Clustering Item Data Sets with Association-Taxonomy Similarity

Ching-Huang Yun, Kun-Ta Chuang+ and Ming-Syan Chen
Department of Electrical Engineering
Graduate Institute of Communication Engineering+
National Taiwan University
Taipei, Taiwan, ROC
E-mail: chyun@arbor.ee.ntu.edu.tw, doug@arbor.ee.ntu.edu.tw, mschen@cc.ee.ntu.edu.tw

Abstract

We explore in this paper the efficient clustering of item data. Different from those of the traditional data, the features of item data are known to be of high dimensionality and sparsity. In view of the features of item data, we devise in this paper a novel measurement, called the association-taxonomy similarity*, and utilize this measurement to perform the clustering. With this association-taxonomy similarity measurement, we develop an efficient clustering algorithm, called algorithm* AT *(standing for Association-Taxonomy), for item data. Two validation indexes based on association and taxonomy properties are also devised to assess the quality of clustering for item data. As validated by both real and synthetic datasets, it is shown by our experimental results that algorithm AT devised in this paper significantly outperforms the prior works in the clustering quality as measured by the validation indexes, indicating the usefulness of association-taxonomy similarity in item data clustering.*

1 Introduction

Data clustering is an important technique for exploratory data analysis. Data clustering is an application dependent issue and certain applications may call for their own specific requirements. Different from those of the traditional data, the features of market-basket data are known to be of high dimensionality and sparsity. There are several clustering technologies which addressed the issue of clustering market-basket data [2][3][4][5][6].

Explicitly, the support of item i is defined as the percentage of transactions which contain i. Note that in mining association rules, a *large item* is basically an item with frequent occurrence in transactions. Thus, item i is called a large item if the support of item i is larger than the pre-given minimum support threshold. In market-basket data, the taxonomy of items defines the generalization relationships for the concepts in different abstraction levels.

In view of the features of item data, we devise in this paper a novel measurement, called the *association-taxonomy similarity*, and utilize this measurement to perform the clustering for shelf-space organization. With this association-taxonomy similarity measurement, we develop an efficient clustering algorithm, called algorithm *AT* (standing for *Association Taxonomy*), for item data. Two validation indexes, *association index* (abbreviated as *AI*) and *taxonomy index* (abbreviated as *TI*), are also devised in this paper for clustering item data. As validated by real data, it is shown by our experimental results, with the association and taxonomy information, algorithm AT devised in this paper significantly outperforms the prior works [2][3] in the clustering quality.

2 Preliminaries

In market-basket data, a database of transactions is denoted by $D = \{t_1, t_2, ..., t_v\}$, where each transaction t_h is a set of items $\{i_1, i_2, ..., i_w\}$. In mining association rules [1], the minimum support Sup is given to identify the large itemsets. In addition, the support of an itemset in database D is defined as the number of transactions which contain this itemset in database D. An itemset is called a *large itemset* if its support is larger than or equal to the minimum support Sup. In this paper, an *association itemset* is defined as a large itemset that contains at least two items and is not contained by any other large itemset. The set of association itemsets is denoted by $L_A = \{I_1, I_2, ..., I_m\}$. Items in the transactions can be generalized to multiple concept levels of the taxonomy and represented as a taxonomy tree. In the taxonomy tree, the leaf nodes are called the *item nodes* and the internal nodes are called the *category nodes*.

In view of the features of item data, the items are categorized into three kinds of items which are *association items* (represented as I_A), *single large items* (represented as I_S), and *rare items* (represented as I_R). An association item is an item which appears in at least one association itemset. A single large item is a large item but not an association item. In essence, a single large item can be viewed as a large 1-item which is not contained by any large 2-itemset. A rare item is not a large item (i.e., not frequently purchased). Explicitly, the rare item is an item whose support is smaller than the minimum support.

In this paper, a clustering $U =< C_1, C_2, ..., C_k >$ is a partition of items into k clusters, where C_j is a cluster consisting of a set of items. Note that purchasing relationships (i.e., association) and taxonomy relationships are important for the shelf-space organization. In this paper, the objective of clustering item data is to cluster the items with high association relationships and high taxonomy relationships together.

In view of the features of item data, we propose association index and taxonomy index, which are defined below, to assess the qualities of the clustering results.

Definition 1: (**Association Index**) The association index of the clustering U is defined as:

$$AI(U) = \frac{\sum\limits_{C_p \in U} \left(\frac{\sum\limits_{(i_x, i_y) \in C_p} A(i_x, i_y)}{\frac{1}{2}(|C_p|)(|C_p|-1)} \right)}{|U|},$$

where $A(i_x, i_y)$ is the association value of item i_x and item i_y. Explicitly, $A(i_x, i_y) = 1$, if i_x and i_y are in the same association itemset based on the minimum support Sup, and $A(i_x, i_y) = 0$, otherwise.

Definition 2: (**Taxonomy Index**) The taxonomy index of the clustering U is defined as:

$$TI(U) = \frac{\sum\limits_{C_p \in U} \left(\frac{\sum\limits_{(i_x, i_y) \in C_p} T(i_x, i_y)}{\frac{1}{2}(|C_p|)(|C_p|-1)} \right)}{|U|},$$

where $T(i_x, i_y)$ is the taxonomy value of item i_x and item i_y. Explicitly, $T(i_x, i_y) = 1$, if i_x and i_y are in the same category under the cluster level Lev^C, and $T(i_x, i_y) = 0$, otherwise. In this paper, the cluster level Lev^C is defined as the level where the number of categories is equal to the number of clusters k.

3 Design of Algorithm AT (Association Taxonomy)

In this paper, we devise algorithm AT for clustering item data. The similarity measurement of AT will be described in Section 3.1. Section 3.2 describes the procedure of AT.

3.1 Similarity Measurement

The similarity measurement employed by algorithm AT is called association-taxonomy similarity which consists of the association similarity and the taxonomy similarity. As described before, the set of association itemsets is denoted by $L_A = \{I_1, I_2, ..., I_m\}$. For each association itemset, the association relationships of items can be represented as a complete graph $I_p = \{V_p, E_p\}$, consisting of a set of vertices V_p and a set of edges E_p. In each complete graph, each vertex represents an item in the association itemset and each edge represents the association between two items. In mining association rules, an association rule $i_x \rightarrow i_y$ holds in transaction database D with confidence $Con(i_x \rightarrow i_y)$ if $Con(i_x \rightarrow i_y)$ of transactions in D that contain i_x also contain i_y. In this paper, we use *co-confidence* as the measurement of the association between two items.

Definition 3: (**Co-Confidence between Association Items**) The co-confidence between i_x and i_y is defined as:

$$\begin{aligned} e(i_x, i_y) &= \frac{1}{2}(Con(i_x \rightarrow i_y) + Con(i_y \rightarrow i_x)) \\ &= \frac{1}{2}\left(\frac{Sup(i_x i_y)}{Sup(i_x)} + \frac{Sup(i_x i_y)}{Sup(i_y)}\right) \end{aligned}$$

where $Sup(i_x)$ is the support of item i_x. The co-confidence $e(i_x, i_y)$ represents the association between item i_x and item i_y.

Each association itemset is viewed as a cluster of items (i.e., $C_p = I_p$). For notational simplicity, the union cluster of C_p and C_q is denoted as $C_{p,q}$. The set of overlapped items in $C_{p,q}$ is denoted as $C^o_{p,q}$ and the set of non-overlapped items in $C_{p,q}$ is denoted as $C^n_{p,q}$. In addition, $E_{C_{p,q}}$ denotes the set of edges in $C_{p,q}$, $E^{oo}_{C_{p,q}}$ denotes the set of edges connecting the overlapped items in $C_{p,q}$, $E^{on}_{C_{p,q}}$ denotes the set of edges connecting the overlapped items and non-overlapped items in $C_{p,q}$, and $E^{nn}_{C_{p,q}}$ denotes the set of edges connecting the non-overlapped items in $C_{p,q}$.

Definition 4: (**Association Similarity between overlapped items**) The association similarity between overlapped items of C_p and C_q is defined as:

$$AS_{oo}(C_p, C_q) = \frac{\sum\limits_{i_x \in C^o_{p,q}, i_y \in C^o_{p,q}} e(i_x, i_y)}{|E^{oo}_{C_{p,q}}| + |E^{nn}_{C_{p,q}}|}$$

Definition 5: (**Association Similarity between overlapped items and non-overlapped items**) The association similarity between overlapped items and non-overlapped items of C_p and C_q is defined as:

$$AS_{on}(C_p, C_q) = \frac{\sum\limits_{i_x \in C^o_{p,q}, i_z \in C^n_{p,q}} e(i_x, i_z)}{|E^{on}_{C_{p,q}}| + |E^{nn}_{C_{p,q}}|}$$

For the similarity measurements in Definition 4 and Definition 5, $|E^{nn}_{C_{p,q}}|$ is a normalization factor for considering the effect of the edges of non-overlapped items in decreasing the similarity between two clusters. Explicitly, the existence of non-overlapped items represents the dissimilarity between two clusters. Thus, an edge between the non-overlapped items increases the association dissimilarity between two clusters.

Definition 6: (**Association Similarity**) The association similarity between C_p and C_q is defined as:

$$AS(C_p, C_q) = \alpha_{oo} * AS_{oo}(C_p, C_q) + \alpha_{on} * AS_{on}(C_p, C_q),$$

where α_{oo} is the weight of the association similarity between overlapped items and α_{on} is the weight of the association similarity between overlapped items and non-overlapped items.

Definition 7: (**Taxonomy similarity of an overlapped Item**) The taxonomy similarity of overlapped item i_x to union cluster $C_{p,q}$ is defined as:

$$T_o(i_x, C_{p,q}) = \sum_{k=1}^{N^{Lev}} \frac{|C_{p,q}(i_x, k)|}{k},$$

where N^{Lev} is the number of levels in the taxonomy tree and $C_{p,q}(i_x, k)$ is the set of items which is in the same category with item i_x in level k in $C_{p,q}$.

Definition 8: (**Taxonomy Similarity of overlapped items**) The taxonomy similarity of overlapped items of C_p and C_q is defined as:

$$TS_o(C_p, C_q) = \frac{\sum_{i_x \in C^o_{p,q}} T_o(i_x, C_{p,q})}{|C^o_{p,q}| * (|C_{p,q}| - 1)},$$

Definition 9: (**Taxonomy similarity of a non-overlapped Item**) Let i_y be an item in C_p and i_y is not overlapped with any item in C_q. The taxonomy similarity of non-overlapped item i_y in cluster C_p to cluster C_q is defined as:

$$T_n(i_y, C_q) = \sum_{k=1}^{N^{Lev}} \frac{|C_q(i_y, k)|}{k},$$

where $C_q(i_y, k)$ is the set of items which is in the same category with item i_y in level k in C_q.

Definition 10: (**Taxonomy Similarity of non-overlapped items**) The taxonomy similarity of non-overlapped items of C_p and C_q is defined as:

$$TS_n(C_p, C_q) = \frac{\sum_{i_y \in C^n_p} T_n(i_y, C_q) + \sum_{i_z \in C^n_q} T_n(i_z, C_p)}{|C^n_p| * |C_q| + |C^n_q| * |C_p|},$$

Definition 11: (**Taxonomy Similarity**) The taxonomy similarity between C_p and C_q is defined as:

$$TS(C_p, C_q) = \beta_o * TS_o(C_p, C_q) + \beta_n * (TS_n(C_p, C_q) - \frac{1}{N^{Lev}}),$$

where β_o is the weight of the taxonomy similarity of overlapped items and β_n is the weight of the taxonomy similarity of non-overlapped items. If each item in C_p and each item in C_q only have the root node as the same category, C_p is totally dissimilar to C_q according to the taxonomy tree and $TS(C_p, C_q)$ should be zero. Hence, because there are no overlapped item between C_p and C_q, the constant $\frac{1}{N^{Lev}}$ is subtracted in the non-overlapped part for normalization purpose.

Definition 12: (**Association-Taxonomy Similarity**) The association-taxonomy similarity between C_p and C_q is denoted as $SIM(C_p, C_q)$ defined as:

$$SIM(C_p, C_q) = \varpi_A * AS(C_p, C_q) + \varpi_T * TS(C_p, C_q),$$

where ϖ_A is the weight of the association similarity and ϖ_T is the weight of the taxonomy similarity. The determination of values of ϖ_A and ϖ_T is in fact application-dependent.

3.2 Procedure of Algorithm AT

Algorithm AT is designed to consist of three phases: the segmentation phase, the association-taxonomy phase, and the pure-taxonomy phase. Note that the association items consist of the elements in association itemsets. The overall procedure of algorithm AT is outlined as follows.

Procedure of Algorithm AT (Association-Taxonomy)

(1) The Segmentation Phase:

Step 1. Identify the set of association itemsets, the set of single large items, and the set of rare items.

(2) The Association-Taxonomy Phase:

Step 2. For each pair in the set of the association itemsets, calculate the corresponding association-taxonomy similarity.

Step 3. Merge the pair which has the largest association-taxonomy similarity as a new cluster.

Step 4. Repeat Step 2 and Step 3 until the dendrogram is constructed.

(3) The Pure-Taxonomy Phase:

Step 5. Identify k clusters in the dendrogram.

Step 6. For each single large item, allocate it to the cluster with the largest taxonomy similarity.

Step 7. For each rare item, allocate it to the cluster with the largest taxonomy similarity.

Step 8. Repeat Step 6 and Step 7 until no item is moved between clusters.

The advantageous features of algorithm AT are twofold. The first one is on employing the association-taxonomy similarity to effectively improve the quality of clustering association items. The second one is to allocate the single large items and rare items into clusters by calculating the taxonomy similarity. As such, these items can be efficiently and effectively allocated into the clusters. Note that the numbers of single large items and rare items are usually large as compared to the number of association itemsets. If we take each single large item (or each rare item) as a cluster and put them into the procedure from Step 2 to Step 4, the execution time will be prohibitive. In addition, lack of large association similarity with other clusters, these clusters with only one single large item (or one rare item) would never be merged until most of the association itemsets are merged. These problems are avoided in algorithm AT.

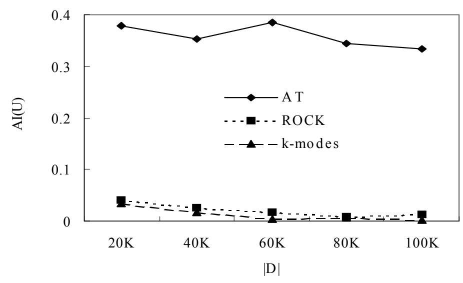
4 Experimental Studies

To assess the efficiency of AT, we conducted experiments to compare AT with the k-modes algorithm [3] and the ROCK algorithm [2]. We use the real market-basket data from a large bookstore company for performance study. In this real data set, there are $|D| = 100K$ transactions, $|I| = 58909$ items, and $N^{Lev} = 3$ levels. In addition, the number of the taxonomy level in this real data set is 3. In the real data, the items with the same category are usually purchased together. Thus, the association relationships and taxonomy relationships are related to each other.

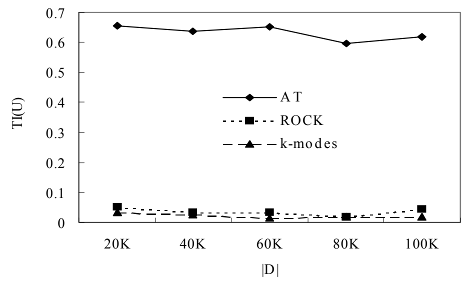
Figure 1 shows the relative quality of clustering results of AT, ROCK, and k-modes in real data set where the database size $|D|$ varies from 20K to 100K. When we vary $|D|$ from 20K to 100K in ROCK, the numbers of clusters are, respectively, 576, 524, 468, 413, and 519. With association-taxonomy similarity measurement, AT significantly outperforms other algorithms as validated by $AI(U)$ in Figure 1(a) and by $TI(U)$ in Figure 1(b). In this real data set, because the items with high taxonomy relationships are usually purchased together while the items with low taxonomy relationships are not, AT has higher taxonomy index than association index, i.e., $AI(U) > AI(U)$.

5 Conclusion

In this paper, with the association-taxonomy similarity measurement proposed, we developed algorithm AT for item data. Two validation indexes based on association and taxonomy features of items was also devised in this paper to assess the quality of clustering for item data. As validated by real data, it was shown by our experimental results that algorithm AT devised in this paper significantly outperforms the prior works in the clustering quality of item data.



(a) Association Index



(b) Taxonomy Index

Figure 1. $AI(U)$ and $TI(U)$ for algorithms when $|D|$ varies.

References

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 478–499, September 1994.

[2] S. Guha, R. Rastogi, and K. Shim. ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Proceedings of the 15th International Conference on Data Engineering*, 1999.

[3] Z. Huang. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3):283–304, September 1998.

[4] K. Wang, C. Xu, and B. Liu. Clustering Transactions Using Large Items. *Proceedings of ACM CIKM International Conference on Information and Knowledge Management*, 1999.

[5] C.-H. Yun, K.-T. Chuang, and M.-S. Chen. Self-Tuning Clustering: An Adaptive Clustering Method for Transaction Data. *Proc. of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2002)*, Sep. 2002.

[6] C.-H. Yun, K.-T. Chuang, and M.-S. Chen. Using Category-Based Adherence to Cluster Market-Basket Data. *Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM 2002)*, Dec. 2002.