Similarity of Transactions for Customer Segmentation

Ke Lu, Tetsuya Furukawa

Similarity of Transactions for Customer Segmentation

Ke Lu and Tetsuya Furukawa

Department of Economic Engineering, Kyushu University,
Hakozaki 6–19–1, Higashi-ku, Fukuoka 812–8581 Japan
{looker,furukawa}@en.kyushu-u.ac.jp

Abstract. Customer segmentation is usually the first step towards customer analysis and helps to make strategic plans for a company. Similarity between customers plays a key role in customer segmentation, and is usually evaluated by distance measures. While various distance measures have been proposed in data mining literature, the desirable distance measures for various data sources and given application domains are rarely known. One of the reasons lies in that semantic meaning of similarity and distance measures is usually ignored. This paper discusses several issues related to evaluating customer similarity based on their transaction data. Various set distance measures for customer segmentation are analyzed in several imaginary scenarios, and it is shown that each measure has different characteristics which make the measure useful for some application domains but not for others. We argue that no measure always performs better than other measures, and suitable measures should be adopted for specific purposes depending on applications.

Keywords: Customer Segmentation, Transaction Similarity, Set Distance

1 Introduction

Intense commercial competition induces companies to pay increasing attention to understand their customers more deeply in order to support decision making. For example, e-commerce companies usually offer distinct home pages and recommend relative products to customers based on predictive models built on customer data. Most financial companies construct their own risk models based on the analysis of customer data to prevent customer credit risk.

Data mining techniques have been widely applied in customer relationship management (CRM) [10]. As an important topic in CRM, customer segmentation, which is based on analysis of customer similarity, has drawn increasing attention, and the similarity between customers is an unavoidable issue. While customer segmentation is highly expected to help companies make commercial plans, it does not seem that existing analysis methods work well enough. Consumers still receive significant amount of mails recommending products that they are not interested in, and online recommendations are still far from acceptable

[14]. An important reason is that customers are segmented improperly due to unsuitable similarity measures. In order to make customer segmentation more adaptable and flexible, it is necessary for companies to understand their target of customer segmentation and which similarity measure is needed for a specific application.

Clustering is usually employed to segment customers, and it is critical to find suitable distance measures to evaluate the similarity between customers with various types of data sources. Customer data is the corner stone of customer segmentation, and can be briefly separated into two categories. The one is demographic data, which is relatively static in long term. Demographic data may include customers' natural properties, *e.g.*, age and gender, or social properties, *e.g.*, marital status and income. The other one is transaction data, which is relatively dynamic compared with demographic data. Generally, transaction data may include much information in purchasing action. Other types of data, *e.g.*, lifestyle data, psychographic data and marketing action data, can be derived from demographic data and transaction data through some statistic methods. Transaction data is merely available for current customers, so that it is necessary to utilize demographic characteristics that are observable in advance for targeting potential customers who are similar to current customers. It has been found that transaction data is the most powerful and reliable data for predicting future customer purchase behavior [8][15]. This paper focuses on the issues of segmenting customers base on transaction data and the issues related to demographic data are not included.

Considerable efforts in finding appropriate distance measures for transaction data have been conducted throughout different applications, because distance measures are fundamentally important for clustering data. However, such endeavors pay little attention to the problem: for a similarity that is evaluated from certain perspective, which distance measures are desirable. For example, some similarities are desired to be evaluated by proportion of affinity items to transactions, while other may require a specific distance. Without explicit understanding the meaning of similarity between customers, it is difficult to select the adaptable distance measures against diverse types of customer data and applications. This paper presents formal discussion on several possible perspectives of measuring the similarity between transactions. Set distances are introduced for evaluating the similarity between transactions. Some measures partially focus on pairwise item distance, while others are affected by assignment of items greatly. It is argued that for different applications, different measures should be adopted and various segmentation results may come out. To the best of our knowledge, this is the first paper that introduces set distances to evaluate the similarity between transactions.

The rest of this paper is organized as follows. Section 2 discusses some preliminary problems and gives some description about the data mentioned in this paper. Similarity between transactions based on *Affinity Items* is discussed in Section 3. Section 4 concerns about the application of set distance measures partially focusing on pairwise item distance. Section 5 refers to the discussion

of set distance measures that take assignment of items into consideration. The conclusion of this paper is presented in Section 6.

2 Preliminaries

Customer data is the first word to segment customer. The description of transaction data mentioned in this paper is formally given as follows. Let a customer transaction database D contain all of transactional records of customers. Let $I = \{i_1, i_2, ..., i_r\}$ be the set of product items included in D, where i_k $(1 \leqslant k \leqslant r)$ is the identifier for the k_{th} item. For items i_1 and i_2, let the distance be denoted by $d(i_1, i_2)$. This paper assumes that the distance between pairwise items is given in advance [1][3]. A transaction, denoted by T, is a subset of I. For a distance measure and a threshold σ, if the distance between two transactions is shorter than σ, they are said similar to each other. Customers can be segmented by analyzing the similarity between their transactions.

Segmenting customers based on transaction data has been a long overdue issue for a public debate. Motivated by [9], the so called Customer-Oriented Catalog Segmentation problem, which concerns the problem of segmenting customer based on transactions, has been discussed in [2][6]. The issues related to segmenting customers by transaction data with concept hierarchy have been addressed in [7][12]. As an important association study, clustering transactions has drawn increasing attention [14][16].

The literature mentioned above measures the similarity between transactions based on co-occurrence items. Intuitively, two transactions are deemed to be similar if most items in one transaction have the same item in the other transaction. Hence, counting the co-occurrence items of two transactions is a general method to evaluate the similarity between two transactions, and follows the conventional understanding of similarity. However, it may face a predicament of differentiation dilemma and overlook the relationship between individual items. Nowadays, companies differentiate their products to tackle the problem of homogenization, so that the total kinds of items in transactions are doubled in the past decadeswhile different items may denote very similar products or highly related products. Therefore, what is needed is the error-tolerant measure, *s.t.*, if two items are similar to some predefined extent, they can be regarded as equal to each other in certain sense.

Based on pairwise distance between items, set distance measures are introduced to evaluate the similarity between transaction. Some topics related to set distance measure have been deeply discussed in [5][13].

From the mathematical point of view, distance is defined as a quantitative degree of how far two entities are from each other. The concept of distance mentioned in this paper, both pairwise item distance and transaction distance, obeys the following mathematical meaning of distance.

Definition 1. *Given a set S, a real-valued function $d(x, y)$ on the Cartesian product $S \times S$ is a distance if for any $x, y \in S$, it satisfies the following conditions:*

1. $d(x, y) \geq 0$ (non-negativity),
2. $d(x, y) = d(y, x)$ (symmetry),
3. $d(x, y) = 0$ *if and only if* $x = y$ (self-identity).

A great number of distance measures have been proposed for various applications [4], and the selection of distance measures should depend on specific data and applications. Employing set distance to evaluate the similarity between transactions can give us more precise information than the similarity given by co-occurrence based methods.

3 Transaction Similarity based on Affinity of Items

The degree of how similar two items are can be evaluated simply by the distance between them. When the distance is short within a certain range, they are called *Affinity Items* in this paper. The definition of *Affinity Items* is formally given as follows.

Definition 2. *If* $d(i_1, i_2) \leqslant \sigma$*, where* σ *is a threshold given in advance,* i_1 *and* i_2 *are regarded as Affinity Items to each other, denoted by* $Aff(i_1, i_2)$*.*

Compared with co-occurrence, employing distance measures enriches the meaning of similarity between transactions. For example, both substitutes and complements can be deemed similar to each other, or highly related to each other in other words, *e.g.*, both the distance between $Coke$ and $Pepsi$ (as substitutes) and the distance between *computer* and $software$ (as complements) can be deemed very short. Transaction T_1 can be treated similar to transaction T_2 that consists of substitutes or another transaction T_3 that consists of complements. The details of this problem are not considered in this paper, because this paper assumes that the pairwise distance is given in advance.

Example 1. Tom and Jerry meet at a super market and found that they bought $Coke$ and $Pepsi$, respectively. Even though $Coke$ and $Pepsi$ are different product items, Tom and Jerry may improve the identity between them mentally, because $Coke$ and $Pepsi$ are both soft drink.

The above scenario may appear in our daily life, and this mental phenomenon promotes the following definition about similarity.

Definition 3. *Transactions* T_1 *and* T_2 *are loosely similar to each other if* $\exists i_1 \in \mathrm{T}_1$*,* $\exists i_2 \in \mathrm{T}_2$*, s.t.,* $Aff(i_1, i_2)$*.*

Similar product items, even the same item, appear in different transactions occasionally is a general phenomenon. However, how two customers are said to be similar to each other, is from the perspective of their purchase behavior, which indicates that all of the items in a transaction should be taken into consideration. Similar customers may used to buy some products together. Motivated by this analysis, the definition that two transactions are similar to each other in the strictest term is given as follows.

Definition 4. *Transactions* T_1 *and* T_2 *are strictly similar to each other if* $\forall i_1 \in \mathrm{T}_1$, $\exists i_2 \in \mathrm{T}_2$, *s.t.,* $Aff(i_1, i_2)$, *and* $\forall i_2 \in \mathrm{T}_2$, $\exists i_1 \in \mathrm{T}_1$, *s.t.,* $Aff(i_1, i_2)$.

Example 2. Consider the example with two transactions T_1 and T_2 shown in Figure 1 whose items are denoted by $\times$ and $\bigcirc$, respectively. If two items are circled by dotted line together, they are *Affinity Items* for each other. In (a), even there is only a pair of item are *Affinity Items*, they can be deemed loosely similar to each other. While in (b), every item in T_1 has at least one *Affinity Item* in T_2, and vice versa, so that they are strictly similar to each other.

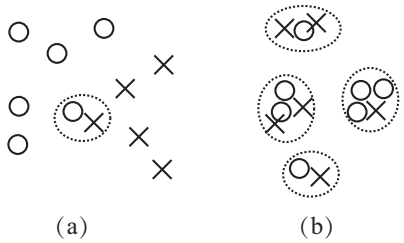

Fig. 1. Loose Similarity and Strict Similarity.

In practical applications, transactions that satisfy strict similarity come out from time to time, and loose similarity is not so acceptable well enough. In most cases, companies concern about the fraction of items that have *Affinity Items* in another transaction, the following measure is an acceptable one.

Definition 5. *Let* $|R(T_1, T_2)|$ *denote the total number of items in transaction* T_1 *that have* Affinity Items *in transaction* T_2. *The Cardinal Transaction Similarity between transactions* T_1 *and* T_2, *denoted by* $S_c(T_1, T_2)$, *is the fraction of items in either transaction that has* Affinity Items *in the other transaction, i.e.,* $S_c(T_1, T_2) = \frac{|R(T_1,T_2)|+|R(T_2,T_1)|}{|T_1|+|T_2|}$. *For a specified similarity threshold* δ, T_1 *and* T_2 *are cardinally similar to each other, if* $S_c(T_1, T_2) \geqslant \delta$.

Clearly, alternative definitions of measures are possible and the above measure, though intuitive, is only one among possibly several reasonable similarity definitions between sets of itemsets. Cardinal Transaction Similarity is a simple and straightforward measure. However, it may lose some other important information about the similarity between transactions.

4 Transaction Similarity based on Distance of Items

While Cardinal Transaction Similarity can approximately reflect how similar two transactions are, the specific distance between items is not involved in evaluating the similarity. The measures, which just vaguely evaluate similarity between

transactions, can not satisfy diverse practical application in daily scenarios of companies.

Example 3. Let $\times$ and $\bigcirc$ denote the items of T_1 and T_2, respectively. In the example shown in Figure 2, items are *Affinity Items* for each other if they are circled by dotted lines. Cardinal Similarities between transactions T_1 and T_2 are all 0.67 in (a), (b), (c) and (d). However, there are also some differences between them should not be ignored. In (a), there is only one item of T_1 that does not have *Affinity Items* in T_2, and in (b), there are nearly half of items in T_1 that does not have *Affinity Items*. In both (a) and (c), for every item, there is at most one *Affinity Items*, while in both (b) and (d), an item may have several *Affinity Items*. Even for (a) and (c), there is an obvious difference. In (a), all of items in either T_1 or T_2 are relatively similar to each other, while in (c), there is an item in T_2 is far from the rest items in T_1 or T_2.

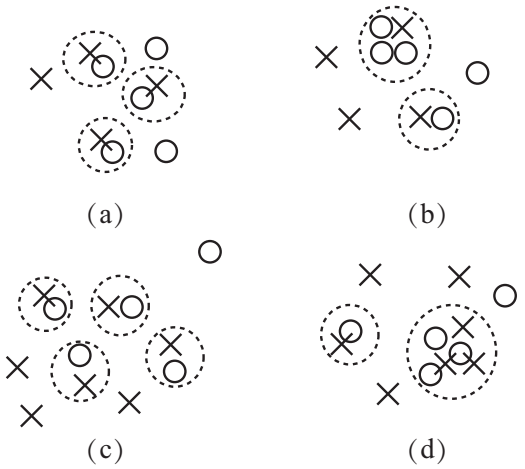

Fig. 2. Examples of Different Conditions for the Same Cardinal Similarity.

As shown in above example, Cardinal Transaction Similarity is not well suitable for applications that require precise information about similarity between transactions. It is very desirable to find measures that at least take the following two factors into consideration:

1. The pairwise distance between items.
2. The assignment determining pairs of items that are involved in calculating the distance between transactions.

This section introduces set distance measures, which determine distance from the perspectives corresponding to those two factors mentioned above, to measure

the similarity between transactions. The following discussion starts by introducing the general form of set distance measures. Referencing some concepts of bipartite graph, if two items are assigned together and the distance between them is involved in calculating the distance between two transactions, it is said that they are connected and there is an edge between these two items. For two transactions T_1 and T_2, the general form of set distance measures between them can be written in the following way

$$\mathcal{D}_s(T_1, T_2) = F(\frac{\sum_{(i_1,i_2)\in M} d(i_1, i_2)}{|M|}),$$

where $M \subseteq T_1 \times T_2$ defines an assignment between T_1 and T_2, $|M|$ denotes the number of edges in M, and F is an aggregation function against the normalized sum of pairwise distance. Maximum, minimum and average are the general options for F. By combining different assignments and aggregation functions, we can get various set distance measures referring to divers factor options.

Assignment is not necessary to be considered together with pairwise distance. Companies may prefer pairwise distance while ignore the assignment in some cases. Consider the following scenario.

Scenario 1:
For a given transaction T_1, companies hope to find a transaction T_2 in which there is an item that has the shortest distance to an item in T_1 than any item in other transactions, and deem T_1 and T_2 similar to each other. Corresponding to this scenario, Single-link Distance [11] introduced as follows is a good choice.

Single-link Distance:

$$D_{sl}(T_1, T_2) = \min_{i_1\in T_1, i_2\in T_2} d(i_1, i_2)$$

In Scenario 1, companies only require one pair of items to satisfy a given constraint. On the opposite, the following scenario need that every items satisfy some conditions.

Scenario 2:
For a given transaction T_1, companies hope to find a transaction T_2, *s.t.*, distances of every pairwise items between T_1 and T_2 shorter than a given threshold, and deem T_1 and T_2 similar to each other. Corresponding to this scenario, Complete-link Distance [11] introduced as follows is suitable.

Complete-link Distance:

$$D_{cl}(T_1, T_2) = \max_{i_1\in T_1, i_2\in T_2} d(i_1, i_2)$$

Another well known distance measure is Hausdorff Distance.

Hausdorff Distance:

$$D_h(T_1, T_2) = \max(h(T_1, T_2), h(T_2, T_1)),$$

where $h(T_1, T_2)$, the so-called one-sided Hausdorff distance from T_1 to T_2, is formally defined as follows.

$$h(T_1, T_2) = \max_{i_1\in T_1} (\min_{i_2\in T_2} d(i_1, i_2))$$

$h(T_1, T_2)$ and $h(T_2, T_1)$ are asymmetric.

If two transactions are deemed similar based on Complete-link Distance, the distances of every pair of items are within a constrained range. Compared with Complete-link Distance, Hausdorff Distance only guarantees every item in one transaction can find an item in the other transaction, *s.t.*, the distance between two items is in a limited range. However, they may cause another type of ambiguous result. Two transactions may be deemed dissimilar to each other due to a pair of items that are distant to each other, while the rest of pairwise distances are very short. If a new constraint is added to calculate Complete-link distance and Hausdorff distance, which requires every pair of items involved in calculating those two distances must be *Affinity Items*, it seems more reasonable. New measures are formally given as follows.

Affinity Complete-link Distance:

$$D_{acl}(T_1, T_2) = \max_{i_1 \in T_1, i_2 \in T_2, Aff(i_1, i_2)} d(i_1, i_2)$$

Affinity Hausdorff Distance:

$$D_{ah}(T_1, T_2) = \max(h(T_1, T_2), h(T_2, T_1)),$$

where $h(T_1, T_2)$, the so-called one-sided Affinity Hausdorff distance from T_1 to T_2, is formally defined as follows.

$$h(T_1, T_2) = \max_{i_1 \in T_1} (\min_{i_2 \in T_2, Aff(i_1, i_2)} d(i_1, i_2))$$

$h(T_1, T_2)$ and $h(T_2, T_1)$ are asymmetric.

These five measures do not take much information about the items into consideration, and are determined by the distance of certain pair of items with extreme condition. For example, two transactions may be deemed similar if there is a pair of items that are very similar while the rest of pairwise distances are far apart from each other. Different from those five measures that are determined by certain pairwise items, Average Distance takes the distances of pairwise items into consideration. For two transactions, the upper limit and lower limit of their Average Distance are Complete-link Distance and Single-link Distance, respectively.

Average Distance:

$$D_{avg}(T_1, T_2) = \frac{\sum_{i_1 \in T_1, i_2 \in T_2} d(i_1, i_2)}{|T_1||T_2|}$$

Despite various weaknesses, the set distance measures mentioned in this section are very straightforward, and adhere to conventional thinking way of clustering.

5 Assignment of Items between Transactions

It is obvious that there is not fixed structure for transaction, *e.g.*, the size of different transactions may be different, and items are not necessary to be corre-

sponding to any attribute. This characteristic induces assignment to be a noteworthy factor in calculating set distance between transactions. Actually, assignment refers to structures of transactions, *e.g.*, a transaction may mainly consist of soft drinks and alcoholic drinks, while another transaction may conclude pastry and vegetable. The original intention of assignment is to connect items that are as similar as possible.

Example 4. As shown in Figure 3, let □, ◯ and × indicate the items of three different transactions, respectively. If the assignment of items is not taken into consideration, for transactions denoted by □ and ×, Single-link Distance, Complete-link Distance and Hausdorff Distance between transaction denoted by ◯ and them are the same as shown in Figure 3(a). However, if assignment of items requires that every item must be connected to at least one item in another transaction, the transaction denoted by □ is closer to the one denoted by ◯ than the one denoted by × as shown in Figure 3(b).

This is just a simple example that every transaction has equal size. It becomes more complicated when the size of various transactions are different.

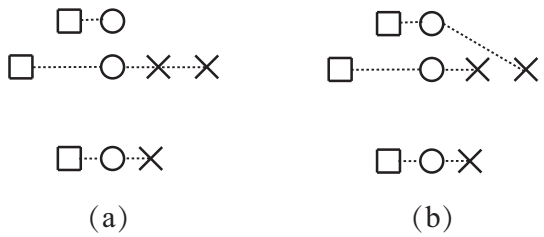

Fig. 3. Examples of Assignment in Set Distance.

This section discusses set distance measures that take assignment of items into consideration. The following discussion goes with various scenarios which companies may encounter.

Scenario 3:
Companies have a target transaction T_1 which consists of some picked up products. They hope to find another transaction T_2 in which there is at least one distinct similar item for as many as possible items in T_2, and the average pairwise distance is as short as possible. Corresponding to this scenario, Matching Distance introduced as follows is a good choice.

Matching Distance:
For two transactions T_1 and T_2, if every item in T_1 is connected to at most one item in T_2, and vice versa, it is said that there is a matching between T_1 and T_2. For a matching ζ between two transactions T_1 and T_2, if $\nexists$ matching ζ' *s.t.* $|\zeta'| > |\zeta|$, ζ is a maximum matching of T_1 and T_2, and $\min\{|T_1|, |T_2|\} \geqslant |\zeta|$. It should be noted that there is not necessarily only one maximum matching

for two transactions. The Matching Distance measure is given as follows and it actually refers to the minimum-weighted maximum matching problem.

Definition 6. *Let ζ be a maximum matching between T_1 and T_2, Matching Distance between T_1 and T_2 is defined as follows.*

$$D_m(T_1, T_2) = \min_{\zeta} \frac{\sum_{(i,j)\in\zeta} d(i,j)}{\min\{|\mathrm{T_1}|, |\mathrm{T_2}|\|}$$

Matching Distance does not take all of items into consideration. Especially, when the difference of size between two transactions is very large, a large proportion of items in the large transaction are not involved in calculation. It is desired to find some other measures that take all of items into consideration. Consider the following scenario.

Scenario 4:
For two transactions T_1 and T_2, suppose $|T_1| \geqslant |T_2|$. Companies hope to find a similarity measure by which every item in T_1 is compared with the most similar item in T_2 and every item in T_2 is compared with at least one item in T_1. The items that are far from any items in the other transaction are also taken into consideration of similarity between transactions as a penalty. Surjection Distance is introduced to against this scenario.

Surjection Distance:
For two transactions T_1 and T_2, here suppose $|T_1| \geqslant |T_2|$, if every item in T_1 is only connected to one item in T_2, and every item in T_2 is connected to at least one item in T_1, it is said that there is a surjection between T_1 and T_2. Based on the distance between items, Surjection Distance is given as follows.

Definition 7. *Let η be a surjection between transactions T_1 and T_2. Surjection Measure between T_1 and T_2 is defined as follows.*

$$D_s(T_1, T_2) = \min_{\eta} \frac{\sum_{(i,j)\in\eta} d(i,j)}{\max(|T_1|, |T_2|)}$$

In Surjection Distance measure, every item in the transaction that has a larger size is constrained to be connected to at most one item in the other transaction. However, in practical application, an item in the transaction with larger size may be very similar to some items in the other transaction. This application requires that a measure should take all of items into consideration while an item can be connected to multiple items in another transaction. Link Distance measure is such a measure that satisfies above requirements.

Link Distance:
Link is another assignment that every item in one transaction is connected with the other transaction. For two transactions T_1 and T_2, if every item in T_1 is connected to at least one item in T_2, and vice versa, it is said that there is a link between T_1 and T_2. Link Distance [5] is given as follows.

Definition 8. *Let τ be a link between T_1 and T_2, Link Distance between T_1 and T_2 is*

$$D_l(T_1, T_2) = \min_{\tau} \sum_{(i,j)\in\tau} d(i, j).$$

It should be noted that Link Distance is not normalized in definition in order to avoid being effected by some items that have many similar items in another transaction. Link Distance must be normalized before we employ Link Distance to compare the distance between diverse transactions.

Example 5. As shown in Figure 4, transactions T_1 and T_2 are denoted by $\times$ and $\bigcirc$, respectively. According to the definition of Link Distance, the link τ that determines the assignment of items is described by the dotted line in Figure 4(a). If Link Distance is normalized before it is applied to compare the similarity between transactions, adding some other pairwise link, *e.g.*, the solid line shown in 4(b), can shorten the Link Distance between transactions. However, it disobeys the original intention of assignment.

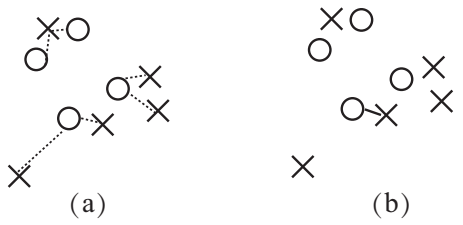

Fig. 4. Why Link Distance cannot be normalized in advance.

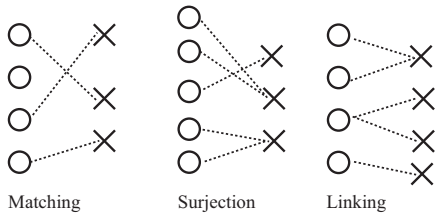

Fig. 5. Examples of Matching, Surjection and Linking.

Some examples of Matching, Surjection and Linking are visualized as examples in Figure 5. Set Distance measures that take assignment into consideration enrich the meanings the similarity between transactions. Employing Set Distance measures flexibly can help companies solve various problems in their daily business.

6 Conclusion

This paper refers to the issue that how to evaluate the similarity between customers based on various customer data. Various measures helping to segment customers based on transaction data were discussed. Set distance measures were introduced to evaluate the similarity between transactions from two perspectives: (1) the pairwise distance between items and (2) the assignment of items. The applications of set distances were discussed under various imaginary business scenarios for companies. No similarity measure performs better over all of other measures, and understanding the semantic meaning of similarity measures is critical for customer segmentation.

An obvious limitation of this paper is that we have not yet verified our analysis on real transaction data. The future work includes verifying the results of this paper, and proposing specific methods for segmenting customers based on transaction data.

References

1. Aggarwal, C.C., Procopiuc, C.M., Yu, P.S.: SFinding Localized Associations in Market Basket Data. IEEE Trans. Knowl. Data Eng. 14(1), 51–62 (2002)
2. Amiri, A.: Customer-oriented Catalog Segmentation: Effective Solution Approaches. Decision Support Systems 42(3), 1860–1871 (2006)
3. Das, G., Mannila, H.: Context-Based Similarity Measures for Categorical Databases. In: The 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp.201–210. Bilbao (2000)
4. Deza, E., Deza, M.: Dictionary of Distances. North-Holland, Amsterdam (2006)
5. Eiter, T., Mannila, H.: Distance Measures for Point Sets and Their Computation. Acta Inf. 34(2), 109–133 (1997)
6. Ester, M., Ge, R., Jin, W., Hu, Z.J: A Microeconomic Data Mining Problem: Customer-oriented Catalog Segmentation. In: The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 557–562. Seattle (2004)
7. Hsu, F.M., Lu, L.P., Lin, C.M.: Segmenting Customers by Transaction Data with Concept Hierarchy. Expert Syst. Appl. 39(6), 6221–6228 (2012)
8. Kim, S.Y., Jung, T., Suh, E.H., Hwang, H.S.: Customer Segmentation and Strategy Development based on Customer Lifetime Value: A Case Study. Expert Syst. Appl. 31(1), 101–107 (2006)
9. Kleinberg, J.M., Papadimitriou, C.H., Raghavan, P.: A Microeconomic View of Data Mining. Data Min. Knowl. Discov. 2(4), 311–324 (1998)
10. Ngai, E.W.T., Li, X., Chau, D.C.K.: Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification. Expert Syst. Appl. 36(2), 2592–2602 (2009)
11. Niiniluoto, I.: Truthlikeness. D. Reidel Pub. Comp., Dordrecht (1987)
12. Wang, M.T., Hsu, P.Y., Lin, K.C., Chen S.S.: Clustering Transactions with an Unbalanced Hierarchical Product Structure. In: 12th International Conference on Data Warehousing and Knowledge Discovery, pp.251–261. Bilbao (2007)
13. Woznica, A., Kalousis A.: Adaptive Distances on Sets of Vectors. In: The 10th IEEE International Conference on Data Mining, pp. 579–588. Sydney (2010)

14. Yang, Y.H., Padmanabhan B.: Segmenting Customer Transactions Using a Pattern-Based Clustering Approach. In: The 3th IEEE International Conference on Data Mining, pp. 411–418. Florida (2003)
15. Yen, S.F., Lee, Y.S.: An Efficient Data Mining Approach for Discovering Interesting Knowledge from Customer Transactions. Expert Syst. Appl. 30(4), 650–657 (2006)
16. Yun, C.H., Chuang K.T., Chen, M.S.: Clustering Item Data Sets with Association-Taxonomy Similarity. In: The 3th IEEE International Conference on Data Mining, pp. 697–700. Florida (2003)