposed Association rule mining in figure1. Firstly the dataset is given to the system as the input. The discretization is applied over the input dataset to obtain the association rules from the dataset. Classification rule set is obtained from the discretization. Later Classification Rule Neural Network (CRNN) construction and the parameter learning set are obtained from the classification rule set. This constructor and learning parameter is now used to form the classification model of the CRNN. Here now we apply the testing dataset with the unknown output as the input to the classification model of the CRNN to find the known rules from the input. The output obtain from the CRNN classification is the desired output dataset.

Figure:



Block Diagram of Association rule mining

V. EXPERIMENTAL RESULT

Once In paper we have Top K Rules for Association Rule Mining. We collected demo datasets of online retailing e-commerce websites Flipkart and Amazon. Each dataset contains 1000, 2000, 5000 and 10000 entries, and is available for Association Rule Mining. We applied these datasets to Top K Rule Mining the algorithms and found out the results in terms of memory and time complexity.

We evaluated system with a different length of input dataset in order to get the desired data mining results. The mining results obtained are shown as follows,

Table 1. Experimental results of proposed Top K Rules Method

Dataset	Time	Memory

1000	24	8.24
2000	171	15.489
5000	133	10.48
10000	235	16.92

VI. CONCLUSION

In this paper we successfully study and implement the data mining system using Top K Rules for the association rule mining on e-commerce dataset. We applied these datasets to Top K Rule Mining the algorithms and found out the results in terms of memory and time complexity.

REFERENCES

1. *Neelamadhab Padhy, Dr. Pragnyaban Mishra, Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope" at International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.*
2. *Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.*
3. *Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.*
4. *Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006*
5. *Liu Jian, Wang Yan-Qing, "Web Log Data Mining Based on Association Rule" at Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011.*
6. *Changchien, S. W. and Lu, T.-C. (2001). Mining association rules procedure to support on-line recommendation by customers and products fragmentation. Expert Systems with Applications, 20(4):325–335.*
7. *Gupta, G., Strehl, A., and Ghosh, J. (1999). Distance based clustering of association rules.*
8. *Chai Wenguang, Tan Chunhui, Duan Yuting, "Research of Intelligent Intrusion Detection System Based On Web Data Mining Technology", at Fourth International Conference on Business Intelligence and Financial Engineering, 2011.*
9. *R. Agrawal, T. Imielminski and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," Proc. ACM Intern. Conf. on Management of Data, ACM Press, June 1993, pp. 207-216.*
10. *P. Tzvetkov, X. Yan and J. Han, "TSP: Mining Top-k Closed Sequential Patterns", Knowledge and Information Systems, vol. 7, no. 4, 2005, pp. 438-457.*
11. *C. Kun Ta, J.-L. Huang and M.-S. Chen, "Mining Top-k Frequent Patterns in the Presence of the Memory Constraint," VLDB Journal, vol. 17, no. 5, 2008, pp. 1321-1344.*
12. *J. Wang, Y. Lu and P. Tzvetkov, "Mining Top-k Frequent Closed Itemsets," IEEE Trans. Knowledge and Data Engineering, vol. 17, no. 5, 2005, pp. 652-664.*
13. *A. Pietracaprina and F. Vandin, "Efficient Incremental Mining of Top-k Frequent Closed Itemsets," Proc. Tenth. Intern. Conf. Discovery Science, Oct. 2004, Springer, pp. 275-280.*
14. *G. I. Webb and S. Zhang, "k-Optimal-Rule-Discovery," Data Mining and Knowledge Discovery, vol. 10, no. 1, 2005, pp. 39-79.*
15. *G. I. Webb, "Filtered top-k association discovery," WIREs Data Mining and Knowledge Discovery, vol.1, 2011, pp. 183-192.*
16. *Y. You, J. Zhang, Z. Yang and G. Liu, "Mining Top-k Fault Tolerant Association Rules by Redundant Pattern Disambiguation in Data Streams," Proc. 2010 Intern. Conf. Intelligent Computing and Cognitive Informatics, March 2010, IEEE Press, pp. 470-473.*