# An Approach to Clustering Marketing Data

Dongwoo Won
Computer Science Department
University of Southern California
dwon@usc.edu

Bo Mi Song
Computer Science Department
University of Southern California
bsong@usc.edu

Dennis McLeod
Computer Science Department
University of Southern California
mcleod@usc.edu

***ABSTRACT:*** *Association rule mining has been the main technique to identify meaningful patterns in market basket data (also known as synthetic transactional data). However, this approach has several limitations. Especially the large number of rules that have been generated from association creates difficulties in extracting and analyzing useful information. So, clustering association rules is introduced to overcome the problem. In this paper, we discuss three combined approaches to cluster association rules. First, we create hierarchical concepts and relationships of domain ontologies to merge and simplify items into more general concepts. This significantly reduces the total number of rules that are created. Adding to this, subspace clustering further reduces the resulting association rules to a more relevant problem space according to the level of details. Finally, a new ranking function is introduced to give more emphasis to the meaningful clustered groups via frequency, recency, and monetary value, which are the most important factors in the marketing domain. Our experiment shows that the total number of rules has decreased while the value of lift has increased. Also, the result of hierarchical subspace clustering shows that our approach groups data effectively to more simple, understandable and relevant problem space that is more suitable for further analysis.*

***Keywords:*** *Data mining, association rules, subspace hierarchical clustering, ranking function, domain ontology*

## 1. INTRODUCTION

Currently, large market basket data has been collected by many organizations, since this data is considered to be the best information for marketing analysis. Each row of data has a transaction that includes a set of items and customer information. Mining association rules [1] are frequently used to identify important patterns in such a transactional database. However, association rules mining tends to produce too many rules to obtain relevant and useful information because of the high dimensionality and sparseness of market basket data [2]. In response, clustering association rules were introduced to reduce the total number of rules needed. Pruning, grouping, and combining are carried out to create more general rules [3, 4]. A drawback of these approaches is that they still produce a large number of rules. Also, the relationship and relevance among the clusters tends to be difficult to understand.

In this paper, we discuss three combined approaches to cluster association rules. First, we suggest two methods: pre- and post-operative methods. Our use of ontologies allows us to have pre- and post- knowledge about the item set. Ontologies provide a means to represent information or knowledge that includes the key concepts and the inter-relationships between them, depending on the domain. We have applied this idea in generalizing and reducing the item set which as a result produces fewer, but more closely associated rules. Next, the second approach is hierarchical subspace clustering. Subspace clustering searches relevant sets of attributes in market basket data and localizes the search space to find clusters that allow for overlapping subspaces [5, 6]. We discover the most relevant attributes using clustered association rules in each level of our domain ontologies. To reduce the total number of rules, we combine similar association rules by clustering association rules [3]. As a result, this hierarchical subspace clustering approach can analyze market basket data efficiently and accurately. Finally, we propose a new cluster ranking function. There have been few attempts in the past to rank clusters as most prior work has been in the area of document ranking [7, 8]. In this paper, we rank the cluster resulting from subspace clustering. In the field of molecular biology, an attempt was made to rank interesting subspaces for high dimensional data to cluster data [9], but this effort is mathematically too complicated and difficult to apply in a marketing application. Here, we introduce a cluster ranking function, S = Frequency * Recency * Monetary (*FRM*), that relies on Frequency, Recency, and

Monetary factors, which are considered key terms in the field of marketing. In doing this, we have a new simple ranking function that is easy to understand and a better fit to the marketing domain.

The combination of our three approaches produces as a result, more relevant clusters of item sets that are easy to understand and easy to analyze for a marketing application. The rest of this paper is organized as follows. Our three approaches of using domain ontologies, subspace hierarchical clustering, and cluster-ranking functions are presented in section 2. In section 3 we discuss our evaluation and experimental results. We summarize our conclusion and discuss our future work in section 4.

## 2. APPROACH

### 2.1 Domain ontology

In our first approach, we use ontologies to reduce association rules by generalizing the existing item sets and compressing association rules. We suggest two techniques: pre- and post-operative methods. A pre-operative process involves building domain ontologies manually and using these ontologies throughout the operation. The post-operative step entails modifying existing ontologies automatically using the resulting subspace. Once we build ontologies, we only have to refer to them instead of needing to subspace cluster the whole item set every time we perform the analysis. In this and the next section, we concentrate on describing pre-operative methods, and we mention post-operative methods in section 4 as our future work.

In Figure 1, we have shown the sample item domain ontologies. All items with general concepts are located in the second level of the tree. Each item has its local item ontology with all the detail concepts as children nodes. The dashed line refers to the omitted nodes and vertices between them.

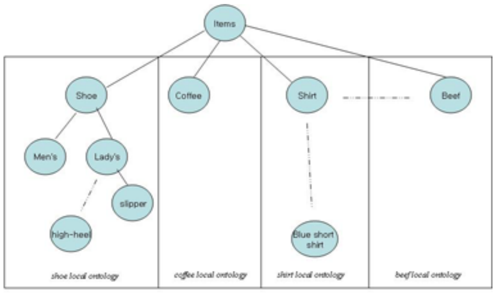
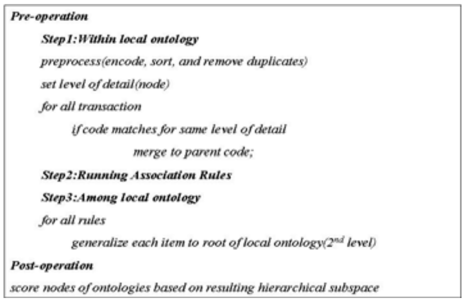


Figure 1 Item domain ontologies

Our pre-operative method requires two steps. First, the items per transaction are generalized. If at least two items in a row belong to a same local ontology, we merge the leaf node concept to a parent node concept. In this example, "high-heel" and "slipper" are both included in the more general category of "lady's shoe". If the level of the tree is really high, this process does not have much influence on the overall analysis. For example, "blue leather 5.6 size 'A' brand high-heel" and "red leather 7.5 size 'A' brand high-heel" merges to parent node "leather 'A' brand high-heel". Hence, the parent node includes detail concepts of leaf nodes within the local ontology. Next, we generate the association rules with the new item concepts. These rules can be used for more detailed analysis such as locating similar items in same section in a market. The second step is to generalize the resulting association rules to the second level concept. For "Men's Shoe", "Coffee", and "Blue short shirt", these concepts are all from a different local ontology. In this case, we use the generalized concept ($2^{nd}$ level) that results in "Shoe", "Coffee", and "Shirt". These generalized rules are used in our second approach for subspace clustering. We discuss this in more detail in Section 2.2. The simple steps of this process are shown below in figure 2.



```
Pre-operation
    Step1:Within local ontology
    preprocess(encode, sort, and remove duplicates)
    set level of detail(node)
    for all transaction
        if code matches for same level of detail
                merge to parent code;
    Step2:Running Association Rules
    Step3:Among local ontology
    for all rules
        generalize each item to root of local ontology($2^{nd}$ level)
Post-operation
score nodes of ontologies based on resulting hierarchical subspace
```

Figure 2 Steps for generalizing association rules

### 2.2 Hierarchical subspace clustering

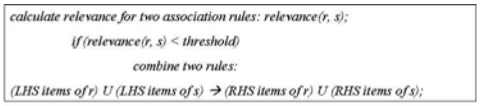
As we have briefly mentioned in Section 2.1, our second method is to create hierarchical clustered association rules based on the rules generated by approach 1. We generalized the association rules via the highest level of items such as shoe, coffee, shirt, beer, and so on in our item domain ontologies. From the generalized association rules, we can find uncorrelated generalized rules and separate them when we find a subspace and pruned small frequency generalized rules. For example, D -> D and B -> B don't have any correlation so we don't need to consider the association rules generalized to B -> B when we find a subspace within D -> D.

Then, we need to calculate the relevance of the two association rules for clustering relevant association rules within each generalized rule. The relevance between two association rules $r_1$ and $r_2$ is

$$relevance(r_1, r_2) = \frac{2\{LHS\, in\, r_1 \cap LHS\, in\, r_2\} + \{LHS\, in\, r_1 \cap RHS\, in\, r_2\} + \{RHS\, in\, r_1 \cap LHS\, in\, r_2\}}{\{item\, in\, r_1\} \cup \{item\, in\, r_2\}}$$

$$where\ r_1 : LHS \Rightarrow RHS, r_2 : LHS \Rightarrow RHS \qquad \ldots (1)$$

Here in our formula (1), $r_1$ and $r_2$ are association rules. The relevance is the total number of common items in the two association rules over the number of total items. With this formula, we give more weight to LHS (Left hand side) common items in the association rules. We also need to decide the threshold to cluster rules. Two association rules that have a larger relevance compared to the threshold are combined and become a clustered association rule. Figure 3 shows the general algorithm for constructing hierarchical clustered association rules.



```
calculate relevance for two association rules: relevance(r, s);
        if (relevance(r, s) < threshold)
                combine two rules:
(LHS items of r) U (LHS items of s) → (RHS items of r) U (RHS items of s);
```

Figure 3 General algorithms for hierarchical clustering

For example, there are two association rules: A & B → C and B & D → C. If the relevance of the two rules is larger than the threshold, the two rules are compressed to A & B & D → C. Through the algorithm in Figure 3, we can get the hierarchical association rule subspace within each generalized rule. In addition, the hierarchical subspace can be overlapped when two generalized rules are correlated. For instance, X & Y → Z and X → Z are generalized nodes and there is a common generalized item X in both nodes. In this case, the hierarchical subspace is overlapped between them. At last, we need to consider only subspaces for clustering. Figure 4 shows clusters searched within subspaces.

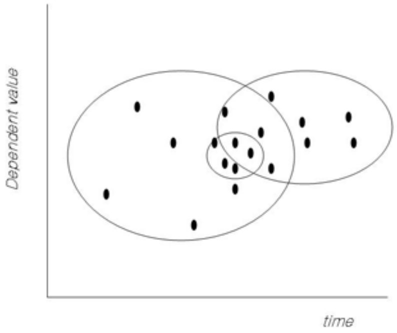


Figure 4 Hierarchical clustering

The hierarchical subspace clustering can handle significant outliers. Since the outlier association rule is meaningful in a market data set, we define higher minimum threshold values for support and confidence that are very high compared to a normal support and confidence value that would take care of this case. Therefore, if there is an association rule with a high support and confidence value, even though the rule should be combined or pruned via the approaches above, we do not compress the rule to the upper level or prune and preserve the rule itself. In this case, a promotion event can be one way of handling this instance in real market.

## 2.3 Cluster ranking

As a result of hierarchical clustering, it is possible to look at the rules by controlling the view level. But, there are many clusters, and it is difficult to know which cluster has more importance than another. Ranking the cluster, however allows us to differentiate. We introduce a new scoring function that stems from the definition of a magnetic field in physics. Looking at Figure 5, a magnetic field can be compared to a cluster. Each data point, in our case the resulting rules can be the electric charge, *q*. Each data point tends to move with high velocity, *v* toward the other magnetic field when it has more force, *F*.

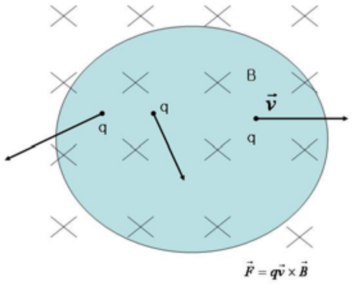


Figure 5 Magnetic field compared to a cluster

Hence, we define our ranking function grounding on magnetic field theory,

$$S = FRM \qquad \ldots (2)$$

*S* stands for *score*, *F* for *frequency*, *R* for *recency*, and *M* for *monetary*. These three factors are considered highly important in marketing applications. *Frequency* means how often particular items have been bought during a specific time period. So in our case, *frequency* depends on the number of data points and the area of the cluster. We consider high *frequency* when the number of data points is high and the area is small. So, we give a compact area with more data points more weight. For the area, we consider an eclipse shaped cluster and solve for that area. *Recency* means how recently the items were bought, so it depends on the value of time. We consider the average time value for all data points in the cluster. *Monetary* is the price factor. In this paper, we do not consider the price set and hence we set this value to the default value 1. From equation (2), we can derive equation (3),

$$S = [number\ of\ \ points\ in\ cluster/(area\ of\ \ cluster)\ ][\ time\ ][\ price\ ]$$

$$S = [\ N_c\ /\ area\ \ of\ \ ecllipse\ \ ] \\ *[(\sum_{x=1}^{N_c}(t_x)\ /\ N_c)] \qquad \ldots (3)$$

$N_c$: number of points in a cluster
$t$: value of time

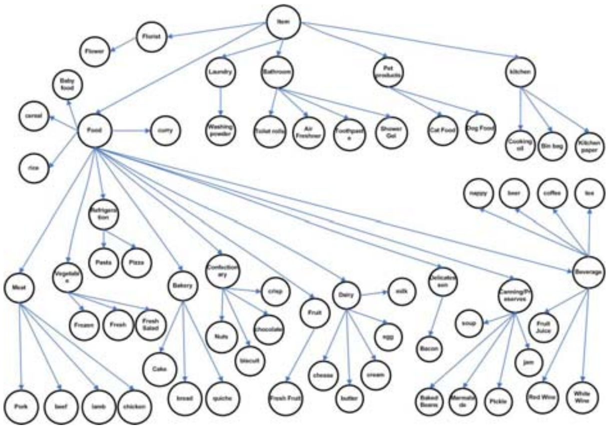


Figure 6 Supermarket Item ontologies

## 3. RESULTS

Our experiment is performed with a transaction based market basket data set with 600 rows of main data that includes customer id, time, and item codes. We also maintain an index table of 49 item sets that includes the name and code of items. First, we did an experiment with SAS Enterprise Miner [10] to see the general association rules for our item set. As we expected, a large rule set totaling 4018 rules was created. For our approach, we have represented our market basket item set as domain ontologies (Figure 6). Using these ontologies, we generate association rules by reducing the number of items through generalizing the leaf node concept to its parent node as we have explained in Section 2.1.

We sort result by confidence, support, and lift which is provided by Enterprise Miner. In Table 1, we compare the number of rules created and the values produced with our approach. Although the number of rules has been reduced dramatically, the values of confidence and support have been kept constant as before. Normally, we can expect a lower value of confidence and support for the reduction of items. But as we can see, it does not make any difference with our approach. Add to this, we have an increase in lift value. Lift refers to the percentage increase in observed co-occurrences over expected co-occurrences i.e. higher lift, higher association [10].

| | | SAS | Our Approach |
|---|---|---|---|
| # Of Rules Created | | 4018 | 2618 |
| Confidence | 10th row | 100 | 100 |
| | 100th row | 71.43 | 71.43 |
| | 500th row | 55.56 | 50 |
| Support | 10th row | 2.5 | 2 |
| | 100th row | 2.5 | 2.5 |
| | 500th row | 2.5 | 2 |
| Lift | 10th row | 2.56 | 3.13 |
| | 100th row | 2.38 | 2.38 |
| | 500th row | 1.04 | 1.52 |

Table 1 Association Rule comparison

After generalizing and pruning the association rules, the total number of generalized rules is reduced a somewhat from 35 to 19 and we just consider an average of 135 association rules to create a hierarchical association rule and subspace within each generalized rule instead of considering all 2618 association rules.

The fundamental idea is to merge the association rules when the relevance value is reasonably large and then create a subspace using combined association rules. Figure 7 shows an example of how to create a subspace within a generalized rule: Food & Bathroom -> Bathroom. In the generalized rule, there are 32 association rules (r1 – r32) and there are several duplicate association rules among the 32 rules. The first step is to remove duplicate association rules. Since the relevance value between the same association rules is definitely larger than the relevance value between different association rules, according to the above relevance formula, we can remove duplicate rules by combining the same rules. After the first step, we get 18 rules (s1 – s18) and calculate the value of relevance between the two association rules again. Because the rules are already generalized, the value of relevance does not vary much. Therefore, in this example, there are only two relevance values; 1.33 and 0.5. We choose 1.0 as a threshold in this second step and combine the two association rules whose relevance value is 1.33 (>1.0). The second row of Figure 4 shows the combined rules and we get 10 association rules (t1-t10). At last, we calculate the relevance values among 10 association rules and the value is the same among the 8 association rules (t1- t8). Among the 8 rules, the relevance value of any two association rules is 1.5 (> 1.0). Finally, we get a large clustered hierarchical association rule called subspace.

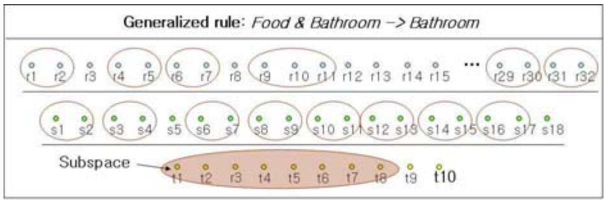

Generalized rule: *Food & Bathroom –> Bathroom*
r1 r2 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 ... r29 r30 r31 r32
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18
Subspace
t1 t2 r3 t4 t5 t6 t7 t8 t9 t10

Figure 7 Subspace clustering

## 4. Conclusion and future work

In this paper, we have showed three combined approaches to cluster association rules. First, we created domain ontologies to merge and simplify items into more general concepts. Our result shows that this highly reduces the amount of association rules that are generated. Next, subspace clustering once more reduces the resulting association rules to a more relevant problem space, according to the level of details. Our experiment represents the possibility of removing unrelated rules and grouping the closely related rules together for a simple and easy analysis. Finally, a new ranking function was introduced.

This work is still in the early stage. Future work will provide mathematical weight and meaning to the nodes and edges for our ontologies. This will provide us more plausible calculations for creating generalized association rules. At this stage, we have done top-down hierarchical clustering, but we will also make an approach for bottom-up hierarchical clustering. Developing a plausible threshold is one of our tasks. Lastly, we will conduct a thorough experiment on our ranking function. This will also be used in our post-operative ontologies creation that will automatically help to enhance our existing ontologies.

## Acknowledgements

This research was supported in part by the Integrated Media Systems Center, A National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152, and in part by NASA's Computational Technologies Project - portions of this work were carried out by the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA.

## References

[1] Agrawal R., Imielinski T., and Swami A., “Mining association rules between sets of items in large databases”, In *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 207-216, Washington, DC, May 1993.

[2] Strehl A. and Ghosh J., “A Scalable Approach to Balanced, High-dimensional Clustering of Market-baskets”, In *Proc. of the 7th International Conference on High Performance Computing*, December 2000.

[3] Lent B., Swami A., and Widom J., “Clustering Association Rules”, In *Proc. of 13th International Conference on Data Engineering*, pp. 1-19, U.K., April 1997.

[4] Toivonen H., Klemettinen M., Ronkainen P., Hdtvnen K., and Mannila H., "Pruning and grouping discovered association rules", In *MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases*, pp. 47-52, Heraklion, Crete, Greece, April 1995.

[5] Parsons L., Haque E., and Liu H., “Subspace Clustering for High Dimensional Data: a Review”, *ACM SIGKDD Explorations Newsletter*. 6(1), pp. 90-105, 2004

[6] Agrawal R., Gehrke J., Gunopulos D., and Raghavan P., “Automatic subspace clustering of high dimensional data for data mining applications”, In *Proc. of the 1998 ACM SIGMOD international conference on Management of data*, pp. 94-105, ACM Press, 1998.

[7] Tombros A., Jose J.M., Ruthven I., “Clustering Top-Ranking Sentences for Information Access”, In *Proc. of the 7th European Conference on Digital Libraries*, pp. 523-528, Trondheim, Norway, 2003.

[8] Evans D. and McKeown K., “Identifying similarities and differences across english and arabic news”, In *Proc. of International Conference on Intelligence Analysis*, pp. 23-30, McLean, VA., 2005.

[9] Kailing K., Kriegel H.-P., Kröger P., Wanka S., “Ranking Interesting Subspaces for Clustering High Dimensional Data”, In *Proc. 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'03)*, Vol. 2838, pp.241-252, Dubrovinic, Croatia, 2003.

[10] Enterprise Miner, SAS Institute, http://www.sas.com/technologies/analytics/datamining/miner/

[11] Gupta G., Strehl A., and Ghosh J., “Distance based clustering of association rules”. In *Proc. of Intelligent, Engineering Systems through Artificial Neural Networks (ANNIE)*, Vol. 9, pp.759-764. ASME Press, St. Louis, Nov. 1999.

[12] Jorge A., “Hierarchical Clustering for thematic browsing and summarization of large sets of association rules”, *International Conference on Data Mining, SIAM SDM 04*, Orlando, Florida, 2004