Using Category-Based Adherence to Cluster Market-Basket Data

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

Abstract

In this paper, we devise an efficient algorithm for clustering market-basket data. Different from those of the traditional data, the features of market-basket data are known to be of high dimensionality, sparsity, and with massive outliers. Clustering transactions across different levels of the taxonomy is of great importance for marketing strategies as well as for the result representation of the clustering techniques for market-basket data. In view of the features of market-basket data, we devise in this paper a novel measurement, called the category-based adherence*, and utilize this measurement to perform the clustering. The distance of an item to a given cluster is defined as the number of links between this item and its nearest large node in the taxonomy tree where a large node is an item (i.e., leaf) or a category (i.e., internal) node whose occurrence count exceeds a given threshold. The category-based adherence of a transaction to a cluster is then defined as the average distance of the items in this transaction to that cluster. With this category-based adherence measurement, we develop an efficient clustering algorithm, called algorithm* CBA *(standing for Category-Based Adherence), for market-basket data with the objective to minimize the category-based adherence. A validation model based on* Information Gain *(*IG*) is also devised to assess the quality of clustering for market-basket data. As validated by both real and synthetic datasets, it is shown by our experimental results, with the taxonomy information, algorithm CBA devised in this paper significantly outperforms the prior works in both the execution efficiency and the clustering quality for market-basket data.*

1 Introduction

Data clustering is an important technique for exploratory data analysis [19]. Explicitly, data clustering is a well-known capability studied in information retrieval [7], data mining [8], machine learning [11], and statistical pattern recognition [18]. In essence, clustering is meant to divide a set of data items into some proper groups in such a way that items in the same group are as similar to one another as possible. Most clustering techniques utilize a pairwise similarity for measuring the distance of two data points. Recently, there has been a growing emphasis on clustering very large datasets to discover useful patterns and/or correlations among attributes [3][4][12][33]. Clustering large spatial databases tries to find the densely populated regions in the feature space [21][26]. Note that clustering is an application dependent issue and certain applications may call for their own specific requirements.

Market-basket data (also called transaction data) has been well studied in mining association rules for discovering the set of frequently purchased items [5][15][27]. Different from the traditional data, the features of market-basket data are known to be of high dimensionality, sparsity, and with massive outliers. Cobweb is a conceptual clustering technique by utilizing a clustering dendrogram called classification tree to characterize each cluster with a probabilistic description [11]. ROCK is an agglomerative hierarchical clustering algorithm by treating market-basket data as categorical data and using the links between the data points to cluster categorical data [13]. The authors in [22] proposed an EM-based algorithm by using the maximum likelihood estimation method for clustering transaction data. OPOSSUM is a graph-partitioning approach based on a similarity matrix to cluster transaction data [28]. The work in [29] proposed a K-Mean-based algorithm by using large items as the similarity measurement to divide the transactions into clusters such that transactions with similar large items are grouped into the same clusters. OAK in [30] combined hierarchical and partitional clustering techniques. SLR in [31] utilized a fixed small to large item ratio to perform the clustering of market-basket data. In market-basket data, the taxonomy of items defines the generalization relationships for the concepts in different abstraction levels [16]. Item taxonomy (i.e., *is-a* hierarchy) is well addressed with respect to its impact to mining association rules of market-basket database [15][27] and can be represented as a tree, called *taxonomy tree*. Similar techniques for extracting synonyms, hypernyms (i.e., a *kind* of) and holonyms (i.e., a *part* of) words from the lexical database are derived in [10][25].

In view of the features of market-basket data, we devise in this paper a novel measurement, called the *category-based adherence*, and utilize this measurement to perform the clustering. The *distance* of an item to a given cluster is defined as the number of links between this item and its nearest large node in the taxonomy tree. In the taxonomy tree, the leaf nodes are called the item nodes and the internal nodes are called the category nodes. For the example shown in Figure 1, "War and Peace" is an item node and "Novel" is a category node. As formally defined in Section 2, a *large item (category)* is basically an item (category) with its occurrence count in transactions exceeding a given threshold. If an item (or category) is large, its corresponding node in the taxonomy tree is called a *large node*. For the example shown in Figure 1, nodes marked gray are assumed to large nodes. The *category-based adherence* of a transaction to a cluster is then defined as the average distance of the items in this transaction to that cluster[1]. With this category-based adherence measurement, we develop an efficient clustering algorithm, called algorithm *CBA* (standing for *Category-Based Adherence*), for market-basket data. Explicitly, CBA employs the category-based adherence as the similarity measurement between transactions and clusters, and allocates each transaction to the cluster with the minimum category-based adherence. To the best of our knowledge, without explicitly considering the presence of the taxonomy, previous efforts on clustering market-basket data. [13][22][28][29][30][31] unavoidably restricted themselves to deal with the items in the leaf level (also called item level) of the taxonomy tree. However, clustering transactions across different levels of the taxonomy is of great importance for marketing strategies as well as for the result representation of the clustering techniques for market-basket data. Note that in the real market-basket data, there are volume of transactions containing only single items, and many items are purchased infrequently. Hence, without considering the taxonomy tree, one may inappropriately treat a transaction (such as the one containing "parallel compiler" in Figure 1) as an outlier. However, as indicated in Figure 1, purchasing "parallel compiler" is in fact instrumental for the category node "computer science" to become a large node. In contrast, by employing category-based adherence measurement for clustering, many transactions will not be mistakenly treated as outliers if we take categorical relationships of items in the taxonomy tree into consideration, thus leading to better marketing strategies. The details of CBA will be described in Section 3. A validation model based on *Information Gain* (IG) is also devised in this paper for clustering market-basket data. As validated by both real and synthetic datasets, it is shown by our experimental results, with the taxonomy information, algorithm CBA devised in this paper significantly outperforms the prior works [17][29] in both the execution efficiency and the clustering quality for market-basket data.

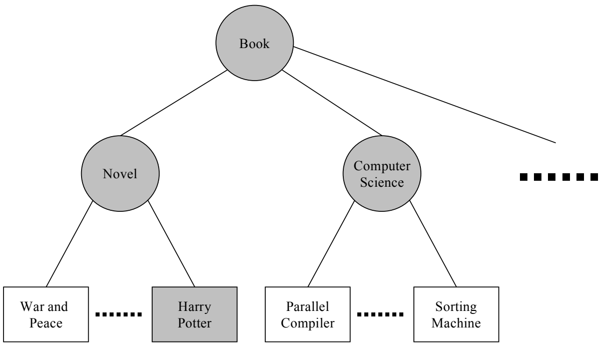


Figure 1. An example taxonomy tree for books.

This paper is organized as follows. Preliminaries are given in Section 2. In Section 3, algorithm CBA is devised for clustering market-basket data. Experimental studies are conducted in Section 4. This paper concludes with Section 5.

2 Preliminary

The problem description will be presented in Section 2.1. In Section 2.2, we describe a new validation model, *IG* validation model, for the assessment to the quality of different clustering algorithms.

2.1 Problem Description

In this paper, the market-basket data is represented by a set of transactions. A database of transactions is denoted by $D = \{t_1, t_2, ..., t_h\}$, where each transaction t_j

[1] The formal definitions of these terms will be given in Section 2.1.

is represented by a set of items $\{i_1, i_2, ..., i_h\}$. A clustering $U =< C_1, C_2, ..., C_k >$ is a partition of transactions into k clusters, where C_j is a cluster consisting of a set of transactions.

Items in the transactions can be generalized to multiple concept level of the taxonomy. In the taxonomy tree, the leaf nodes are called the *item nodes* and the internal nodes are called the *category nodes*. The root node in the highest level is a virtual concept of the generalization of all categories. In this taxonomy structure, item g *is-a* category B, category B *is-a* category A, and item h *is-a* category B, etc. In this paper, we use the measurement of the occurrence count to determine which items or categories are major features of each cluster.

Definition 1: The count of an item i_k in a cluster C_j, denoted by $Count(i_k, C_j)$, is defined as the number of transactions in cluster C_j that contain this item i_k. An item i_k in a cluster C_j is called a *large item* if $Count(i_k, C_j)$ exceeds a predetermined threshold.

Definition 2: The count of a category c_k in a cluster C_j, denoted by $Count(c_k, C_j)$, is defined as the number of transactions containing items under this category c_k in cluster C_j. A category c_k in a cluster C_j is called a *large category* if $Count(c_k, C_j)$ exceeds a predetermined threshold.

Note that one transaction may include more than one item from the same category, in which case the count contributed by this transaction to that category is still one. In this paper, the minimum support percentage S_p is a given parameter for determining the large nodes of the taxonomy tree in the cluster. For a cluster C_j, the minimum support count $S_c(C_j)$ is defined as follows.

Definition 3: For cluster C_j, the minimum support count $S_c(C_j)$ is defined as:

$$S_c(C_j) = S_p * |C_j|.$$

where $|C_j|$ denotes the number of transactions in cluster C_j.

2.2 Information Gain Validation Model

To evaluate the quality of clustering results, some experimental models were proposed [9][14]. In general, *square error criterion* is widely employed in evaluating the efficiency of numerical data clustering algorithms [9]. In addition, authors in [32] proposed a novel clustering validation scheme which uses the variance and the density of each cluster to measure the inter-cluster dissimilarity and the intra-cluster similarity. Note that the nature feature of numeric data is quantitative (e.g., weight or length), whereas that of categorical data is qualitative (e.g., color or gender) [19]. Thus, validation schemes using the concept of variance are thus not applicable to assessing the clustering result of categorical data. To remedy this problem, some real data with good classified labels, e.g., mushroom data, congressional votes data, soybean disease [2] and Reuters news collection [1], were taken as the experimental data for categorical clustering algorithms [13][20][29][30]. In view of the feature of market-basket data, we propose in this paper a validation model based on Information Gain (IG) to assess the qualities of the clustering results. It is noted that information gain is widely used in the classification problem [23][24]. Explicitly, ID3 [23] and C4.5 [24] used information gain measurement to select the test attribute with the highest information gain for splitting when constructing the decision tree.

The definitions required for deriving the information gain of a clustering result are given below.

Definition 4: The entropy of an attribute J_a in the database D is defined as:

$$I(J_a, D) = -\sum_{i=1}^{n} \frac{|J_a^i|}{|D|} * log_2 \frac{|J_a^i|}{|D|}.$$

where $|D|$ is the number of transactions in the database D and $|J_a^i|$ denotes the number of the transactions whose attribute J_a is classified as the value J_a^i in the database D.

Definition 5: The entropy of an attribute J_a in a cluster C_j is defined as:

$$I(J_a, C_j) = -\sum_{i=1}^{n} \frac{|J_{a,c_j}^i|}{|C_j|} * log_2 \frac{|J_{a,c_j}^i|}{|C_j|}.$$

where $|C_j|$ is the number of transactions in cluster C_j, and $|J_{a,c_j}^i|$ denotes the number of the transactions whose attribute J_a is classified as the value J_a^i in C_j.

Definition 6: Let a clustering U contain C_1, C_2, ..., C_m clusters. Thus, the entropy of an attribute J_a in the clustering U is defined as:

$$E(J_a, U) = \sum_{C_j \in U} \frac{|C_j|}{|D|} I(J_a, C_j).$$

Definition 7: The information gain obtained by separating J_a into the clusters of the clustering U is defined as:

$$Gain(J_a, U) = I(J_a, D) - E(J_a, U).$$

Definition 8: The information gain of the clustering U is defined as:

$$IG(U) = \sum_{J_a \in I} Gain(J_a, U).$$

where I is the data set of the total items purchased in the whole market-basket data records.

For clustering market-basket data, the larger an IG value, the better the clustering quality is. In market-basket data, with the taxonomy tree structure, there are three kinds of IG values, i.e., $IG_{item}(U)$, $IG_{cat}(U)$, and $IG_{total}(U)$, for representing the quality of a clustering result. Specifically, $IG_{item}(U)$ is the information gain obtained on items and $IG_{cat}(U)$ is the information gain obtained on categories. $IG_{total}(U)$ is the total information gain, i.e., $IG_{total}(U) = IG_{item}(U) + IG_{cat}(U)$. In general, market-basket data set is typically represented by a 2-dimensional table, in which each entry is either 1 or 0 to denote purchased or non-purchased items, respectively. In IG validation model, we treat each item in market-basket data as an attribute J_a with two classified label, 1 or 0. It will be shown in Section 4 that with the category-based adherence measurement, algorithm CBA devised outperforms others in the clustering quality based on the IG validation model.

3 Design of Algorithm CBA

The similarity measurement of CBA, called category-based adherence, will be described in Section 3.1. The procedure of CBA is devised in Section 3.2 and the complexity of CBA is analyzed in Section 3.4.

3.1 Similarity Measurement: Category-Based Adherence

The similarity measurement employed by algorithm CBA, called category-based adherence, is defined as follows. In the taxonomy tree, the *nearest large node* of an item i_k is itself if i_k is large and is its nearest large ancestor node otherwise. Then, the distance of an item to a cluster is defined below.

Definition 9: (**Distance of an item to a cluster**): For an item i_k of a transaction, the *distance* of i_k to a given cluster C_j, denoted by $d(i_k, C_j)$, is defined as the number of links between i_k and the nearest large nodes of i_k. If i_k is a large node in cluster C_j, then $d(i_k, C_j) = 0$. Otherwise, the nearest large node is the category node which is the lowest generalized concept level node among all large ancestors of item i_k. Note that if an item or category node is identified as large node, all its high level category nodes will also be large nodes.

Definition 10: (**Adherence of a transaction to a cluster**):

For a transaction $t = \{i_1, i_2, ..., i_p\}$, the adherence of t to a given cluster C_j, denoted by $H(t, C_j)$, is defined as the average distance of the items in t to C_j and shown below.

$$H(t, C_j) = \frac{1}{p}\sum_{k=1}^{p} d(i_k, C_j).$$

where $d(i_k, C_j)$ is the distance of i_k in cluster C_j.

3.2 Procedure of Algorithm CBA

The overall procedure of algorithm CBA is outlined as follows.

Procedure of Algorithm CBA

Step 1. Randomly select k transactions as the seed transactions of the k clusters from the database D.

Step 2. Read each transaction sequentially and allocates it to the cluster with the minimum category-based adherence. For each moved transaction, the counts of items and their ancestors are increased by one.

Step 3. Repeat Step 2 until no transaction is moved between clusters.

Step 4. Output the taxonomy tree for each cluster as the visual representation of the clustering result.

In Step 1, algorithm CBA randomly selects k transactions as the seed transactions of the k clusters from the database D. For each cluster, the items of the seed transaction are counted once in the taxonomy tree. In each cluster, the items and their ancestors are all large in the very beginning because their count is one (which means 100% in the only seed transaction), larger than the minimum support threshold. For each cluster, these large nodes represent the hot sale topics in this cluster. In Step 2, algorithm CBA reads each transaction sequentially and allocates it to the cluster with the minimum category-based adherence. After one transaction is inserted into a cluster C_j, the counts of the items and their ancestors are increased by one in the corresponding nodes in the taxonomy tree of C_j. In addition, the minimum support count of C_j is updated. In Step 3, algorithm CBA repeats Step 2 until no transaction is moved between clusters. In Step 4, algorithm CBA outputs the taxonomy tree of the final clustering result for each cluster, where the items, categories, and their corresponding counts are presented.

3.3 Complexity Analysis of Algorithm CBA

The time complexity and the space complexity of algorithm CBA are analyzed by the following two theorems.

Theorem 1: The time complexity of CBA is $O(lk|D|vn)$, where l is the number of iterations of Step2, k is the given cluster number, $|D|$ is the database size, v is the average transaction length, and n is the maximum number of levels in the taxonomy tree. **Proof:** We first define following symbols to analyze the complexity of Step 2 in detail: I_t is the item set in transaction t, $1 \leq t \leq |D|$, I_t^m is the mth item in transaction t, $cost(I_t^m, C_k)$ is the cost for I_t^m to find the nearest large node in cluster C_k, and $x(I_t^m)$ is the number of levels from item I_t^m to its highest ancestor, $1 \leq x(I_t^m) \leq n$. Hence, for each transaction t, the adherence of t to every cluster is obtained in Step 2. Thus, the time complexity of this sub-step is

$\sum_{k} \sum_{I_t \in D} \sum_{I_t^m} cost(I_t^m, C_k)$. After obtaining the cluster C_a in which t has the minimum adherence, t is allocated to C_a and the count of items and related categories in the taxonomy tree of C_a will be increased by one. Thus, the time complexity of this sub-step is $\sum_{I_t^m} x(I_t^m)$. With l iterations for running Step 2, the total time complexity is therefore

$$\sum_{l}[\sum_{k} \sum_{I_t \in D} \sum_{I_t^m} cost(I_t^m, C_k) + \sum_{I_t^m} x(I_t^m)]$$
$$\leq lk|D|vn + l|D|vn = l(k+1)|D|vn$$
$= O(lk|D|vn)$ where l is the number of iterations of Step2, k is the given cluster number, $|D|$ is the database size, v is the average transaction length, and n is the maximum number of levels in the taxonomy tree. **Q.E.D.**

Theorem 2: The space complexity of CBA is $O(|D| + kA)$, where $|D|$ is the database size, k is the given cluster number, and A is the all nodes, including category nodes and item nodes, in the taxonomy tree. **Proof:** First, before CBA is executed, all data must be loaded and the space requirement is $O(|D|)$. In each cluster, there is only a array structure needed to store the counts of all nodes, whose space requirement is $O(A)$. Because the number of clusters is k, the space requirement is $O(kA)$ for all clusters. Thus, the overall space complexity of CBA is $O(|D| + kA)$. **Q.E.D.**

4 Experimental Results

To assess the efficiency of CBA, we conducted experiments to compare CBA with a traditional hierarchical clustering algorithm, called *CL* (standing for *Complete Link*) [17] and another algorithm proposed in [29] (for the convenience, the algorithm is named as *Basic* in this paper). In general, CL performs well when the data is well-separated. However, because market-basket data is very sparse, CL will suffer from the existence of outliers. We therefore adopt the outlier eliminating method in [12] in our implementation of CL algorithm. In addition, algorithm Basic is a K-Mean-based process which utilized a cost function to minimize the overlap of large items (corresponding to inter-cluster cost) and minimize the union summation of small items (corresponding to intra-cluster cost). By extending both previous approaches with taxonomy consideration in market-basket data, we also implement algorithm *CLT* (standing for *Completed Link with Taxonomy*) and algorithm *BasicT* (standing for *Basic with Taxonomy*) for comparison purposes. Note that the algorithms mentioned above will generate a fixed number of clusters for fair comparisons. However, algorithm ROCK [13] has very high time complexity and generates the uncertain number of clusters. In our experiments for clustering market-basket data, ROCK generates too many clusters (in the range between 500 to 600 clusters in our experiments of $|D| = 10K$ transactions) and most of them have few transactions. These small-size clusters have little implication for marketing strategies. We thus remove ROCK from our following comparisons. The details of data generation are described in Section 4.1. The experimental results are shown in Section 4.2.

Notation	Meaning
$\|D\|$	The database size
$\|T\|$	Average size of the transactions
$\|L\|$	Number of large itemsets within database
N^I	Number of items in database
N^R	Number of the roots
N^L	Number of the taxonomy levels

Table 1. The meanings of various parameters used in experimental results.

4.1 Data Generation

The meanings of various parameters used in our experiments are shown in Table 1. We take the real market-basket data from a large bookstore company for performance study. In this real data set, there are $|D| = 100K$ transactions and $N^I = 21807$ items. Note that in this real data, there are volume of transactions containing only single items, and many items are purchased infrequently. In addition, the number of the taxonomy level in this real data set is 3. In addition, to provide more insight into this study, we use a well-known market-basket synthetic data generated by the IBM Quest Synthetic Data Generation Code [5], as the synthetic data for performance evaluation. This code will generate volumes of transaction data over a large range of data characteristics. These transactions mimic the transactions in the real world retailing environment. This generation code also assumes that people will tend to buy sets of items together, and each such set is potentially a maximal large itemset. An example of such a set might be sheets, pillow case, comforter, and ruffles. However, not all items purchased by the customer are large itemsets. The average size of the transactions, denoted by $|T|$, is set to 5 as default. The average size of the maximal potentially large itemsets, denoted by $|I|$, is set to 2 as default. The number of maximal potential large itemsets, denoted by $|L|$, is set to 2000. The number of items in database, denoted by N^I, is set to 5000 as default. The number of roots, denoted by N^R, is set to 100 and the number of the taxonomy level, denoted by N^L, is set to 3.

4.2 Performance Study

We conduct four experiments in this section for performance study and the clustering quality is evaluated by the IG values. For algorithms CBA, Basic, and BasicT, the minimum support percentage S_p is set to 0.5%. Recall that there are three kinds of IG values, i.e., IG_{item}, IG_{cat}, and IG_{total}, for evaluating the quality of the clustering result. IG_{item} is the information gain obtained on items and IG_{cat} is the information gain obtained on categories. $IG_{total} = IG_{item} + IG_{cat}$.

4.2.1 Experiment One: Comparison on the clustering results in real data

Figure 2 shows the relative quality of clustering results of CBA, CL, CLT, Basic, and BasicT in real data set where $|D| = 100K$, $N^L = 3$, and $N^I = 21807$. With category-based adherence measurement, CBA emerges as the winner among all algorithms evaluated. Note that as described in [6] a term with a higher discrimination value will be associated with a longer distance between data points in the database. Because different items may belong to the same categories, the discrimination values of categories are lower than those of items for the transactions in the database. Note that CL can obtain better clustering results when the average discrimination values of the terms, including the items in the transactions, increase. With the taxonomy information, CLT obtains less information gain than CL because CLT obtains the lower average discrimination values of the terms, including both the items and the categories in the transactions. However, for identifying the large and small terms in Basic and BasicT, the discrimination values of the items and the categories are aggregated in the similarity measurements for clustering market-basket data. Thus, BasicT can obtain higher IG values than Basic. By considering the item similarities across their category levels, algorithm CBA allocates each transaction to a proper cluster so that CBA in general outperforms other algorithms in the three IG values. This reasoning accounts for the results shown in Figure 2. In this real data set, because the number of taxonomy levels is 3 and the number of item nodes is larger than that of category nodes, CBA obtains more information gain on items than categories, e.g., $IG_{item} > IG_{cat}$. However, it will be shown in the third experiment that when the number of taxonomy levels increases, CBA is able to obtain more information gain on categories than on items.

4.2.2 Experiment Two: When the database size $|D|$ varies

In this experiment, the scalability of CBA is evaluated by both the real data and the synthetic data. Note that the time complexity of CBA is $O(lk|D|vn)$, where l is the number of iterations of Step 2, k is the given cluster number, $|D|$ is the database size, v is the average transaction length, and n is the maximum number of levels in the taxonomy tree. By varying the real database size $|D|$ from $5K$ to $20K$, it is shown in Figure 3 that CBA significantly outperforms other algorithms in execution efficiency. Note that the logarithmic scale with base 10 is used in the y-axis of Figure 3 since the execution time of CBA is significantly shorter than those of other algorithms and the execution times of CBA increase linearly as the database size increases, indicating the good scale-up feature of algorithm CBA.

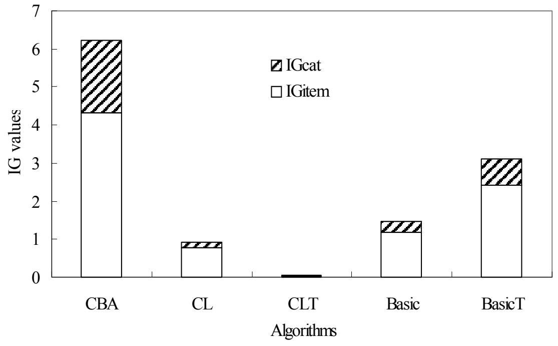


Figure 2. The comparison of IG values on the real database.

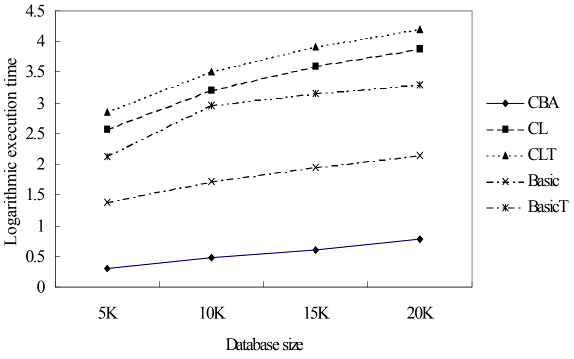
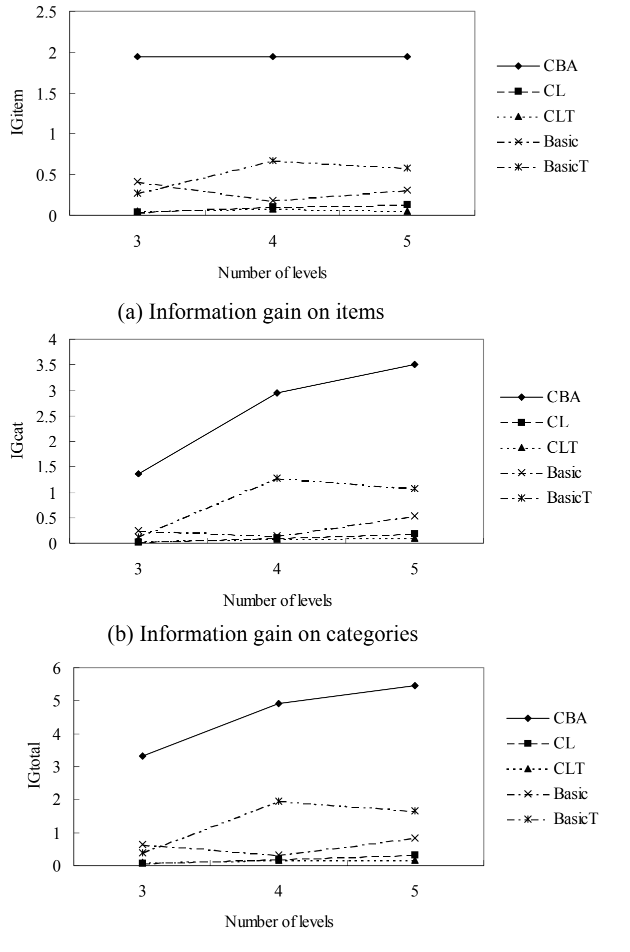


Figure 3. Execution time in logarithm for CBA, CL, CLT, Basic, and BasicT when the database size $|D|$ varies.

4.2.3 Experiment Three: When the number of taxonomy levels N^L varies in synthetic data

In the synthetic data experiment shown in Figure 4, we set $|D| = 100K$, $|T| = 5$, $|I| = 2$, $|L| = 2000$, $N^I = 5000$,



(a) Information gain on items

(b) Information gain on categories

(c) Information gain in total

Figure 4. The IG values when the number of taxonomy levels N^L varies.

$N^R = 100$, and N^L varies from 3 to 5. When the number of taxonomy levels increases, the number of internal (i.e., category) nodes also increases. Thus, the IG_{cat} increases so that CBA can obtain more information gain on categories than on items, indicating the advantage of CBA by employing the category-based adherence as the measurement.

4.2.4 Experiment Four: When the number of maximal potential large itemsets $|I|$ varies in synthetic data

In the synthetic data experiment shown in Figure 5, we set $|D| = 100K$, $|T| = 5$, $|L| = 2000$, $N^I = 5000$, $N^R = 100$, $N^L = 3$, and $|I|$ varies from 1 to 4. Note that the category-based adherence of a transaction to a cluster is the average distance of the items in this transaction to that cluster. In algorithm CBA, when $|I|$ increases, the IG_{total} value also increases. This can be explained by the reason that when $|I|$ increases, the number of transactions containing co-occurrence itemsets increases and thus these transactions are allocated to the corresponding clusters with lower category-based adherences. Explicitly, many members of the transactions containing an item i_k are allocated to a cluster C_j because these transactions also contain other frequently-purchased items which are purchased together with i_k. When $|I|$ increases, the number of such items as i_k also increases so that more transactions containing i_k are allocated to one cluster instead of being allocated to several clusters separately. Therefore, the IG_{total} value increases.

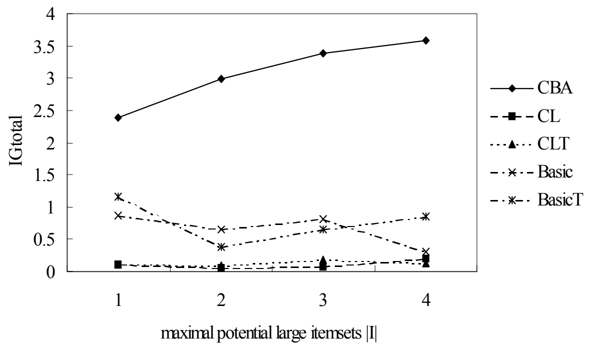


Figure 5. The IG_{total} value when the maximal potential large itemset $|I|$ varies.

5 Conclusion

In this paper, we proposed an efficient algorithm for clustering market-basket data. In view of the features of market-basket data, we devised in this paper a novel measurement, called the category-based adherence, and utilize this measurement to perform the clustering. With this category-based adherence measurement, we developed an efficient clustering algorithm, called algorithm CBA (standing for Category-Based Adherence), for market-basket data with the objective to minimize the category-based adherence. A validation model based on Information Gain (IG) was also devised in this paper to assess the quality of clustering for market-basket data. As validated by both real and synthetic datasets, it was shown by our experimental results, with the taxonomy information, algorithm CBA devised in this paper significantly outperforms the prior works in both the execution efficiency and the clustering quality for market-basket data.

References

[1] Reuters-21578 news collection, http://www.research.att.com/ lewis/reuters21578.html.

[2] UCI Machine Learning Repository. *http:://www.ics.uci.edu/~mlearn/MLRepository.html.*

[3] C. C. Aggarwal, C. M. Procopiuc, J. L. Wolf, P. S. Yu, and J.-S. Park. Fast Algorithms for Projected Clustering. *ACM SIGMOD International Conference on Management of Data*, pages 61–72, June 1999.

[4] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *ACM SIGMOD International Conference on Management of Data*, 27(2):94–105, June 1998.

[5] 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.

[6] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. *Addison-Wesley*, 1999.

[7] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental Clustering and Dynamic Information Retrieval. *Proceedings of the 29th ACM Symposium on Theory of Computing*, 1997.

[8] M.-S. Chen, J. Han, and P. S. Yu. Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):866–833, 1996.

[9] R. Duda and P. Hart. Pattern Classification and Scene Analysis. *Wiley, New York*, 1973.

[10] C. Fellbaum. WordNet: An Electronic Lexical Database. *The MIT Press*, 1998.

[11] D. H. Fisher. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning*, 1987.

[12] S. Guha, R. Rastogi, and K. Shim. CURE: An Efficient Clustering Algorithm for Large Databases. *ACM SIGMOD International Conference on Management of Data*, 27(2):73–84, June 1998.

[13] 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.

[14] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 2001.

[15] J. Han and Y. Fu. Discovery of Multiple-Level Association Rules from Large Databases. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 420–431, September 1995.

[16] J. Han and M. Kamber. Data Mining: Concepts and Techniques. *Morgan Kaufmann*, 2000.

[17] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. *Prentice Hall*, 1988.

[18] A. K. Jain, R. P. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 4–37, Jan. 2000.

[19] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Computer Surveys*, 31(3), Sept. 1999.

[20] F.-X. Jollois and M. Nadif. Clustering Large Categorical Data. *PAKDD02*, 2002.

[21] R. T. Ng and J. Han. Efficient and Effective Clustering Methods for Spatial Data Mining. *Proceedings of 20th Annual International Conference on Very Large Data Bases*, pages 144–155, 1994.

[22] C. Ordonez and E. Omiecinski. A Fast Algorithm to Cluster High Dimensional Basket Data. *Proceedings of the 1st IEEE International Conference on Data Mining (ICDM 2001)*, Nov./Dec. 2001.

[23] J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1986.

[24] J. R. Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufmann*, 1993.

[25] S. Scott and S. Matwin. Text Classification Using WordNet Hypernyms. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 38–44, 1998.

[26] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. *Proceedings of 24th Annual International Conference on Very Large Data Bases*, pages 428–439, 1998.

[27] R. Srikant and R. Agrawal. Mining Generalized Association Rules. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 407–419, September 1995.

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

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

[30] Y. Xiao and M. H. Dunham. Interactive Clustering for Transaction Data. *Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2001)*, Sept. 2001.

[31] C.-H. Yun, K.-T. Chuang, and M.-S. Chen. An Efficient Clustering Algorithm for Market Basket Data Based on Small-Large Ratios. *Proceedings of the 25th International Computer Software and Applications Conference (COMPSAC 2001)*, October 2001.

[32] O. R. Z., A. Foss, C.-H. Lee, and W. Wang. On Data Clustering Analysis: Scalability, Constraints and Validation. *PAKDD02*, 2002.

[33] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. *ACM SIGMOD International Conference on Management of Data*, 25(2):103–114, June 1996.