Customer segmentation of multiple category data in e-commerce using a soft-clustering approach

Roung-Shiunn Wu [a,*], Po-Hsuan Chou [b]

[a] Department of Information Management, National Chung Cheng University, No. 168 University Road, Min-Hsiung, Chia Yi 62102, Taiwan
[b] Graduate Institute of Information Management, National Chung Cheng University, No. 168 University Road, Min-Hsiung, Chia Yi 62102, Taiwan

ARTICLE INFO

ABSTRACT

The segmentation of online consumers into multiple categories can contribute to a better understanding and characterization of purchasing behavior in the electronic commerce market. Online shopping databases consist of multiple kinds of data on customer purchasing activity and demographic characteristics, as well as consumption attributes such as Internet usage and satisfaction with services. Information about customers uncovered by segmentation enables company administrators to establish good customer relations and refine their marketing strategies to match customer expectations. To achieve optimal segmentation, we developed a soft clustering method that uses a latent mixed-class membership clustering approach to classify online customers based on their purchasing data across categories. A technique derived from the latent Dirichlet allocation model is used to create the customer segments. Variational approximation is leveraged to generate estimates from the segmentation in a computationally-efficient manner. The proposed soft clustering method yields more promising results than hard clustering and greater within-segment clustering quality than the finite mixture model.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

The updating of current Internet technologies and the creation of new ones allow for the continued growth of e-commerce. There has been a tendency for customers to move from traditional shopping outlets to the Internet as a new shopping channel. The constraints engendered by opening times or geographical boundaries are no longer much of an issue. Even the styles of shopping are continually changing as more businesses and customers convert to e-commerce, regardless of whether the transactions are business-to-business or business-to-customer.

Understanding the interaction patterns between online companies and their customers is important in the face of this online market growth. Companies often set goals for profit-making. To achieve these goals, they must perform an analysis of how they manage their customer relations and adjust their marketing strategies. A new transaction model based on customer service and satisfaction shows that price is not the only major determinant of whether a customer decides to buy a product (Bhatnagar and Ghose 2004). It is also important that the company and the customer reach a consensus about the value of the product and of good customer service. It follows from this point of view that companies should not try to develop a single product that satisfies

every customer. Instead, they must learn about the shopping behavior of different kinds of customers and develop separate products for each segment. In other words, customer classification based on buying behavior is essential for developing a successful marketing strategy, which in turn creates and maintains competitive advantage.

Dividing customers into subgroups or segments through the application of clustering techniques can be used to analyze the relationships between customers and Internet shopping channels. The results of such analyses can provide companies with information about what customers expect when they buy a product over the Internet. By acting on the results of these analyses, companies can improve their customer-relations management by enhancing customer satisfaction and loyalty.

Previous methods have divided customers into segments by considering how frequently they make purchases and how much money they spend (Chen et al. 2009). Other information, such as customer satisfaction with the services and their concerns about using the Internet for shopping, can also contribute to a better understanding of customer buying behavior. Segmentation derived from considering customers' buying behavior in multiple categories demonstrates cross-category dependency (Heilman and Bowman 2002). Such data are usually collected separately and placed in different categories in the model.

To place all of the relevant customer information in a single category is computationally infeasible when the numbers of

---

* Corresponding author. Tel.: +886 5 2720411x34621; fax: +886 5 2721501.
  E-mail address: roungwu@ccu.edu.tw (R.-S. Wu).

items and customers are large. Thus, in the present study, we organized our online shopping data under multiple headings. We used a latent mixed-class membership (LMCM) clustering technique, a probabilistic model for analyzing latent class interaction patterns and mixed-class membership of customers, to classify customers in terms of their responses to these items for the purpose of exploring their buying behavior. We also used a technique derived from the latent Dirichlet allocation (LDA) model to create our customer segments. This is often used to find latent-class structures in large samples of real-world data. To generate estimates from the segmentation, we chose variational approximation because of its computational efficiency. This method makes use of an expectation–maximization (EM) procedure for parameter estimation.

In full-membership modeling, a customer's data are classified into one of segments. However, it is also possible to use mixed-membership modeling, also called partial-membership modeling, in which a given customer's data can be included in multiple segments (Erosheva 2004). Our segmentation method uses the mixed-membership approach because it corresponds better to the real world of business.

The paper is structured as follows. In Section 2 we review the relevant literature. The segmentation model is presented in Section 3 and the variational approximation method in Section 4. In Section 5, we describe a soft-clustering approach for customer segmentation. In Section 6, we present the design, methodology, results, and interpretive discussion of the experiments we designed to test our customer segmentation model. The managerial implications and conclusions are presented in Sections 7 and 8.

2. Literature review

We next review the relevant literature pertaining to the recency and frequency of shopping and the monetary (RFM) model, segmentation of online customers, multiple category segmentation, and the latent Dirichlet allocation model.

2.1. RFM models and customer relationship management

Customer relationship management (CRM) refers to the procedures companies use to gain lifetime customer loyalty and thereby increase competitive advantage and profits. The mantra for CRM is "the right time, the right channel, the right price, and the right customers." Due to the maturity of online shopping markets and advances in Internet technology, companies have increasingly been adopting CRM to manage their customer relations (Cheng and Chen 2009). A successful web-based CRM not only can strengthen a company's virtual interactions with its customers, but it also provides a way to generate more revenue by developing an e-commerce strategy and redesigning its web pages.

According to Pareto's 80/20 rule (Bult and Wansbeek 1995, Stone and Jacobs 2001), a small percentage of customers make the major contribution to a company's revenue. Thus, it is better to retain those customers who spend the most or have stayed with the company the longest than to acquire new customers. But how does one retain valuable customers? Companies must learn about customer buying behavior and then adopt the most effective marketing strategy for each customer segment. The most commonly-used strategies are one-to-one marketing and the recommendation system (Jiang and Tuzhilin 2006). Both strategies personalize products and services. With the recommendation system, the idea is to let customers search for detailed information of interest to them on personalized web pages that offer content tailored to their particular market segment.

CRM enhances opportunities to use information to understand customers and create customer loyalty (Zabah et al. 2004). Its successful application requires a cross-functional integration of the processes, customers, operations, and marketing capabilities enabled by information, technology, and applications (Payne and Frow 2005).

The best-known model of customer shopping characteristics is the RFM model. This simple and easy to understand model divides customers into segments in terms of the following shopping characteristics: recency of shopping, frequency of shopping, and money spent while shopping. In its simplest form, the RFM model makes use of a point system. Customers are given scores that are used to assign them to segments. In empirical research by Hughes (1996), customer records in a database were generally divided into five equal quintiles for each of the three RFM characteristics. However, this quintile method creates potential problems, because the arbitrary cutoffs are applied to the three RFM characteristics rather than to the customers (Migkautsch 2000). Using the means of continuous distributions of customer scores rather than quintiles not only yields greater sensitivity at both the top and bottom of the distribution, but it also isolates single customers.

Another shortcoming of the RFM model is that, because of its three-dimensional nature, its predictive capacity is inferior to that of more sophisticated methods such as χ^2 automatic interaction detection and regression analysis (McCarty and Hastak 2007).

2.2. Segmentation of online customers

The ever-increasing computational capacities for data storage and processing, coupled with advanced data mining techniques, increase the opportunities to obtain information from online shopping databases (Ngai et al. 2009, Regielski et al. 2002). Data mining provides for the extraction of hidden or predictive information from large databases, thereby enabling companies to identify valuable customers, predict their buying behavior, and make proactive, knowledge-based decisions.

In supervised classification methods (Vellido et al. 1999) such as neural networks, linear discriminant analysis, and decision-tree induction, the available observations or samples have class labels. The aim is to construct a model that assigns one of these class labels to each new observation. When the available observations are not identified as belonging to one of the pre-defined classes, unsupervised clustering methods can be used to infer the class information from the distribution of observations. Partition-based clustering algorithms group observations that are close to one another in distance (Bose and Chen 2010), and model-based clustering approaches estimate membership probabilities for the purpose of assigning observations to the appropriate clusters. Latent class clustering, also referred to as finite mixture model clustering, was designed for the analysis of grouped categorical data (Magidson and Vermunt 2002). Latent classes are unobservable subgroups or segments.

In the finite mixture model, although the classes overlap, each observation is considered as belonging to one class only. One can determine only the probability that an observation belongs to a given class. In many real-world situations, the classes are defined in such a way that class membership is not exclusive, but rather a matter of degree. Fuzzy logic enables one to use the non-numerical attributes to capture the imprecision of human perception. Examples are terms such as "high loyalty" and "low loyalty." Both qualitative and quantitative attributes can be expressed by the membership functions. With fuzzy clustering, an online customer can be treated as a member of several different classes simultaneously. Fuzzy customer classes provide differentiated assessments of customers and customer segments (Meier and Werro 2007), allowing company administrators to improve customer