Electronic Commerce Research and Applications 10 (2011) 331–341

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

# 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

*Article history:*
Received 19 October 2009
Received in revised form 8 November 2010
Accepted 8 November 2010
Available online 25 November 2010

*Keywords:*
Customer segmentation
Clustering
Customer relationship management
Data mining
Expectation–maximization (EM) procedure
Latent class

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.



## 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).

1567-4223/$ - see front matter 
doi:10.1016/j.elerap.2010.11.002

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 $\chi^2$ 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 relations and refine their marketing campaigns. In fuzzy clustering, the classes themselves are not necessarily well-defined or mutually-exclusive. The latent-class model can be turned into a fuzzy-class model by embedding fuzzy clustering methods that provide for the assessment of overlapping clusters of categorical data (Yang and Yu 2005).

### 2.3. Multiple category segmentation

Many real-world applications of classification, such as e-commerce (Heilman and Bowman 2002) and medical informatics (Spangler et al. 1999), use multi-category data. Online shopping databases contain many different kinds of information about, for example, customers' buying behavior. Customer segmentation involves assigning sample customers to clusters or segments by noting their behavior patterns and the relationships among the items in the different categories. For multi-category data, the data points can be modeled as a *k*-partite graph of heterogeneous types of nodes. Spectral clustering techniques are used to partition graphs into clusters (Long et al. 2006). Spectral clustering techniques make use of similarity graphs to represent data points. Each vertex in the graph represents a data point. Two vertices are connected if the similarity between the corresponding data points is larger than a certain threshold value. The similarity graphs are used to reformulate the clustering problem. As we are dealing with a full-membership model, each customer can be assigned to only one of the clusters. However, it is also possible to cluster customers' shopping behavior by using a grade of membership (GoM) model, in which the same customer can be assigned to more than one cluster. This kind of model is an extension of the latent-class model (Bhatnagar and Ghose 2004), and it incorporates the features of mixed-membership models (Erosheva and Feinberg 2005).

Consider a multiple classification scheme (Spangler et al. 1999) in a medical context. Multiple classification means that the categories are well-defined and mutually-exclusive but the observations themselves transcend categorical boundaries. A single patient has multiple memberships across categories. The patients' records include diagnoses, the procedures performed, and demographic and other case-specific information. A patient's diagnoses are classified by the International Classification of Disease (ICD-9) coding system. A procedural taxonomy is classified by common procedural terminology (CPT). Scheduling surgery requires that patterns be uncovered linking each patient's ICD-9 and demographic information to the CPT outcomes. Data mining in this situation is a multiple-classification task, because several procedures may be performed on a patient during a single operation. Spangler et al. (1999) compared the success of three supervised learning methods for which the class categories labeled by CPT codes were predefined.

Because in the present experiment the class labels were unknown, we had to use unsupervised learning techniques for segmentation. Our soft-clustering approach created not only overlapping clusters but also mixed-memberships for each customer. In contrast, in multiple classification schemes, the clusters are well-defined and mutually-exclusive, but the observations themselves transcend cluster boundaries.

### 2.4. Latent Dirichlet allocation

The Dirichlet distribution is one of the class of distributions called the exponential family. Given a set of multinomial observations, the parameters for the Dirichlet distribution can be estimated by maximizing the log-likelihood of the data. Although there is no closed-form solution for maximizing log-likelihood (Minka 2001), it is convex in the Dirichlet parameter $\alpha$ and guaranteed to have a unique optimal value. The Dirichlet distribution is used as prior to the parameters of a multinomial distribution, which is also a member of the exponential family. The multinomial and Dirichlet distributions form a conjugate pair.

The latent Dirichlet allocation (LDA) model is a generative probabilistic model for the collection of discrete data (Attias 2000). The basic idea is that customers are randomly assigned to latent classes using the mixed-membership principle. Based on the exchangeability assumption (Blei et al. 2004), LDA has been applied to text modeling, content-based image retrieval, and bioinformatics (Barnard et al. 2003, Blei et al. 2003, Airoldi et al. 2006). Exact inference is intractable with LDA, but approximate inference can be used for parameter estimation (Nowicki and Snijders 2001).

The main estimation problem that must be solved to use LDA is how to compute the posterior probabilities for hidden random variables. Given the model parameters, the marginal probabilities are obtained by integrating over hidden random variables the joint probability of a set of latent classes and the individual observations. The problem of computationally-intractable marginal probabilities can be solved by using variational approximation. Specifically, we can use Jensen's inequality to obtain an adjustable lower bound, indexed by a set of variational parameters, for the log-likelihood (Minka 1998). A simple way to obtain a usable family of lower bounds is to modify the LDA model. By endowing the simplified LDA model with variational parameters, we obtain a family of variational probabilities for the latent variables.

Finding a strict lower bound for the log-likelihood can be viewed as an optimization problem. The optimal values for the variation parameters can be obtained by minimizing the Kullback–Leibler (KL) divergence between the approximate probability and the true probability. The minimization is achieved by implementing an expectation–maximization procedure.

## 3. Segmentation model formulation for multi-category data

The multi-category data set used in our experiments on customer segmentation is illustrated in Fig. 1. The data were obtained from an online customer survey questionnaire, in which the questions can be organized in sections or categories. Questions concerning customers' reactions to the service they received online are included in the *Satisfaction with Service* category. The *Shopping Behavior* category uses data on customers' online shopping frequency and money spent. Questions about customers' experience using the Internet are listed under the *Internet Usage* category. The *Demographics* category includes questions such as the customer's age and gender. The categories and the individual items are presented in Table 1.

The responses in each category were transformed according to the following formulas: for *Satisfaction with Service*: $X^{(1)} = \{x_i^{(1)}\}_{i=1}^N$; for *Shopping Behavior*: $X^{(2)} = \{x_i^{(2)}\}_{i=1}^N$; for *Internet Usage*: $X^{(3)} = \{x_i^{(3)}\}_{i=1}^N$, and for *Demographics*: $X^{(4)} = \{x_i^{(4)}\}_{i=1}^N$.$X^{(r)}$ denotes the $r$ th category and $x_i^{(r)}$ is the score of the $i$ th customer in the $r$ th category. The scores for the individual items in each category are represented as $x_{ij}^{(r)}$.

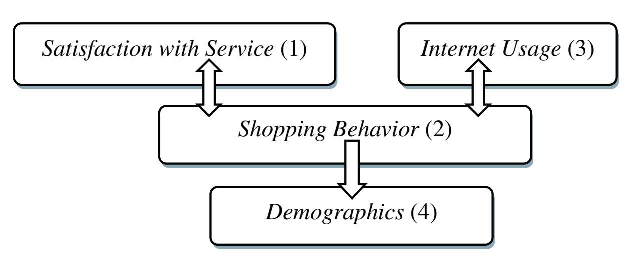


**Fig. 1.** The multiple categories.

**Table 1**
The online customer survey questionnaire.

| Items | Diversity of products | Quality of products | Security of transactions | After-sale service | Delivery service | Personalized service | Pricing |
|---|---|---|---|---|---|---|---|
| *Category 1: Satisfaction with Service* | | | | | | | |
| Response options | 1: highly satisfied, 2: satisfied, 3: neutral, 4: unsatisfied, 5: very unsatisfied, 6: no opinion | | | | | | |

| Items | Frequency of online shopping | Total money spent shopping online |
|---|---|---|
| *Category 2: Shopping Behavior* | | |
| Response options | 1: 1–2 times, 2: 3–4 times<br>3: 5–7 times, 4: 8–10 times<br>5: 11 times or more | 1: <$1000, 2: $1000–1999, 3: $2000–4999<br>4: $5000–9999, 5: $10,000–19,999<br>6: $20,000–29,999, 7: $30,000–39,999<br>8: $40,000–49,000, 9: $50,000 or more |

| Items | Years of Internet use | Internet use per week | Times per week browsing for products | Expected Internet use in next year |
|---|---|---|---|---|
| *Category 3: Internet Usage* | | | | |
| Response options | 1: < 1/2 | 1:<2, 2: 2–5, | 1: never, 2: 1–2 | 1: much less often |
| | 2: 1/2–1 | 3: 6–10, 4: 11–15, | 3: 3–5, 4: 4–10 | 2: less often, 3: same, 4: more often |
| | 3: 1–2, 4: 2–3 | 5: 16–20, 6: 21–25, | | |
| | 5: 3 or more | 7: 26–30, 8: >30 | 6: 16 or more | 5: much often |

| Items | Age | Gender | Income | Marital status |
|---|---|---|---|---|
| *Category 4: Demographics* | | | | |
| Response options | 1: under 15, 2: 16–20, 3: 21–25, 4: 26–30 | 1: female | 1:<$15 K, 2: $15 K–35 K | 1: married |
| | | 2: male | 3: $35 K–55 K, 4: $55 K–75 K | |
| | 5: 31–35, 6: 36–40 | | 5: $75 K–95 K, 6: $95 K–115 K | 2: single |
| | 7: 41–50, 8: over 50 | | 7: $115 K or more | |

### 3.1. Interactions between the latent classes for the categories

For estimating the latent mixed-class membership of each customer and the interactions between the latent classes of the different categories, a hierarchy of probabilistic assumptions was created as a basis for deriving how the customers in the various categories interact with one another. The customers can interact in multiple ways. We describe a probabilistic model of customer interaction patterns, using multi-category data sets as working examples.

Assume that we observe cross-category interactions among $N$ customers. The interaction patterns are represented by the $N \times N$ adjacency matrix $A$, where $A_{ij}$ represents the interaction between the $i$th customer in one category and the $j$th customer in another category. For each pair of customers, the presence or absence of an interaction is determined by choosing a latent class for the customers from a customer-specific distribution with parameters associated with the pair of latent classes. The similarity of the customers is inferred from the interactions.

For a model with $K$ latent classes, there is a set of $K$-dimensional Dirichlet parameters $\alpha = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$ and a matrix parameter $\eta_{KxK}$. Assume that discrete responses to $J$ questions are recorded for $N$ customers. Let $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})$ denote a response, where $x_{ij}$ is a categorical random variable denoting the response of the $i$ th customer to the $j$ th question. $x_{ij}$ takes on values $l_j \in \{1, 2, \ldots, L_j\}$, where $L_j$ denotes the number of categorical values under which the $j$ th question falls. Let $\theta_i = (\theta_{i1}, \ldots, \theta_{ik}, \ldots, \theta_{iK})$ be a latent class vector with element $\theta_{ik}$ representing the membership scores of the $i$th customer in latent class $k$. The mixed-membership score is the proportion of a customer's memberships that falls in each latent class, given the constraint $\sum_{k=1}^{K} \theta_{ik} = 1$.

The membership scores $\theta_{ik}$ are the elements of a random vector forming the Dirichlet distribution $Dir(\alpha)$, which is parameterized by $\alpha = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$ Consider vector $z_i = (z_{i1}, \ldots, z_{ij}, \ldots, z_{iJ})$ as consisting of $J$ latent variables, one for each observable item. Each latent variable $z_{ij}$ takes a value from $\{1, 2, \ldots, K\}$. To specify the conditional distribution of the items, we make the assumption that $x_{ij}$ is directly influenced by the latent variable $z_{ij}$.

Let each $\theta_i$ be a Dirichlet distribution of the parameter $\alpha$ and each $z_i$ be a multinomial distribution with the parameter $\theta_i$. $z_i$ and $\theta_i$ are hidden random variables. The generative process of the observations is as follows:

1. For each customer category $i = 1, \ldots, N$, the membership score is defined as: $\theta_i / Dir(\alpha)$.
2. For each customer in category $i, j, k = 1, \ldots, N$, the latent class membership is defined as: $z_i \sim Mult(1, \theta_i)$; $z_j \sim Mult(1, \theta_j)$; and $z_k \sim Mult(1, \theta_k)$.
3. For each customer in category $i, j, k = 1, \ldots, N$, the responses to individual items are defined as: $x_i \sim P(x_i|z_i, \eta)$; $x_j \sim P(x_j|z_j, \eta)$; and $x_k \sim P(x_k|z_k, \eta)$.

### 3.2. Procedures for computing mixed-membership scores

The parameters $\alpha = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$ are normalized by $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. $\alpha$ represents the spread of the Dirichlet distribution within the hidden set determined by the latent classes. $\alpha_k$ represents the proportion of the customer memberships that is assigned to the $k$ th latent class. In our experiments, we assumed that the initial Dirichlet distribution $\alpha$ is uniform $(1/K)$, and we set all the elements of the interaction matrix $\eta_{KxK}$ to zero. The procedure used for computing the mixed-membership scores is as follows:

- *Step 1:* The customers in each category were randomly assigned to $K$ latent classes.
- *Step 2:* For each latent class, $K$-means were calculated for each item as representing values for the latent class. For *Shopping Behavior* and *Internet Usage* categories, the Tanimoto distance was used to measure the similarity between the customer scores in the categories and the corresponding values in the latent classes. For *Satisfaction with Service* category, item's categorical values adjacency, which could take on the values 1, 3/8,

1/8, and 0, was used to measure similarity. The similarities between customers' item and latent class scores were evaluated based on the items. The joint item similarity scores were then normalized and used as a distribution of probabilities for each latent class. These distribution probabilities were summed and represented as vectors $\alpha_1, \alpha_2, \ldots, \alpha_K$.

- *Step 3:* A Dirichlet distribution probability $p(\theta_i|\alpha)$ and a multinomial distribution probability $p(z_i|\theta_i)$ were computed for each customer. The joint probability for the two distributions, $p(z_i|\theta_i)p(\theta_i|\alpha)$, was calculated and then normalized to create the mixed-membership scores for each latent class.
- *Step 4:* The model parameters, $\eta_{gh}^{(12)}$ and $\eta_{hl}^{(23)}$, representing the interactions between the latent classes for the different categories, were evaluated according to formulas (13) and (14), which are described in Section 4.

Steps 2–4 were computed iteratively until there was convergence.

The Tanimoto distance $(x, y)$ is defined as

$$\frac{x^t \cdot y}{x^t \cdot x + y^t \cdot y - x^t \cdot y} \tag{1}$$

For instance, the similarity between the customer item scores (3, 3, 6, 2) and the corresponding latent-class values (3, 8, 6, 3) is $\frac{3*3+3*8+6*6+2*3}{(3^2+3^2-3*3)+(3^2+8^2-3*8)+(6^2+6^2-6*6)+(2^2+3^2-2*3)} = 0.7425$. The similarity of adjacent category values is evaluated based on the items. For instance, if a customer's item scores are (3, 3, 3, 4, 3, 3, 6) and the corresponding latent-class values are (4, 4, 3, 3, 2, 2, 3), their similarity is (3/8 + 3/8 + 1 + 3/8 + 3/8 + 3/8 + 0)/7 = 0.4107.

The elements of matrix $\eta_{gh}^{(12)}$ denote the degree of interaction between the latent-class pairs $(g, h)$ for *Satisfaction with Service* and *Shopping Behavior*. The elements of matrix $\eta_{hl}^{(23)}$ denote the interaction between the latent-class pairs $(h, l)$ for *Shopping Behavior* and *Internet Usage*. The adjacent matrices $A_{ij}^{(12)}$ and $A_{jk}^{(23)}$ were evaluated using:

$$\begin{aligned} A_{ij}^{(12)} &\sim z_i \eta_{gh}^{(12)} z_j^T \\ A_{jk}^{(23)} &\sim z_j \eta_{hl}^{(23)} z_k^T \end{aligned} \tag{2}$$

The elements of the matrix $A_{ij}^{(12)}$ represent the interaction between the $i$th customer under the *Satisfaction with Service* category and the $j$th customer under the *Shopping Behavior* category. The elements of the matrix $A_{jk}^{(23)}$ represent the interaction between the $j$th customer under *Shopping Behavior* category and the $k$th customer under *Internet Usage* category. The dimensions of matrices $z_i, z_j$, and $z_k$ are $N \times K$, whereas the dimensions of matrices $\eta_{gh}^{(12)}$ and $\eta_{hl}^{(23)}$ are $K \times K$. There can be a very large number of customers ($N$) in each category. To do the computations, there is no need to evaluate the product of the $N \times N$ matrices, because the number of latent classes ($K$) is much smaller than $N$. The proposed method allows scaling for a very large numbers of customers.

To determine the model parameters $\alpha$ and $\eta$, it is required to first determine the likelihood. The generative process leads to a joint probability distribution over observations and the latent variables. The joint distribution is

$$\begin{aligned} p(\theta, z, x|\alpha, \eta) = &\prod_{i=1}^{N} p\left(z_i^{(1)}|\theta_i^{(1)}\right) p\left(\theta_i^{(1)}|\alpha^{(1)}\right) \prod_{j=1}^{N} p\left(z_j^{(2)}|\theta_j^{(2)}\right) p\left(\theta_j^{(2)}|\alpha^{(2)}\right) \\ &\times \prod_{k=1}^{N} p\left(z_k^{(3)}|\theta_k^{(3)}\right) p\left(\theta_k^{(3)}|\alpha^{(3)}\right) p\left(A_{ij}^{(12)}|z_i^{(1)}, z_j^{(2)}, \eta_{gh}^{(12)}\right) \\ &\times p\left(A_{jk}^{(23)}|z_j^{(2)}, z_k^{(3)}, \eta_{hl}^{(23)}\right) \end{aligned} \tag{3}$$

The marginal probability of the observations is calculated by integrating over the latent variables $\theta$ and z (Minka 2001).

$$p(x|\alpha, \eta) = \int_\theta \int_z p(\theta, z, x|\alpha, \eta) dz\, d\theta \tag{4}$$

## 4. Parameter estimation using variational approximation

Because it is not possible in practice to compute the marginal probabilities in Eq. (4), we developed a method to deal with this problem that yields a variational approximation and corresponding parameter estimates. The aim is to determine the posterior probabilities for the mixed-membership scores of the customers and the latent class interactions.

### *4.1. A graphical representation of the variational approximation*

Fig. 2 presents the variational approximation graphically.

The full factorization of the variational distribution is:

$$\begin{aligned} q(\theta, z|\gamma, \varphi) &= \prod_{i=1}^{N} q(\theta_i|\gamma) \prod_{i=1}^{N} q(z_i|\varphi_i) \\ &= \prod_{i=1}^{N} Dir\left(\theta_i^{(1)}|\gamma^{(1)}\right) Mult\left(z_i^{(1)}|\varphi_i^{(1)}\right) \prod_{j=1}^{N} Dir\left(\theta_j^{(2)}|\gamma^{(2)}\right) \\ &\quad \times Mult\left(z_j^{(2)}|\varphi_j^{(2)}\right) \prod_{k=1}^{N} Dir\left(\theta_k^{(3)}|\gamma^{(3)}\right) Mult\left(z_k^{(3)}|\varphi_k^{(3)}\right) \end{aligned} \tag{5}$$

Given the model parameters and observations, it is necessary to compute the posterior probabilities to determine the mixed-membership for each latent class. By using variational approximation, we can determine a lower bound for the log-likelihood and an approximate posterior distribution for each customer's membership vector. The approximation assumes a latent-variable variational distribution $q(\theta, z)$ that is a close fit to the final Kullback–Leibler (KL) convergence. This value, corresponding to the maximum lower bound $KL(\gamma, \phi||\alpha, \eta)$ for the log probabilities of the observations, is calculated by using Jensen's inequality:

$$\begin{aligned} \log p(x|\alpha, \eta) \geqslant &\sum_{j=1}^{3} \sum_{i=1}^{N} \mathrm{E}_q\left[\log p\left(\theta_i^{(j)}|\alpha^{(j)}\right)\right] \\ &+ \sum_{j=1}^{3} \sum_{i=1}^{N} \mathrm{E}_q\left[\log p\left(z_i^{(j)}|\theta_i^{(j)}\right)\right] \\ &+ \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{E}_q\left[\log p\left(A_{ij}^{(12)}|z_i^{(1)}, z_j^{(2)}, \eta_{gh}^{(12)}\right)\right] \\ &+ \sum_{j=1}^{N} \sum_{k=1}^{N} \mathrm{E}_q\left[\log p\left(A_{jk}^{(23)}|z_j^{(2)}, z_k^{(3)}, \eta_{hl}^{(23)}\right)\right] \\ &- \mathrm{E}_q[\log q(\theta, z)], \end{aligned} \tag{6}$$

where the expectations ($E$) are evaluated based on $q(\theta, z)$. The true expectation values are the first four terms on the right side of Eq. (6), whereas the fifth term is the approximation expectation value, which can be calculated with:

$$\begin{aligned} E_q[\log q(\theta, z)] &= E_q[\log q(\theta|\gamma)] + E_q[\log q(z|\varphi)] \\ &= \sum_{j=1}^{3} \sum_{i=1}^{N} \mathrm{E}_q\left[\log q(\theta_i^{(j)}|\gamma^{(j)})\right] + \sum_{j=1}^{3} \\ &\quad \times \sum_{i=1}^{N} \mathrm{E}_q\left[\log q(z_i^{(j)}|\varphi_i^{(j)})\right] \end{aligned} \tag{7}$$

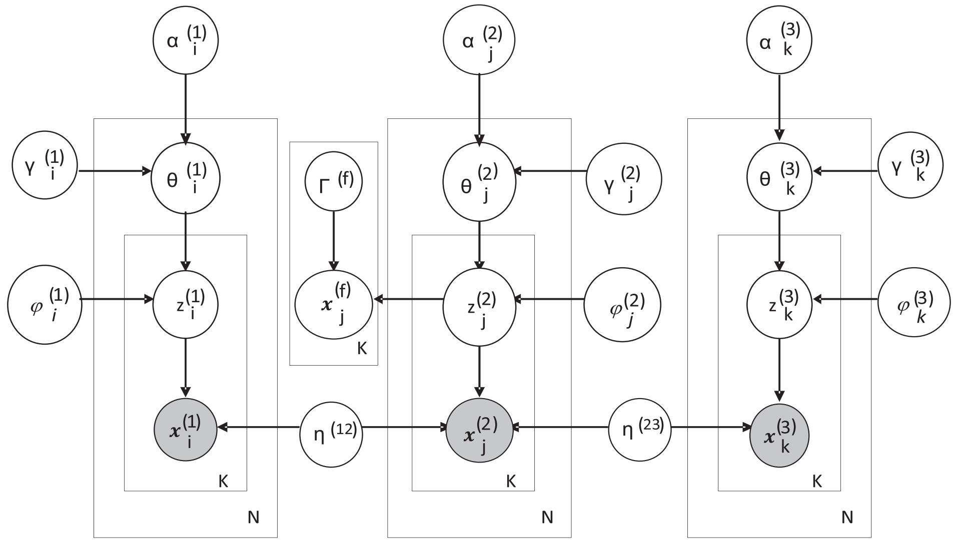


**Fig. 2.** The variational approximation model.

### *4.2. Procedures for determining the optimal expectation and for parameter estimation*

To maximize the lower bound for the likelihood, the variational approximate was estimated for the fixed values of the model parameters $\alpha$ and $\eta$. Then, we maximized the lower bound with respect to $\alpha$ and $\eta$. This required a quantified optimization for the maximum likelihood estimate of the Dirichlet parameters $\alpha$. The variational parameters, $\phi$ and $\gamma$, and the model parameter, $\eta$, were approximated using the following EM procedure:

*E-step:* The lower bound for $KL(\gamma, \phi||\alpha, \eta)$ was maximized using functions from the exponential family. This led to the following updates of the variational parameters $\gamma_g$, $\gamma_h$, and $\gamma_l$ for latent classes $g$, $h$ and $l$, respectively:

$$\gamma_g^{(t+1)} = \alpha_g + \sum_{i=1}^{N} \varphi_{ig}^{(t)} \tag{8}$$

$$\gamma_h^{(t+1)} = \alpha_h + \sum_{j=1}^{N} \varphi_{jh}^{(t)} \tag{9}$$

$$\gamma_l^{(t+1)} = \alpha_l + \sum_{k=1}^{N} \varphi_{kl}^{(t)} \tag{10}$$

where the superscript $(t+1)$ denotes the new estimate and $(t)$ denotes the previous estimate. Note that the parameters $\gamma$ can be thought of as pseudo-counts for the latent classes.

The updates of the variational parameter pairs $(\phi_{ig}, \phi_{jh})$ and $(\phi_{jh}, \phi_{kl})$ are:

$$\begin{aligned} \varphi_{ig}^{(t+1)} &\propto \exp\left\{\psi\left(\gamma_{ig}^{(t+1)}\right) - \psi\left(\sum_{g=1}^{K} \gamma_{ig}^{(t+1)}\right)\right\} \prod_{h=1}^{K} (\eta_{gh}^{(t)})^{n_{gh}} \\ \varphi_{jh}^{(t+1)} &\propto \exp\left\{\psi\left(\gamma_{jh}^{(t+1)}\right) - \psi\left(\sum_{h=1}^{K} \gamma_{jh}^{(t+1)}\right)\right\} \prod_{g=1}^{K} \left(\eta_{hg}^{(t)}\right)^{n_{hg}} \end{aligned} \tag{11}$$

$$\begin{aligned} \varphi_{jh}^{(t+1)} &\propto \exp\left\{\psi\left(\gamma_{jh}^{(t+1)}\right) - \psi\left(\sum_{h=1}^{K} \gamma_{jh}^{(t+1)}\right)\right\} \prod_{l=1}^{K} \left(\eta_{hl}^{(t)}\right)^{n_{hl}} \\ \varphi_{kl}^{(t+1)} &\propto \exp\left\{\psi\left(\gamma_{kl}^{(t+1)}\right) - \psi\left(\sum_{l=1}^{K} \gamma_{kl}^{(t+1)}\right)\right\} \prod_{h=1}^{K} \left(\eta_{lh}^{(t)}\right)^{n_{lh}} \end{aligned} \tag{12}$$

where $\psi$, known as the digamma function, is the first derivative of the log $\Gamma$ function. $n_{gh}$ is the number of customers in latent class $g$ when the same customers are also in latent class $h$, and vice versa for $n_{hg}$. Likewise, $n_{hl}$ is the number of customers in latent class $h$ when the same customers are in latent class $l$, and vice versa for $n_{lh}$. The vectors $\varphi_{ig}^{(t+1)}$, $\varphi_{jh}^{(t+1)}$, and $\varphi_{kl}^{(t+1)}$ are normalized to sum to 1 for each customer $i$, $j$, and $k$.

*M-step*: The approximate maximum likelihood estimates for the matrix parameters $\eta$ for latent-class pairs $(g, h)$ and $(h, l)$ are:

$$\eta_{gh}^{(t+1)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \delta(i-j)\varphi_{ig}^{(t+1)}\varphi_{jh}^{(t+1)}}{\sum_{i=1}^{N}\sum_{j=1}^{N} \varphi_{ig}^{(t+1)}\varphi_{jh}^{(t+1)}} \tag{13}$$

$$\eta_{hl}^{(t+1)} = \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} \delta(j-h)\varphi_{jh}^{(t+1)}\varphi_{kl}^{(t+1)}}{\sum_{j=1}^{N}\sum_{k=1}^{N} \varphi_{jh}^{(t+1)}\varphi_{kl}^{(t+1)}}, \tag{14}$$

where $\delta(i-j) = 1$ when customer $i$ and customer $j$ are the same; otherwise, $\delta(i-j) = 0$. $\delta(j-k) = 1$ if customer $j$ and customer $k$ are the same; otherwise, $\delta(j-k) = 0$. The customers belonging to the latent-class pairs $(g, h)$ are the same as those belonging to the latent-class pairs $(h, l)$. We define them as cluster $(g, h)$ and cluster $(h, l)$, respectively.

We generated the variational estimates by iterating the E-step and the M-step until there were no customer differences in any of the latent classes.

## 5. Customer segmentation

In soft clustering, each customer has a mixed-membership score associated with each latent class. We assume that customers are assigned to multiple latent classes if their membership scores for these classes are sufficiently similar. The elements of the matrices $\eta_{gh}^{(12)}$ or $\eta_{hl}^{(23)}$ yield the interactions between the latent classes. As implied by Eqs. (13) and (14), the larger the mixed-membership score, the more the associated latent classes share the same customers. During the iteration process, the interactions between the latent classes for *Satisfaction with Service* and *Internet Usage* became linked to the latent classes for *Shopping Behavior*. We used *Shopping Behavior* as the pivotal category for deriving the following formula for customer segmentation.

Let $g, h, l = \{1, \ldots, K\}$ represent the latent classes for the categories. For each latent class $h$ for Shopping Behavior, the element $(g, h)$ of the matrix $\eta_{gh}^{(12)}$ is multiplied by the element $(h, l)$ of the matrix $\eta_{hl}^{(23)}$ to generate an interaction term for the component $\eta_{gh}^{(12)} * \eta_{hl}^{(23)}$. The customers belonging to this component are represented by a combination of clusters $(g, h)$ and $(h, l)$. The components of $\eta^{(12)}_{*h} * \eta^{(23)}_{h*}$ are defined as the segment (∗-*h*-∗), where ∗ denotes the *g* or *l* latent class. For instance, the element (4, 2) of the matrix $\eta^{(12)}_{gh}$ is multiplied by the element (2, 3) of the matrix $eta^{(23)}_{hl}$ to create an interaction score for the component (4-2-3) of segment (∗-2-∗). A high interaction score represents more in common in terms of the behaviors sampled among the customers assigned to the component. Within the segment (∗-*h*-∗), we sorted the components according to the interaction scores. The segment (∗-*h*-∗) was further partitioned into micro-segments based on the sorted list of components. Components were discarded from the segment (∗-*h*-∗) if their interaction scores were significant low compared to other components.

## 6. Experimental results

In the first experiment, we demonstrated empirically that our variational approximation is KL-convergent. Using a soft-clustering approach for customer segmentation, we compared the soft-clustered customer distributions with several latent classes. The mixed-memberships of these latent classes were evaluated and compared using several sets of membership-score-percentage differences. We then compared our soft-clustering results with the results from an EM clustering algorithm implemented in Weka.

As noted previously, the data for the study were retrieved from an online shopping questionnaire stored in the Survey Research Data Archive. The data consisted of responses to 2329 questionnaires collected through an online survey in July 2006. The responders were presented with six product categories available through online retail: books, computers, electronics, software, entertainment, and clothing. The buying frequency (number of purchase decisions) and the money spent were counted from July 2005 to June 2006.

### 6.1. Convergence of the variational approximations

Initially, we set the number of latent classes at 5, the value chosen by the RFM model. The Dirichlet parameters $\alpha$ were uniformly set to 0.2 (or 1/5). This number is small, because we needed to avoid expectation values of zero in Eq. (6) for small dataset. The true and approximate expectation values were calculated from the log-likelihood probabilities represented by the first four terms on the right sides of inequality (6) and Eq. (7). We denote the true expectation as E1 and the approximate expectation as E2. The KL divergence ($E$) is defined as E1–E2.

First, we employed a hard-clustering approach by assigning each customer to one and only one latent class. The expectation values converged after just a few iterations, as shown in Fig. 3. Second, we explored a soft-clustering approach in which customers could have membership in more than one latent class. In this case, we defined a customer as having mixed-membership if the customer's other scores on the membership vector differed from the highest score on the vector by less than 5%. For instance, if a customer's membership vector had the values 0.27, 0.27, 0.06, 0.28, and 0.12, the customer was assigned to latent classes 1, 2, and 4. The expectations derived from the first 16 of the 32 iterations are shown in Fig. 4. The soft clustering required more iterations than the hard clustering to reach convergence. The KL divergence decreased monotonically for both hard and soft clustering.

### 6.2. Customer segmentation using soft clustering

We decided to continue our investigation of customer segmentation by using a soft-clustering approach. We again set the number of latent classes for each category to 5; that is, 5 segments were created. There were 25 (5 × 1 × 5) components per segment, all selected on the basis of the interaction scores. We intuitively divided each segment into two micro-segments using the mean value of the components of the segment as the cutoff. Alternatively, other statistical criteria (e.g., the results of a *t*-test) can be applied to the micro-segmentation. As one of our segments had only 61 customers, we did not partition this segment further. This left a final sample of 2329 customers partitioned into nine micro-segments.

Table 2 summarizes these final customer distributions, giving the means and standard deviations for buying frequency and money spent by all the customers in each micro-segment. The results of the online customer satisfaction survey are provided in Table 3. The customers in micro-segments 1 and 2 are both frequent shoppers and high spenders. However, they differ in their habits of *Internet Usage* and *Satisfaction with Service* they receive when shopping online. The customers in micro-segments 3, 4, and 9 spend little money because they are unsatisfied with the services. We labeled the customers in micro-segments 5 and 6 as reluctant shoppers: they neither shop frequently nor spend a great deal of money. We labeled the customers in micro-segments 7 and 8 as efficient shoppers: they do not shop frequently, but when they do, they spend a lot. To characterize the typical customer in each micro-segment, we used data from the *Demographics* category to calculate the corresponding modes for age, gender, income, and marital status. These data are presented in Table 4.

The results indicate that the customers in micro-segment 1 have used the Internet more than 18 h per week for at least three years. They are satisfied with the company's online shopping services and expect to use the Internet more often in the future.

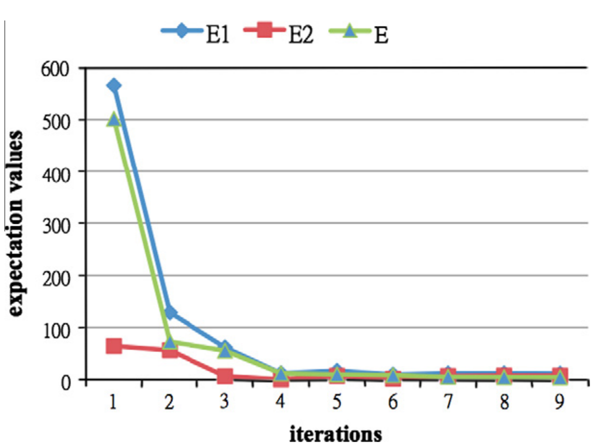


**Fig. 3.** Iterations of the expectation values for hard clustering with five latent classes.

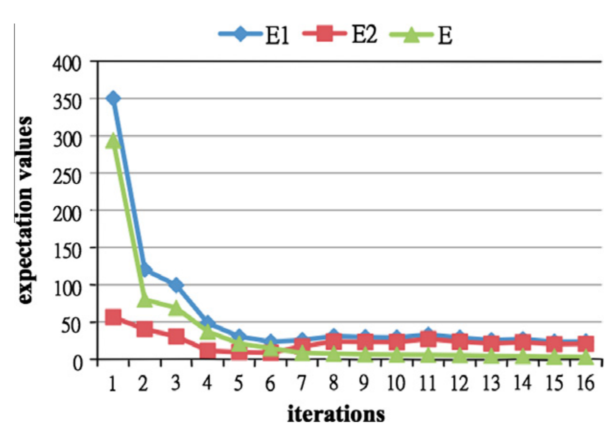


**Fig. 4.** Iterations of the expectation values for soft clustering with five latent classes.

**Table 2**
Micro-segments for shopping behavior (buying frequency and money spent).

| Micro-segments | Number of customers | Frequency (1–5) | | Money spent (1–9) | |
|---|---|---|---|---|---|
| | | Mean | s.d. | Mean | s.d. |
| 1 | 155 | 3.9290 | 0.9125 | 5.1097 | 1.7602 |
| 2 | 105 | 3.7905 | 0.9167 | 5.0571 | 1.6102 |
| 3 | 743 | 1.2283 | 0.1032 | 1.4803 | 0.1682 |
| 4 | 254 | 1.0264 | 0.0185 | 1.3462 | 0.0353 |
| 5 | 417 | 2.1799 | 0.4372 | 2.9161 | 0.7937 |
| 6 | 152 | 2.2105 | 0.4693 | 2.7961 | 0.8484 |
| 7 | 480 | 1.4012 | 0.2867 | 2.5271 | 0.8470 |
| 8 | 180 | 1.2685 | 0.2016 | 2.6389 | 1.3894 |
| 9 | 61 | 2.4590 | 0.7205 | 1.0228 | 0.0164 |

Customers in micro-segment 2 buy with similar frequency and spend as much money as the customers in micro-segment 1. However, the micro-segment-2 customers have used the Internet for only two to three years. They spend 15 h per week browsing the Internet, and most are not entirely satisfied with the services the company provides them.

The customers in micro-segment 3 have one to three years experience using the Internet and browse the Internet often. The customers in micro-segment 4 have one to two years experience using the Internet, browsing it occasionally. The customers in micro-segments 3 and 4 do not shop online frequently. Customers in segment 5 have more than three years experience using the Internet. They spend 16 h per week browsing and expect to use the Internet more often in the future. However, they do not shop on the Internet often or spend much money there. These customers are dissatisfied with the services. Although the customers in micro-segment 6 have little experience with the Internet, they expect to use it more often in the future. They are satisfied with the online shopping services.

The customers in micro-segment 7 have two to three years experience using the Internet but are not frequent buyers. They are very satisfied with the online shopping services. The customers in micro-segment 8 have one to two years experience using the Internet. They like to browse the Internet, but they do not buy frequently. The customers in micro-segment 9 are frequent shoppers but do not spend a great deal of money on the Internet. They have one to two years experience using the Internet and spend a moderate amount of time browsing. They are satisfied with the services and expect to use the Internet more often in the future.

### 6.3. Comparing the customer distributions for hard clustering and soft clustering

We chose *Shopping Behavior* to compare the results of hard clustering and soft clustering of the latent-class distributions. The numbers of latent classes were set at 5, 10, 20, and 30. We used the larger numbers of latent classes so we could observe the behavior of distributions in which there were customers in all the cells (no empty cells). Table 5 shows the number of iterations and non-empty latent-class cells for the different numbers of latent classes. Despite our efforts, some of the distributions with large numbers of classes had empty cells, regardless of whether hard clustering or soft clustering was used. However, the number of non-empty cells stabilized when the number of classes reached a certain point.

The number of non-empty cells is a function not only of the number of cases but also of how similar the cases are to one another during the segmentation process. How many customers there are in each cell depends on the number of categories and items in the questionnaire, and there is an upper limit to the number of these classes. The $N \times K$ dimensions of the latent variable matrices $z$, and the $K \times K$ dimensions of the matrices $\eta$, do not grow exponentially as the number of customers ($N$) increases. This is why the soft clustering method can effectively scale very large numbers of customers.

### 6.4. The mixed-membership distributions

Using soft clustering, we initially defined a customer as having mixed-membership if the customer's other scores on the membership vector differed from the highest score on the vector by less than 5%. We then repeated the soft clustering using a 10% difference and then a 15% difference. The results, presented in Table 6, show that the number of micro-segments was shrunk when the percentage cutoff was increased to 10%. We observed that most of the customers in micro-segments 1 and 2 were moved into micro-segments 5 or 6. Some customers in micro-segments 3 and 4 were moved into micro-segment 9, and a few customers in micro-segments 9 and 7 were moved into segment 3. Only two latent classes were generated with the 15% difference. In short, setting a high percentage cutoff increases the number of mixed-membership customers.

The probability of assigning a customer to a given membership class is derived by calculating the distance of each customer from the class center. Customers with similar distances to two class centers are assigned to both classes. The larger the percentage cutoff, the greater the probability that the customer will be assigned to multiple classes. Because this process creates many mixed-membership customers, the classes can merge due to many common members. There is no theoretical basis for deciding what the percentage cutoff should be. In practice, the choice should be based on how much weight a company places on a large number of mixed-membership customers for business development purposes.

### 6.5. A comparison of soft clustering and EM clustering using weka

For comparison purposes, we explored the expectation–maximization (EM) clustering algorithm, implemented in Weka, to cluster the customer data. This algorithm can be applied to multiple items as long as the items can be assumed to be independent. Cus-

**Table 3**
Micro-segments for online customer satisfaction.

| Micro-segments | Diversity of products (1–6) | Quality of products (1–6) | Security of transactions (1–6) | After-sale service (1–6) | Delivery service (1–6) | Personalized service (1–6) | Pricing (1–6) |
|---|---|---|---|---|---|---|---|
| 1 | 2.3290 | 2.4452 | 2.6710 | 2.8710 | 2.2323 | 2.5032 | 2.8839 |
| 2 | 3.2762 | 3.6476 | 3.9429 | 4.0381 | 3.0381 | 3.6952 | 3.4857 |
| 3 | 1.9803 | 2.2105 | 2.2697 | 2.4145 | 1.9276 | 2.3289 | 2.6645 |
| 4 | 1.8543 | 2.1063 | 2.1850 | 2.1969 | 2.0512 | 2.0551 | 2.4606 |
| 5 | 2.8825 | 3.0072 | 3.6115 | 3.8345 | 2.9410 | 3.4556 | 3.1583 |
| 6 | 3.5451 | 3.8358 | 4.5572 | 4.6797 | 4.2625 | 4.3755 | 3.9892 |
| 7 | 3.2354 | 3.3896 | 4.1646 | 4.1750 | 3.03021 | 3.6875 | 3.3917 |
| 8 | 2.1778 | 2.2444 | 2.3944 | 2.7111 | 2.1833 | 2.3278 | 2.6889 |
| 9 | 2.5246 | 2.5902 | 2.8852 | 2.9508 | 2.5246 | 2.7705 | 2.8689 |

**Table 4**
Demographic characteristics of customers.

| Subcategory | Micro-segment | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Age | 21–25 | 26–30 | 16–20 | 31–35 | 26–30 | 31–35 | Over 51 | 21–25 | 21–25 |
| Gender | Female | Male | Male | Female | Female | Male | Male | Male | Male |
| Income ($) | 15,000–35,000 | 35,000–55,000 | Under 15,000 | 35,000–75,000 | 15,000–55,000 | 55,000–75,000 | 35,000–75,000 | Under 15,000 | Under 15,000 |
| Marital status | Single | Single | Single | Married | Single | Single | Married | Single | Single |

**Table 5**
Iterations and non-empty cells for different numbers of latent classes.

| | 5 Classes | | 10 Classes | | 20 Classes | | 30 Classes | |
|---|---|---|---|---|---|---|---|---|
| | Iterations | Non-empty cells | Iterations | Non-empty cells | Iterations | Non-empty cells | Iterations | Non-empty cells |
| Hard clustering | 10 | 5 | 24 | 7 | 26 | 6 | 27 | 7 |
| Soft clustering | 29 | 5 | 34 | 6 | 45 | 9 | 64 | 8 |

**Table 6**
Comparisons of customer distributions using hard clustering and soft clustering with three membership-score-percentage differences.

| Micro-segments | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Hard clustering | 145 | 75 | 737 | 248 | 407 | 107 | 434 | 176 | 0 | 2329 |
| 5% Difference | 155 | 105 | 743 | 254 | 417 | 152 | 480 | 180 | 61 | 2547 |
| 10% Difference | 0 | 0 | 804 | 186 | 494 | 235 | 430 | 180 | 990 | 3219 |
| 15% Difference | 0 | 0 | 0 | 0 | 1051 | 1278 | 0 | 0 | 0 | 2329 |

tomers are assigned to classes probabilistically. To compare the resulting mixed-membership distributions with the nine micro-segments that resulted from the soft clustering, we had the EM algorithm cluster the customers into nine latent classes under *Satisfaction with Service*, *Shopping Behavior*, and *Internet Usage*.

To determine which customers to assign to each of the nine clusters, each latent class under *Shopping Behavior* was linked with each latent class under *Satisfaction with Service* and *Internet Usage* to form a $9 \times 1 \times 9$ components. The customers in these 81 components formed the cluster.

Tables 7 and 8 show the mixed-membership customer distributions for the resulting nine micro-segments and nine clusters. The total membership counts for the nine micro-segments, which appear in the diagonals of the tables, are 2547 for the micro-segments and 2581 for the clusters. The upper triangular matrices of Tables 7 and 8 show the mixed-membership counts, and the lower triangular matrices show the relative closeness estimates described below.

The relative closeness of cluster $C_i$ to cluster $C_j$ is the absolute closeness of the two clusters normalized with respect to the closeness of the members within the cluster among themselves. More precisely, the relative closeness, RC($C_i$, $C_j$), is defined as

$$RC(C_i, C_j) = \frac{S\{C_i, C_j\}}{\frac{|C_i|}{|C_i|+|C_j|}} S\{C_i\} + \frac{|C_j|}{|C_i| + |C_j|} S\{C_j\}, \tag{15}$$

where $S\{C_i, C_j\}$ is the number of customers assigned to clusters $C_i$ and $C_j$ simultaneously, and $S\{C_i\}$ (or $S\{C_j\}$) is the number of customers that belong exclusively to one of these clusters. $|C_i|$ and $|C_j|$ are the cardinalities of clusters $C_i$ and $C_j$ respectively. $S\{C_i\}$ (or $S\{C_j\}$) = $|C_i| - S\{C_i, C_j\}$ or $|C_j| - S\{C_i, C_j\}$.

The larger the relative closeness value, the greater the similarity between the two clusters and the less the similarity within each cluster. The total relative closeness values are 0.5993 for the micro-segments and 0.8037 for the clusters. These values indicate that soft clustering yields greater cohesion within micro-segments

**Table 7**
Mixed-member customer distributions and relative closeness for nine micro-segments, five latency classes and 5% difference in membership scores.

| Micro-segments | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **155** | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.1067 | **105** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | **743** | 51 | 0 | 0 | 23 | 0 | 0 |
| 4 | 0 | 0 | 0.0901 | **254** | 0 | 0 | 12 | 8 | 0 |
| 5 | 0 | 0 | 0 | 0 | **417** | 21 | 9 | 19 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0.0646 | **152** | 15 | 4 | 0 |
| 7 | 0 | 0 | 0.0373 | 0.0308 | 0.0204 | 0.0388 | **480** | 43 | 0 |
| 8 | 0 | 0 | 0 | 0.0372 | 0.0275 | 0.0248 | 0.1211 | **180** | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **61** |

*Note*. The upper triangle gives the mixed-membership counts, the diagonal gives the total counts, and the lower triangle gives the relative closeness scores.

**Table 8**
Mixed-member customer distributions and relative closeness using Weka clusters.

| Clusters | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **153** | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.1267 | **109** | 0 | 0 | 2 | 4 | 0 | 0 | 0 |
| 3 | 0 | 0 | **725** | 62 | 0 | 0 | 19 | 2 | 3 |
| 4 | 0 | 0 | 0.1151 | **287** | 0 | 0 | 18 | 16 | 4 |
| 5 | 0 | 0.0055 | 0 | 0 | **427** | 30 | 0 | 9 | 0 |
| 6 | 0 | 0.0311 | 0 | 0 | 0.0923 | **150** | 19 | 16 | 0 |
| 7 | 0 | 0 | 0.0321 | 0.0520 | 0 | 0.0591 | **409** | 32 | 0 |
| 8 | 0 | 0 | 0.0033 | 0.0620 | 0.0254 | 0.0791 | 0.1004 | **258** | 0 |
| 9 | 0 | 0 | 0.0043 | 0.0153 | 0 | 0 | 0 | 0 | **63** |

*Note*. The upper triangle gives the mixed-membership counts, the diagonal gives the total counts, and the lower triangle gives the relative closeness scores.

than within clusters, because step 4 of the soft clustering procedure determines the interaction between the latent classes iteratively.

## 7. Managerial implications

Analyzing and understanding customer characteristics and buying behavior form the foundation that companies use to develop their CRM strategy. Customer satisfaction, which is referred to the ratio of customers' expectations of being satisfied to their subsequent perceptions of being satisfied, is the central concern for CRM. Many online surveys have focused on the demographic characteristics of online customers. Although demographic information is useful, it provides little diagnostic information about the customers. How online consumers perceive the web to be performing in their shopping context is also important to know for the purpose of delivering value to the customer and maximizing the customer's value to the company.

To get adequate diagnostic information, it is crucial to investigate multiple customer segments categorized on the basis of the customers' buying-behavior patterns. This approach allows marketing managers to develop insights of significant managerial interest. For instance, e-retailers can segment online customers by using data from their online shopping records and online surveys. In these surveys, online customers are asked to evaluate Internet vendors on their services and web-related attributes. The e-retailers require registration and ask their online customers to enter their demographic information. Customers may perceive that the web performs poorly with respect to the security of sensitive personal data, so Internet vendors should ensure that the security of their websites is of the highest quality. After identifying the different characteristics and needs of the different segments, Internet vendors can direct their efforts and allocate their resources in such a way as to attract the typical customer in each segment. If managers are to successfully use CRM to maximize the vendor's profits, it is necessary that they retain and increase the lifetime value of their most valuable customers.

The segments uncovered by our soft-clustering approach provide information about customers' buying behavior in multiple categories. What customers who belong to more than one membership segment find attractive in each of the segments they belong to should be of interest to marketing managers. Marketing managers consider information from a number of segments before making a decision. This deliberate approach identifies customers' characteristics and, thus, their responsiveness to the services in multiple segments. The degree to which a customer spans segments provides information about otherwise hidden buying-behavior patterns and thus provides useful input for marketers who are developing and implementing a marketing strategy. This strategy can be justified by the expected benefits of reaching customers and helping them to satisfy their unique needs and desires in the best possible way. Marketers should take suitable actions to offer these customers appropriate promotions and other incentives that will direct them to the particular channels the customer prefers.

## 8. Conclusion

Although acquiring customers is usually crucial to a company's success, understanding the customers' characteristics and retaining them as customers are more important for its financial success. We analyzed customer buying behavior and attitudes using data from an online shopping website that collects data about a number of factors related to this topic. We divided our customer questionnaire into three main sections or categories (items in parentheses): *Shopping Behavior* (frequency, money spent), *Satisfaction with Service* (product diversity, product quality, transaction security, after-sale service, delivery, personalized service, price), and *Internet Usage* (years of experience, hours per week, frequency of product browsing, expected use in the next year). A fourth category was *Demographics* (age, gender, income, marital status).

The main contribution of this paper is to offer readers a computationally feasible method for segmenting customers across multiple categories. The proposed method also allows scalability of very large numbers of customers. To analyze the data, we employed a mixed-membership model. First, we used soft clustering, through which customers can be assigned to more than one cluster. We found that soft clustering produces more promising results for real-world applications than hard clustering, through which each customer can be assigned to only one cluster. However, soft clustering requires more computational iterations than hard clustering. A variational approximation method was developed to efficiently estimate the parameters. By iteratively adjusting the latent-class structures for each category, our model achieved convergence. To calculate the expectation values, we first set the Dirichlet parameter $\alpha$. The larger this parameter value, the higher the true and approximate expectation values and the greater the number of required computational iterations. The expectation values eventually converge, regardless of how high $\alpha$ is set. For small $N$, the $E_q[\log p(\theta|\alpha)]$ sometimes equals zero; in these cases, $\alpha$ must be small. This means that the number of latent classes must be chosen carefully as a function of the sample size.

Two limitations of our soft clustering method are the difficulty in selecting the number of latent classes for implementation and choosing the cutoff percentage for the membership-score difference. The proposed clustering method automatically generates the optimal number of latent classes from the sample, and choosing a larger latent class number for implementation purposes results in computational inefficiency. The function used to measure the similarity of the objects (customers) to the latent classes affects the mixed-membership scores on the membership vector. The choice of cutoff percentage is influenced by which similarity function is chosen. The other factor that influences the selection of both these parameters is the sampling distribution, which is usually unknown prior to segmentation. Hence, both parameters should be determined empirically.

## References


Airoldi, E., Blei, D., Fienberg, S., and Xing, E. Mixed membership stochastic block models for relational data with application to protein–protein interactions. In *Proceedings of the International Conference on Biometrics Society Annual Meeting*, Tampa, FL, March 26–29, 2006.

Attias, H. A variational Bayesian framework for graphical models. In S. A. Solla, T. K. Leen, and K. R. Mueller (eds.). *Advances in Neural Information Processing Systems*, vol. 12, MIT Press, Cambridge, MA, 2000, 209–215.

Barnard, K., Duygulu, P., de Fraitas Forsyth, N., Blei, D., and Jordan, M. Matching words and pictures. *Journal of Machine Learning Research*, 3, 6, 2003, 1107–1135.

Bhatnagar, A., and Ghose, S. A latent class segmentation analysis of e-shoppers. *Journal of Business Research*, 57, 7, 2004, 758–767.

Blei, D., Ng, A., and Jordan, M. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 1, 2003, 993–1022.

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. K. Saul, and B. Scholkopf (eds.). *Advances in Neural Information Processing Systems*, vol. 16, MIT Press, Cambridge, MA, 2004, 17–24.

Bose, I., and Chen, X. Exploring business opportunities from mobile services data of customers: An inter-cluster analysis approach. *Electronic Commerce Research and Applications*, 9, 3, 2010, 197–208.

Bult, J., and Wansbeek, T. Optimal selection for direct mail. *Marketing Science*, 14, 1995, 378–394.

Chen, Y., Kuo, M., Wu, S., and Tang, K. Discovering recency, frequency, and monetary (RFM) sequential patterns from customer's purchasing data. *Electronic Commerce Research and Applications*, 8, 5, 2009, 241–251.

Cheng, C., and Chen, Y. Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems and Applications*, 36, 1, 2009, 4176–4184.

Erosheva, E. Partial membership models with application to disability survey data. In H. Bozdogan (ed.), *Statistical Data Mining and Knowledge Discovery*, Chapman and Hall/CRC, Boca Raton, FL, 2004, 117–134.

Erosheva, E., and Feinberg, S. Bayesian mixed membership models for soft clustering and classification. In *Classification: The Ubiquitous Challenge*, Springer-Verlag, Berlin, Germany, 2005, 11–26.

Heilman, C., and Bowman, D. Segmenting consumers using multiple-category purchase data. *International Journal of Research in Marketing*, 19, 1, 2002, 225–252.

Hughes, A. Boosting response with RFM. *Marketing Tools*, 3, 1996, 4–10.

Jiang, T., and Tuzhilin, A. Segmenting customers from population to individual: Does 1-to-1 keep your customers forever. *IEEE Transactions on Knowledge and Data Engineering*, 18, 10, 2006, 1297–1311.

Long, B., Zhang, Z., Wu, X., and Yu, P. Spectral clustering for multi-type relational data. In *Proceedings of International Conference on Machine Learning*, Pittsburgh, PA, June 2006, ACM Press, New York, NY, 2006.

Magidson, J., and Vermunt, J. Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research*, 20, 2002, 37–44.

McCarty, J., and Hastak, M. Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, 60, 6, 2007, 656–662.

Meier, A., and Werro, N. A fuzzy classification model for online customers. *Informatica*, 31, 2007, 175–182.

Migkautsch, J. Thoughts on RFM scoring. *Journal of Database Marketing*, 8, 1, 2000, 67–72.

Minka, T. Expectation–maximization as lower bound maximization. Technical Report, Massachusetts Institute of Technology, Cambridge, MA, 1998.

Minka, T. Using lower bounds to approximate integrals. Technical Report, Carnegie Mellon University, Pittsburgh, PA, 2001.

Ngai, E., Xie, L., and Chau, D. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems and Applications*, 36, 2, 2009, 2592–2602.

Nowicki, K., and Snijders, T. A. B. Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association*, 96, 455, 2001, 1077–1087.

Payne, A., and Frow, P. A strategic framework for customer relationship management. *Journal of Marketing*, 69, 4, 2005, 167–176.

Regielski, C., Wang, J. C., and Yen, D. C. Data mining techniques for customer relationship management. *Technology in Society*, 24, 4, 2002, 483–502.

Spangler, W., May, J., and Vargas, L. Choosing data-mining methods for multiple classification: Representational and performance measurement implications for decision support. *Journal of Management Information Systems*, 16, 1, 1999, 37–62.

Stone, B., and Jacobs, R. *Successful Direct Marketing Methods*, 7th edition. McGraw Hill, New York, NY, 2001.

Vellido, A., Lisboa, P., and Meehan, K. Segmentation of the on-line shopping market using neural networks. *Expert Systems and Applications*, 17, 4, 1999, 303–314.

Yang, M., and Yu, N. Estimation of parameters in latent class models using fuzzy clustering algorithms. *European Journal of Operational Research*, 160, 2, 2005, 515–531.

Zabah, A., Bellenger, D., and Johnston, W. An evaluation of divergent perspectives on customer relationship management: Towards a common understanding of an emerging phenomenon. *Industrial Marketing Management*, 33, 6, 2004, 475–489.