Expert Systems with Applications 39 (2012) 6221–6228

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Segmenting customers by transaction data with concept hierarchy

Fang-Ming Hsu [a,*], Li-Pang Lu [b], Chun-Min Lin [b,c]

[a] Department of Information Management, National Dong Hwa University 1, Sec.2, Dahsueh Rd., Shoufeng, Hualien 974, Taiwan, ROC
[b] Department of Information Management, National Dong Hwa University, Taiwan, ROC
[c] General Education Center, Taiwan Hospitality & Tourism College, Taiwan, ROC

ARTICLE INFO

Keywords:
Customer segmentation
Concept hierarchy
Hierarchical clustering

ABSTRACT

The segmentation of customers is crucial for an organization wishing to develop appropriate promotion strategies for different clusters. Clustering customers provides an in-depth understanding of their behavior. However, previous studies have paid little attention to the similarity of different items in transaction. Lack of categories and concept levels of items, results from item-based segmentation methods are not as good as expected. Through employing a concept hierarchy of items, this study proposes a segmentation methodology to identify similarities between customers. First, the dissimilarity between transaction sequences is defined. Second, we adopt hierarchical clustering method to segment customers by their transaction data with concept hierarchy of consumed items. After segmentation, three cluster validation indices are used for optimizing the number of clusters of customers. Through the compassion of normalized index, the segmentation method proposed by this study rendered better results than other traditional methods.

1. Introduction

Knowledge regarding what customers think, what they want, and how to serve them is quite useful for companies wishing to generate suitable strategies in competitive markets. Owing to disparate desires, interests, and needs, gaining a comprehensive understanding of customers is difficult. Since an organization cannot normally serve all customers in a market (Dibb & Stern, 1995), customer segmentation is often used by organizations to categorize customers for marketing purposes. Customer segmentation divides customers into groups, with the members of each group having similar needs, characteristics, or behaviors. Segmentation also represents the key element of customer identification in customer relationship management (Ngai, Xiu, & Chau, 2009). After segmenting customers, companies can then use further strategies such as customer attraction to maintain relationships with customers and gain more profit from them. The selection of the customers' attributes is critical in their segmentation. The attributes for segmentation can be classified into two types: general attributes and transaction-based attributes (Tsai & Chiu, 2004). General attributes include customer base variables such as customer demographics, lifestyle, attitude and psychology (Bloom, 2005; Huang, Tzeng, & Ong, 2007; Kuo, Ho, & Hu, 2002; Lee & Park, 2005; Vellido, Lisboa, & Meehan, 1999). The main goal of these studies is to offer appropriate services or products to people based on different customer status. This approach is known as customer status orientation. Although general attributes are easy to operate and understand, the disadvantage of using general attributes is that customers with similar general attributes do not necessarily have similar purchasing behavior, and information about customer variables is also difficult to collect and is often incomplete (Tsai & Chiu, 2004). Finally, the segmenting results obtained using general attributes may miss some important trends because of its static nature (Böttcher, Spott, Nauck, & Kruse, 2009).

Articles using transaction-based attributes are mostly customer-value oriented (Böttcher et al., 2009; Chen, Chiu, & Chang, 2005; Cheng & Chen, 2009; Hosseini, Maleki, & Gholamian, 2010). These articles focus on high value customers. Several articles have been made to take RFM (Hughes, 1994) as their clustering or mining attributes. Verhoef and Donkers (2001) used socio-demographic information and transaction information to measure customers' potential value, and argued that companies should pay close attention to potentially valuable customers. Hwang, Jung, and Suh (2004) proposed a new lifetime value model by considering past profit contribution, potential benefit, and defection probability of customers, and finally segmenting customers based on their value. Furthermore, Kim, Jung, Suh, and Hwang (2006) used a lifetime value model with current value, potential value, and customer loyalty in segmentation. They also mentioned that current value provides a financial viewpoint, potential value indicates cross-selling opportunities, and customer loyalty estimates durability of the previous two values. These articles focus on customer value in order to gain the

* Corresponding author. Tel.: +886 3 8633107; fax: +886 3 8633100.
E-mail addresses: fmhsu@mail.ndhu.edu.tw (F.-M. Hsu), jodan.lu@msa.hinet.net (L.-P. Lu), jimmy@tht.edu.tw (C.-M. Lin).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.12.005

most profits for firms. However, from a customer retention standpoint, these approaches may be not good enough to ensure firm/customer relationship are maintained in the long-term, because firms do not have any idea what customers like or prefer, and the approaches lack the most important information about products.

There are only a few articles dealing with transaction-based attributes which consider product information as an important factor in segmentation (Lu & Wu, 2009; Tsai & Chiu, 2004; Tsai & Shieh, 2009). Because the studies by Lu and Wu (2009), Tsai and Chiu (2004), and Tsai and Shieh (2009) looked at product information, they can be classified as customer preference-oriented. By discovering customers' preferences, firms can then deliver the right marketing strategy to the right customer cluster, and ultimately can improve the quality of the customer relationship and enhance customer loyalty. Although these three studies considered product information, they did not specifically consider the relationships among items. When there are huge numbers of items provided by an enterprise, this means that similarities between any two customers are often actually very small, owing to customer preferences for similar, but not exactly the same, items. Therefore, segmentation results in these studies have not been as good as expected.

Based on a concept hierarchy of items, this study proposed a segmentation methodology to identify relationships between customers. Generally speaking, two more similar items have strong relationship than that between two less ones. The rest of this paper is organized as follows: Section 2 describes an overview of the related research including data representation, similarity measure and its example, hierarchical clustering algorithm, and clustering criteria functions. Section 3 presents the proposed procedure and briefly discusses its architecture. Section 4 analyzes the experimental results. Finally, a summary and conclusion are presented in Section 5.

2. Related works

In this section, a formal representation for transaction data is illustrated first. Next, the item concept hierarchy is defined. Then, the dissimilarity measurement between two pieces of transaction data can be calculated. After that, hierarchical clustering is adopted to segment transaction data, along with an explanation of why we use it. In the end, three clustering criteria functions are used to verify the cluster results in order to find the best cluster numbers.

2.1. Data representation

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of transaction records and $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. A transaction record t_i is an itemset, represented as $(x_1, x_2, \ldots, x_m)$ where $x_j \in I$ *for* $1 \leqslant j \leqslant m$. A transaction sequence s_i is an ordered list of transaction records, represented as $\langle t_1, t_2, \ldots, t_n \rangle$ where $t_j \in T$ *for* $1 \leqslant j \leqslant n$. An itemset of a transaction sequence is the items in a transaction sequence.

2.2. Concept hierarchy of items

A hierarchy is an arrangement of objects in which the objects are represented as being "above," "below," or "at the same level" as one another. A subsumptive containment hierarchy is a classification of objects from the general to the specific. Other names for this type of hierarchy are "compositional hierarchy", "taxonomic hierarchy" and "IS-A hierarchy". A lower-level object automatically is a member of the higher level. Let $C = \{c_1, c_2, \ldots, c_n\}$ be a set of categories for items and $i_j \subset b_1 \subset b_2 \ldots \subset b_m$ where $b_k \in C$ for $1 \leqslant k \leqslant m$. This study takes seven (7) transaction sequences with four (4) customers and six (6) different items to explain the proposed method. Table 1a contains the original transaction records in which the items are organized into a hierarchical relation as show in Fig. 1. Table 1b represents the transaction sequences and their itemset as retrieved by Table 1a.

2.3. Semantic similarity between words

In the previously mentioned transaction-based research (Lu & Wu, 2009; Tsai & Chiu, 2004; Tsai & Shieh, 2009), items are treated independently. However, most organizations define different categories of items. Moreover, a particular category may have numerous subcategories, and so on. Having categories makes items easy to look up, arrange, classify, and stand in relation to one another.

For expressing relation between items, this study refers to the semantic similarity between words proposed by Li, Bandar, and McLean (2003). They considered a hierarchical semantic knowledge base to calculate the semantic similarity between words. The lexical hierarchy is connected by following trails of super-ordinate terms in "is a" or "is a kind of" (ISA) relations. Fig. 2 shows a portion of such a hierarchical semantic knowledge base:

Note that words are associated with concepts in the ISA hierarchy. If we want to define the semantic similarity between words, there are two important properties to consider, as follows:

2.3.1. Path length

Path length means the shortest length of path connecting the two concepts containing the two words, w_1 andw_2, in the hierarchy of semantic knowledge base as shown in Fig. 2. For example, the shortest path length between boy and girl is boy → male → person → female → girl, so the path length is 4. It can be determined from one of three cases:

(1) If word w_1 and w_2 are in the same concept, the path length between them is 0.
(2) If word w_1 and w_2 are not in the same concept, but the concept for w_1 and the concept for w_2 contain one or more of the same words, the path length between them is 1
(3) If w_1 and w_2 are not in the same concept nor do their concepts contain the same words, the path length is the actual path length distance between w_1 and w_2.

However, the original path length could not be used directly without transforming it into [0,1] to express the similarity between words. It is intuitive that when the path length increases infinitely, the similarity would monotonically decrease to zero.

2.3.2. Scaling depth effect

It is intuitive that concepts at upper layers have more general semantics and less similarity than concepts at lower layers, as measured between any two objects. In biological taxonomy, kingdoms

Table 1
The transaction data example and its converted transaction sequences.

TID	Customer ID	Item
(a) Original transaction data example		
t_1	0001	*A*, *B*, *D*
t_2	0002	*A*, *B*, *C*
t_3	0003	*A*, *F*
t_4	0004	*A*, *B*, *E*
t_5	0001	*C*, *E*
t_6	0002	*A*, *B*
t_7	0004	*C*, *D*
Customer ID	transaction sequence	Item of transaction sequence
(b) Converted transaction sequences		
0001	$\langle t_1, t_5 \rangle$	$\langle (A,B,D), (C,E) \rangle$
0002	$\langle t_2, t_6 \rangle$	$\langle (A,B,C), (A,B) \rangle$
0003	$\langle t_3 \rangle$	$\langle (A,F) \rangle$
0004	$\langle t_4, t_7 \rangle$	$\langle (A,B,E), (C,D) \rangle$