Cybernetics and Systems: An International Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucbs20

CUSTOMER SEGMENTATION AND CLASSIFICATION FROM BLOGS BY USING DATA MINING: AN EXAMPLE OF VOIP PHONE

Long-Sheng Chen [a] , Chun-Chin Hsu [b] & Mu-Chen Chen [c]

[a] Department of Information Management , Chaoyang University of Technology , Taiwan

[b] Department of Industrial Engineering and Management , Chaoyang University of Technology , Taiwan

[c] Institute of Traffic and Transportation, National Chiao Tung University , Taipei, Taiwan

Published online: 04 Sep 2009.

To cite this article: Long-Sheng Chen , Chun-Chin Hsu & Mu-Chen Chen (2009) CUSTOMER SEGMENTATION AND CLASSIFICATION FROM BLOGS BY USING DATA MINING: AN EXAMPLE OF VOIP PHONE, Cybernetics and Systems: An International Journal, 40:7, 608-632, DOI: 10.1080/01969720903152593

To link to this article: http://dx.doi.org/10.1080/01969720903152593

Cybernetics and Systems: An International Journal, 40: 608–632

ISSN: 0196-9722 print/1087-6553 online
DOI: 10.1080/01969720903152593

Taylor & Francis
Taylor & Francis Group

CUSTOMER SEGMENTATION AND CLASSIFICATION FROM BLOGS BY USING DATA MINING: AN EXAMPLE OF VOIP PHONE

LONG-SHENG CHEN[1], CHUN-CHIN HSU[2], and MU-CHEN CHEN[3]

[1]Department of Information Management, Chaoyang University of Technology, Taiwan
[2]Department of Industrial Engineering and Management, Chaoyang University of Technology, Taiwan
[3]Institute of Traffic and Transportation, National Chiao Tung University, Taipei, Taiwan

Blogs have been considered the 4th Internet application that can cause radical changes in the world, after e-mail, instant messaging, and Bulletin Board System (BBS). Many Internet users rely heavily on them to express their emotions and personal comments on whatever topics interest them. Nowadays, blogs have become the popular media and could be viewed as new marketing channels. Depending on the blog search engine, Technorati, we tracked about 94 million blogs in August 2007. It also reported that a whole new blog is created every 7.4 seconds and 275,000 blogs are updated daily. These figures can be used to illustrate the reason why more and more companies attempt to discover useful knowledge from this vast number of blogs for business purposes. Therefore, blog mining could be a new trend of web mining. The major objective of this study is to present a structure that includes unsupervised (self-organizing map) and supervised learning methods (back-propagation neural networks,

This work was supported in part by National Science Council of Taiwan (Grant No. NSC 96-2416-H-324-003-MY2).

Address correspondence to Long-Sheng Chen, Department of Information Management, Chaoyang University of Technology, 168 Jifong East Road, Wufong Township, Taichung County, 41349 Taiwan. E-mail: lschen@cyut.edu.tw

decision tree, and support vector machines) for extracting knowledge from blogs, namely, a blog mining (BM) model. Moreover, a real case regarding VoIP (Voice over Internet Protocol) phone products is provided to demonstrate the effectiveness of the proposed method.

Keywords: Back-propagation neural network; Blog; Data mining; Self-organizing map; Sparse data; Support vector machines

INTRODUCTION

In recent decades, electronic communication has changed to be more convenient and ubiquitous. Once e-mail had been a staple, but as the Internet grew there has been FTP (file transfer protocol), (BBS) electronic bulletin board system, web pages, secured servers, and now the world of wikis, blogs (i.e., weblogs) (Wood-Black and Pasquarelli 2007) and YouTube. Among them, blogs are one of the fastest growing sections of the Internet and are emerging as an important communication mechanism that is used by an increasing number of people (Cohen and Krishnamurthy 2006). Blogs have been regarded as the fourth Internet application that can cause radical changes in the world, after e-mail, instant messaging, and BBS. A blog can be considered as a journal which can continuously allow the users to update their own words and post their work online through software. Bloggers (blogs users) often make a record of their lives and express their opinions, feelings, and emotions through writing blogs (Nardi, Schiano, Gumbrecht, and Swartz 2004). One of the most important features in a blog is the ability for any reader to write a comment on a blog entry. This ability has facilitated the interaction between bloggers and their readers (Chau and Xu 2007). In blogs, many cybercommunities have also emerged (Kumar, Raghavan, Rajagopalan, and Tomkins 1999) and become a new way to discuss specific issues, such as infectious diseases (Larkin 2005), cancer (Oransky 2005), online campaigning (Trammell 2006), research topics (Todoroki, Konishi, and Inoue 2006), online jewelry selling (Hof 2005), tourism promotion (Lin and Huang 2006), and so on.

Blogs can be used in many different ways. One can add multimedia files to improve a blog's attractiveness. What's more, a blog is interactive in the sense that readers can respond with comments in just a few steps. Online forums can also take place on any blog site. For example, the well-known health website WebMD (www.webmd.com) allows site

visitors a large variety of blogs involving topics such as pregnancy, asthma, sleep disorders, and so on to choose from. On a monthly basis, 30 million individual users come to WebMD to make better decisions for themselves or for their loved ones (Mathieu 2007). The political blog talkingpointsmemo.com claims to have more than 3000 unique readers in a month (Cohen and Krishnamurthy 2006). Besides personal blogs, there are also blogs created by companies. For instance, ice.com, an online jewelry seller, has launched three blogs and reported that thousands of people linked to their website from these blogs (Hof 2005). Consequently, blogs are now treated as new marketing channels.

Popularization of blogs means that there is more information out there that can be very useful (Wood-Black and Pasquarelli 2007). According to a survey of the blog search engine, Technorati, approximately 94 million blogs were tracked in August 2007 (on 60 million blogs were searched in January 2007). It also reported that 11% of Internet users visit blogs frequently. Additionally, a whole new blog is created every 7.4 seconds and 275,000 blogs are updated daily. These figures can be used to demonstrate the prosperity of blogs. Most importantly, they motivate lots of commercial or academic activities in blogs. Rosenbloom (2004) indicates that blogs are becoming a new form of mainstream personal communication for millions of people to publish and exchange knowledge/information, and to establish networks or build relationships in the world of all blogs, so-called the "blogsphere." E-business models may have radical changes because of the explosion of bloggers. Therefore, enterprises need to truly understand the power of blogs and deliberate how to utilize blogs as commercial tools to create a profit margin. How to discover useful knowledge from the huge amount of bloggers' comments has become an important business issue from which enterprises can potentially benefit.

Following this trend, researchers have paid more and more attention in studying some issues regarding blogs. For examples, Todoroki et al. (2006) propose to utilize a blog as an electronic research notebook. Chau and Xu (2007) present a semi-automated approach to facilitate the monitoring, study, and research on blogs of online hate groups. Lin and Huang (2006) indicate that blogs can significantly influence browsers and indirectly promote tourism. Du and Wagner (2006) seek to explore blogs' success factors from a technology perspective. Asano (2007) investigated whether a "fiction novel" on blogs describing a girl undergoing epilepsy surgery can potentially facilitate familiarity to epilepsy surgery

among the general Internet users in Japan. In addition, while blogs are not yet a standard public relations tool, practitioners use blogs to enhance their power within their organizations (Porter, Trammell, Chung, and Kim 2007). This available literature has reported that blogs have already influenced human behavior and has become a new channel of exchanging information. However, most researchers focus on how bloggers behave and what the blogs' influence is. Relatively few articles discuss extracting knowledge from blogs, such as usage mining (Facca and Lanzi 2005), structure mining (Chau and Xu 2007), and blogs content mining.

To summarize, more and more companies attempt to discover what bloggers' true thinking is for business purposes. To achieve this goal, data mining from blogs could be considered a new trend. By using data mining techniques, enterprises can profile customers who have responded to previous similar campaigns and these profiles are useful in finding the best customer segments that the company should target (Olafsson, Li, and Wu 2008). For the existing customers, data mining can also be used to predict churn. Therefore, the major objective of this study is to present a structure that includes unsupervised (self-organizing map (SOM)) and supervised learning methods (back-propagation neural networks (BPN) decision tree, C4.5; support vector machines (SVM)) to extract knowledge from bloggers' comments, namely, a blog mining (BM) model, for customer segmentation and customer classification. Finally, a real case regarding Voice over Internet Protocol (VoIP) phone products is provided to illustrate the superiority of our proposed methodology.

RELATED WORKS

Blog Mining

The World Wide Web (WWW) is a tremendously rich knowledge base. The knowledge comes not only from the contents of the web pages, but also from the webs' hyperlink structure and its diversity of content and languages. Analyzing these characteristics often can find interesting patterns and knowledge that can be employed to support decision-making or business management (Chen and Chau 2004; Borzemski 2006). When interacting with the web information, users could encounter some problems such as 1) finding relevant information, 2) creating new knowledge out of the information available on the web, 3) personalization of the information, and 4) learning about consumers or individual users (Kosala and Blockeel 2000). These problems can

be tackled by web mining techniques that are important because the web has provided a vast amount of publicly accessible information which could be useful in security applications (Chau and Xu 2007). It has reported that many terrorists and extremist groups have been using the web for various purposes (Zhou, Reid, Qin, Chen, and Lai 2005; Qin, Zhou, Reid, Lai, and Chen 2006). According to the available works (Kosala and Blockeel 2000; Wang and Liu 2003; Lappas 2007), web mining techniques can be divided into three groups: content mining, structure mining, and usage mining.

Web content mining refers to the discovery of useful information from web contents, including document categorization and clustering, and information extraction from web pages. The extracted key information, such as distinctive menu items, navigation indicators, which is embedded in web pages, can help classify the main contents of web pages and reflect certain taxonomy knowledge (Wang, Lu, and Zhang 2007; HaCohen-Kerner, Stern, Korkus, and Fredj 2007). The purpose of web structure mining, which involves analysis of in-links and out-links of a web page, is to study the web's hyperlink structure. This graph structure can provide information about a page's ranking or authoritativeness and enhance search results through filtering (Kolari and Joshi 2004). Web usage mining, whose main application is learning user profiles, focuses on analyzing the results of user interactions with a web server such as logs, click-streams, and database transactions at a website for finding relevant patterns (Wang and Liu 2003). Usage pattern extracted from web data can be applied to a wide range of applications such as web personalization, system improvement, site modification, business intelligence discovery, usage characterization, and so on (Cho, Kim, and Kim 2002). A survey of web usage mining can be found in the work of Facca and Lanzi (2005).

Blog mining belongs to one of the areas of web mining. However, blog mining could be a new trend due to the explosive growth of blogs and numerous potential applications. Before discussing blog mining, we should clarify the differences between blogs and web pages. According to the work of Cohen and Krishnamurthy (2006), the differences can be summarized as follows:

1. A blog is often a single page site.
2. Blogs are often personal journals or discussion groups on a narrow topic.

3. Active blogs are updated with a frequency significantly higher than a traditional web page. In practice, blogs have turned out to be writings about a variety of topics, typically updated on a more regular basis than home pages.
4. Traditional websites are designed as a coherent view of a subject, where older links may be as relevant as new links. Blogs, in contrast, are modified regularly with new content added while older content is archived away. In blogs, all contents on the main page are new and expected to be relevant. Furthermore, unlike a moderated site, additions to a blog are immediately available to anyone accessing the URL of the blog.

From the above-mentioned, although blogs own their unique characteristics different from web pages, a blog basically can be viewed as a special long web page. This study aims at developing a knowledge acquisition mechanism for discovering useful information and generating desired knowledge from a large amount of bloggers' comments. Therefore, the proposed BM model can be considered as one of web content mining approaches.

Customer Segmentation and Classification

Marketing focuses on the establishment, development, and maintenance of lengthy relationships between buyer and seller. Marketing managers should be able to understand the factors that explain the establishment of continuous relationships to manage their customer portfolio in an effective way. In the existing highly competitive market, the literature (Gil-Saura and Ruiz-Molina 2009) recommends customer segmentation. Customer segmentation refers to the process of dividing similar customers into groups. Companies naturally segment customers to better serve their needs for increasing the value of customers (Olafsson et al. 2008). How the company segments its customers can have an important impact on company profitability (Epstein, Friedl, and Yuthas 2009). The practical benefits of customer segmentation include focusing on customers' needs, building relationships with the most attractive customers, creating barriers for competitors, delivering focused product and service propositions, increasing revenues, determining who not to chase for business, and prioritizing resource allocation and marketing spent on the most worthwhile opportunities (Simkin 2008).

Today, successful businesses must aggressively attack target markets and niche segments. Therefore, how to identify customer segments is very crucial for survival. Customer classification means to classify unknown or new customers into established segments. Based on established segments, the company can identify customer behavior patterns from customer usage data and predict which customers are likely to respond to cross-sell and up-sell campaigns, which are very important to the business (Olafsson et al. 2008). This prediction of new customers enables an organization to selectively target "good" business. To achieve these goals mentioned above, Data mining techniques such as self-organizing map (SOM), decision trees (C4.5), and support vector machines (SVM) are usually employed to do customer segmentation and classification (Wu et al. 2008; Olafsson et al. 2008).

PROPOSED BLOG MINING METHODLOGY

Proposed Blog Mining Procedure

This section describes the proposed BM methodology. The proposed procedure shown in Figure 1 displays knowledge from blog contents, i.e., bloggers' comments. The BM procedure can be listed as follows:

Step 1: Definition of variables
    -Predefine keywords
    -Remove meaningless words
    -Clarify polysemy and synonym problem
Step 2: Data collection
    -Collect bloggers' comments
Step 3: Data preprocessing
    -Data cleaning
    -Data normalization
    -Correlation analysis
    -Frequency analysis
Step 4: Unsupervised learning (SOM)
    -Sparsity analysis
Step 5: Supervised learning Back-Propagation Neural Network (BPN), C4.5, and (SVM)

In step 1, blog mining users need to define a set of keywords to be the variables (attributes) of data. In step 2, every single personal comment in blogs has been considered as an instance. Those predefined

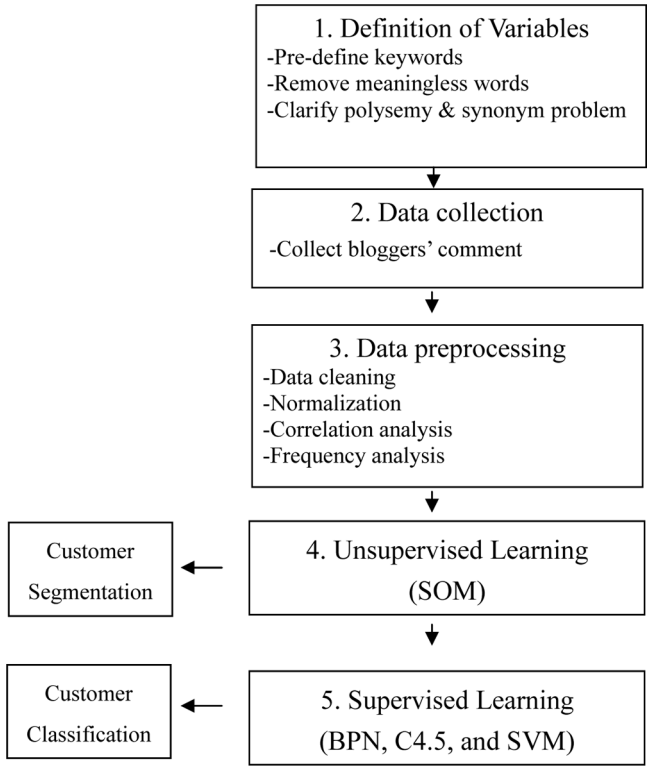

Figure 1. The proposed blog mining procedure.

attributes are used to describe the collected instances. In step 3, some data preprocess techniques such as data cleaning, data normalization, and frequency analysis are applied to the collected data. Steps 4 and 5 are knowledge acquisition phases. In step 4, because the collected data has no class labels, an unsupervised learning tool, SOM, is employed to analyze data. A suitable number of clusters has to be determined by sparsity analysis in this step. Then, different customer segmentations can be found. Depending on customers' preferences, we can design several marketing strategies for each specific type of customer. In step 5, the cluster label defined in step 4 has been viewed as a class label for classification. Therefore, we can acquire knowledge from those instances and use them to predict which segmentation the unknown customer belongs to. Some concise illustrations regarding the proposed procedure are provided in next subsections.

Definition of Variables

The first step of mining knowledge from the blog is to define suitable variables for describing every comment in blogs. There are three jobs needed to be done in this step. Firstly, blog mining users can collect a small size of samples to find a set of keywords that are mentioned in those comments. Then, this step combines these found keywords with which users are interested to construct the predefined variable set. Secondly, we remove some meaningless keywords such as "is," "are," "will," "have," "of," and so on. Thirdly, the polysemy and synonym problems should be considered. We may integrate several words that have exactly the same meaning into a new variable. Besides, many blogger-created netspeaks, including funny words or creative icons, frequently appear in bloggers' comments, such as ":)," ":-)," or ": D," which represent a smile face. Others like "Orz," "OTZ," and "or2" are also typical netspeaks examples. We should carefully clarify them by continuing to understand Internet users' languages.

Data Collection and Preprocessing

In the data collection step, we use the predefined variable set in step 1 to describe the collected data. Every single blogger's comment related to a specific topic is viewed as one instance. Next, if we have any missing value instances, they should be removed in the step of data preprocessing. We also remove some keywords with an extremely low frequency of occurrence. Besides, data normalization is necessary because we utilize Euclidean distance-based tools, neural networks (SOM and BPN) to be knowledge acquisition tools. We need to normalize all attribute values into the same scale. Hence, all values of attributes are normalized to the interval [0, 1] by employing a min-max normalization equation, as expressed by Equation (1). In this equation, $\max_i$ is the maximum and $\min_i$ is the minimum of the ith attribute values, and v_{ij} is the value of the ith attribute of the jth object and v'_{ij} is the normalized value. Besides, we can visualize data by performing the frequency and correlation analysis before implementing clustering and classification

$$v'_{ij} = \frac{v_{ij} - \min_i}{\max_i - \min_i}. \quad (1)$$

Unsupervised Learning

In this subsection, we implement clustering analysis for customer segmentation. The collected instances have no class labels. Therefore, we attempt to analyze them by using a widely used unsupervised tool, SOM. This step primarily finds the customer segmentations and knows what their tastes and interests are. Then we can focus on the discovered knowledge and develop unique marketing strategies for each specific type of customer.

Among numerous clustering algorithms, one neural network model of particular interest is Kohonen's (1990) SOM. The SOM and its variants (Kukolj et al. 2006; Salas, Moreno, Allende, and Moraga 2007) have been very successful in several real application areas (Seiffert and Jain 2002), such as retrieval, clustering, visualization, and classification (Rahman, Pi, Wang, Tommy, and Wu 2007). The SOM is a special kind of artificial neural network with unsupervised learning. The SOM model preserves the topology mapping from the high-dimensional input space onto a low-dimensional display. The original application area is speech recognition and it's a clustering method that can be used to cluster a set of samples $X = (X1, X2, \ldots, Xn)$ into p clusters; and the obtained clusters are arranged on a grid. The SOM can gain insight into the structure of the dataset and observe the relationships between the patterns being originally located in a high dimensional space. Usually, SOM is a fully connected network with two layers whose second layer is usually organized as a 2D grid as shown in Figure 2. The weight vector, W_{np}, for a cluster unit serves as an exemplar of the input patterns associated with that cluster. During the self-organizing process, the cluster unit, whose

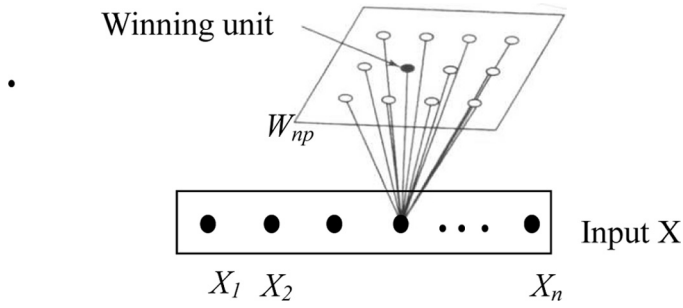

Figure 2. A basic topology of the self-organizing map.

weight vector matches the input pattern most closely, is chosen as the winner. Then, the winning unit and its neighboring units update their weights. We briefly illustrate the operation of SOM by the following steps:

Step 1: Compute a matching value $\|X_n - W_{np}\|$ for each unit in the competitive layer.
Step 2: Find the best match.
Step 3: Identify the neighborhood around the winning unit.
Step 4: Weights are updated for all units that are in the neighborhood of the winning unit.

In SOM, two parameters, learning rate α and radius of the neighborhood R, must be specified. The learning rate is a slowly decreasing function of time (or training epochs). Kohonen (1990) indicates that a linearly decreasing function is satisfactory for practical computations; a geometric decrease would produce similar results. The radius (R) of the neighborhood around a cluster unit also decreases as the clustering process progresses. If the size of a neighbor is reduced to a winning unit only, the SOM essentially becomes a vector quantization network.

After clustering instances, several groups of customers can be found by sparsity analysis and we can understand what their characteristics are. In the next step, the customer segments are considered as class labels. Then, we can use supervised learning tools to extract knowledge.

Supervised Learning

The main purposes of this subsection are not only to extract knowledge for supporting decision-making, but also to build a classifier to predict the type of customers' segmentations for future unknown customers (i.e., customer classification). Three famous supervised learning algorithms, BPN, C4.5, and SVM, are employed in this study.

Back-Propagation Neural Network (BPN). The BPN has been widely used in pattern recognition, function approximation, optimization, and classification. Generally speaking, neural nets can be classified into two categories—feed-forward and feedback networks. In this study, the feed-forward network as shown in Figure 3 was employed because of their superior ability of classification.

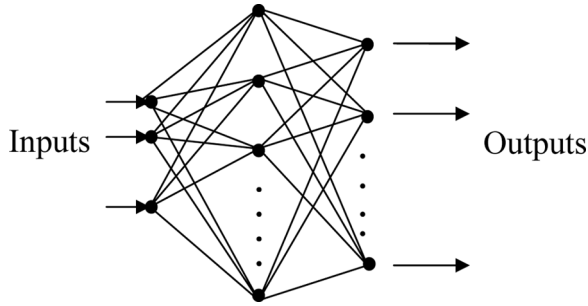

Figure 3. The back-propagation neural network structure.

The back-propagation learning algorithm (Rumelhart and McClelland 1986) is the best-known training algorithm for neural networks and is still one of the most useful. This iterative gradient algorithm is designed to minimize the mean-squared error between the actual output of a multilayer feed-forward perceptron and the desired output. According to the rule of thumb and reports of available published articles, the number of hidden layers should be one or two. The back-propagation algorithm includes a forward pass and a backward pass. The forward pass obtains the activation value, and the backward pass adjusts weights and biases according to the difference between the generated and actual network outputs. These two passes will go through iteratively until the network converges. More detailed information about network training by back-propagation can be found in related references (e.g., Philip (1992) and Freeman and Skapura (1992)).

Decision Trees. Decision tree is one of the most popular knowledge acquisition algorithms and has been successfully applied in many areas. Decision tree algorithms, such as ID3 (Quinlan 1986) and C4.5 (Quinlan 1993), were originally designed for classification purposes. The core of C4.5 contains recursive partitioning of the training examples. Whenever a node is added to a tree, some subsets of the input features are used to pick the logical test at that node. The feature that results in the maximum information gain is selected for testing at that node. In other words, the algorithm chooses the "best" attribute to partition the data into individual classes at each node. After the test has been determined, it is used to partition the

examples, and the process proceeds recursively until each subset contains examples of one class or satisfies some statistical criteria (Su and Shiue 2003).

Support Vector Machines. Support vector machines (Vapnik, Golowich, and Smola 1996) has yielded some of the best accuracy to date. Support vector machines seek to find a separator surface between classes such that a band of maximum possible thickness of the region (also called the margin) around the separator is empty, i.e., no training points in the margin. This leads to better generalization (Chakrabarti, 2000; Lu, Lin, and Ying 2007). Generally speaking, SVM learns a decision boundary between two classes by mapping the training data $\mathbf{x}_i \in R^d$ (through kernel functions ϕ) onto a higher dimensional space, and then finding the maximal margin hyperplane within that space. Finally, this hyperplane can be viewed as a classifier. Figure 4 illustrates the concept of feature mapping and two-class separation.

Considering a classifier, which uses a hyperplane to separate the two-class patterns based on given examples, $S = \{x_i, y_i\}_{i=1}^{n}, y_i \in \{-1, +1\}$, x_i is the attribute set and y_i is the class label. The hyperplane is defined by (w, b), where w is a weight vector and b a bias. Let w_0 and b_0 denote the optimal values of the weight vector and bias. Correspondingly, the optimal hyperplane can be written as:

$$w_0^T x + b_0 = 0. \tag{2}$$

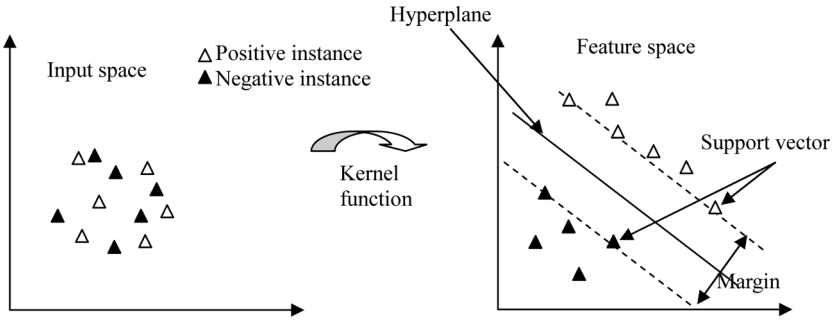

Figure 4. The operations of support vector machine (Cristianini and Shawe-Taylor 2000).

To find the optimum values of w and b, it requires solving the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \\ & \mathbf{y_i}(\mathbf{w^T}\phi(\mathbf{x_i}) + \mathbf{b}) \geq 1 - \xi_i \\ \text{Subject to} \quad & \xi_\mathbf{i} \geq \mathbf{0} \\ & \mathbf{i} = 1, 2, \ldots, \mathbf{n}, \end{aligned} \tag{3}$$

where ξ is the slack variables, C is the user-specified penalty parameter of the error term ($C > 0$), and ϕ is a vector of kernel function. Some popular kernel functions are listed below:

$$\begin{aligned} &\text{Linear kernel} && K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \mathbf{x}_j \\ &\text{Polynomial kernel of degree } g && K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i \mathbf{x}_j + r)^g, \gamma > \mathbf{0} \\ &\text{Radial basis function} && K(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\}, \gamma > \mathbf{0} \\ &\text{Sigmoid kernel} && K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i \mathbf{x}_j + r), \gamma > \mathbf{0}, \end{aligned} \tag{4}$$

where r, γ, and g are kernel parameters and are also user-specified.

In this research, LIBSVM 2.82 (Chang and Lin 2001) is adopted to perform the function of SVM. The optimal SVM-specific parameters can be automatically generated in LIBSVM. According to the user guide of LIBSVM (Hsu, Chang, and Lin 2009), the kernel function is radial basis function (RBF). Therefore, it is necessary to determine two SVM-specific parameters, C and γ. More detailed information about SVM can be found in Wu and Chang (2005) and Cristianini and Shawe-Taylor (2000).

IMPLEMENATION

In this section, a real case of blogs regarding VoIP phone products issues (e.g., the VoIP & Gadgets Blog http://blog.tmcnet.com/blog/tom-keating/; VoIP Blog-Tehrani.com http://voip-forum.tmcnet.com/forum/; VoIP Watch: andyabramson.blogs.com/, and so on) is provided to implement the proposed blog mining procedure.

Defining Variables

In the very beginning, 50 comments on the VoIP phone blog were collected for generating the predefined variables set. Then, some keywords

Table 1. The predefined variables/keywords

Product			
Speaker	Noise	Webcam	Voice quality
Echo	Headset	Hand free	Bluetooth
Bandwidth	Availability	Logoff	Internet call
Skype	MSN	Battery	LCD monitor
Weight	Conference call	Standby time	
Service			
Inspection/Security	Delay/lag	Service	Fare
Charge	Phone number assignment	Bill	International call
Long-distance	Video	Ring tone	Call waiting
Communication time	Clear	Vogue	
Emotion			
Orz/OTZ/or2	囧	^_^/^^”/^”^	XD
:)/ :-)/ : D	>_</ ></ >”<	Expensive	Cheap
: (/ :- (	==/==”/=.=	Convenient	Saving money
Free	Good	Bad	Perfect/wonderful

were added into the predefined set. They are listed in Table 1 and these keywords can be separated into three dimensions: product, service, and emotions. Next, we remove some meaningless words. Finally, we combine some words that have the same meaning. The results are listed in Table 2.

Data Collection

After removing some missing value examples, 200 instances are kept for further analysis. In addition, the variables with very low frequency of

Table 2. Combining keywords according to synonym

Keywords	Combined Keywords
Noise, Voice quality, Delay/lag	Voice quality
Availability, Logoff	Availability
Charge, Fare, Bill	Fare
Orz/OTZ/or2, 囧	Orz
>_</ ></ >”<, ==/==”/=.=	: ((unhappy)
Cheap, Saving money	Saving money

Table 3. The variables with low occurrence frequency

LCD monitor	Standby time	Call waiting	Hand free
Weight	:)/ :-)/ : D	: (/ :- (	

occurrence, listed in Table 3, are removed as well. In total, this study used 35 variables listed in Table 4 for blog mining.

Data Preprocessing

Except data cleaning and normalization, frequency analysis and correlation analysis are also done in this section. The frequency is used to visualize data. The results are described in Table 5. From this table, the top 10 variables contain almost 80% of occurrence frequency. They are "skype," "voice quality," "saving money," "fare," "free," "video," "international call," "handset," "MSN," and "convenient." It's easy to find what customers' major concerns are when using VoIP phone products. The remaining 25 keywords merely involve about 20% of occurrence frequency. They attract much less attention than the top 10.

The correlation analysis is performed to find the strength of relationship between variables. Table 6 shows a part of the correlation matrix of 35 keywords. Most of the correlation coefficients are close to a very small number. It means most of the relationships between variables can be viewed as independent or weak-correlated after combining variables, except the one which is between "inspection/security" and "phone number assignment." Its value is 0.6 which represents the strength of correlation as medium strong. It is easy to understand why customers consider "assigning phone number" as a factor of "inspection/security." Fortunately, the frequencies are 0.83% and 1.83%. Such low frequency does not influence the results of blog mining.

Table 4. The final defined variables/keywords for blog mining

Speaker	Webcam	Echo	Headset
Bandwidth	Internet call	Skype	MSN
International call	Conference call	Service	Phone number assignment
XD	Inspection/Security	Video	Ring tone
Convenient	Long-distance	Clear	Vogue
Free	Communication time	Good	Bad
Perfect/Wonderful	Expensive	Orz	Availability
Fare	Bluetooth	Voice quality	Saving money
: ((unhappy)	^_^/ ^^"/ ^"^ (happy)	Battery	

Table 5. Frequency analysis

Keywords	Frequency (%)	Accumulative Frequency (%)
Skype	21.35	21.35
Voice quality	11.09	32.44
Saving money	8.76	41.20
Fare	8.67	49.87
Free	6.09	55.96
Video	5.59	61.55
Int'l call	5.34	66.89
Handset	5.09	71.98
MSN	4.09	76.06
Convenient	2.50	78.57
Availability	2.09	80.65
Expensive	2.09	82.74
Internet call	1.92	84.65
Phone number assignment	1.83	86.49
Long distance	1.83	88.32
Bluetooth	1.75	90.08
Bandwidth	1.67	91.74
Clear	1.25	92.99
Webcam	1.00	93.99
Echo	0.92	94.91
Inspection/security	0.83	95.75
..	..	..

Experimental Results

Customer Segmentation. In this section, we use clustering analysis to find suitable customer segments. We set the initial and final values of the learning rate (α) and the radius of neighborhood (as % of map width, R)

Table 6. Correlation analysis

	Inspection/ security	Voice quality	Service	Fare	Phone number assignment	
Inspection/Security	1	−0.1	0.11	0.19	0.6	..
Voice quality		1	−0.1	0.02	−0.1	..
Service			1	0.34	−0	..
Fare				1	0.1	..
Phone number assignment					1	..
..	..	..	..	.	.	.

to be 0.6 and 0.1 (α) and 50% and 1.0% (R), respectively. The decay function is exponential. The second layer of SOM is organized as 2D grids. Three grid structures including 1 by 2, 1 by 3, and 2 by 2, are employed to control the number of clusters. The number of input variables is 35 and the data size is 200. The software, NNclust, which is available at http://www.geocities.com/adotsaha/NN/SMMinExcel.html is employed to execute SOM. The learning rate and width of neighborhood decrease geometrically over 100 epochs. Table 7 summarizes the results of sparsity analysis. It shows the number of observations in every single cluster. From this table, when the number of clusters increases from 2 to 4, we can find cluster #4 contains only three objects. The objects in this cluster are too few to be able to represent a specific type of customer. If we merely separate the data into two clusters, it is too rough. Therefore, three clusters as shown in Table 8 seem to be suitable after sparsity analysis.

According to the results of frequency analysis, we merely list the variables till the accumulative frequency reaches 75%. These three clusters can be named as "professional users," "enterprise/company customers," and "economic/public customers." The characteristics of each cluster can be drawn as follows:

Cluster 1: Professional users

Compared with the other two clusters, this type of customer not only considers basic factors such as skype, voice quality, and MSN, but is also concerned with "headset," "fare," "video," "availability," and "phone number assignment." In other words, their major concerns are highly related to professional/advanced functions. They may prefer high-tech or novel-functioned VoIP phone products. Therefore, this cluster can be named "professional users."

Cluster 2: Enterprise/company customers

Table 7. Sparsity analysis

Cluster label / No. of clusters	Cluster #1	Cluster #2	Cluster #3	Cluster #4
2 clusters	168	32	*	*
3 clusters	175	16	8	*
4 clusters	153	28	7	3

Table 8. Customer segmentation

Keywords	Frequency (%)	Accumulative frequency (%)
Cluster 1 Professional users		
Skype	19.08	19.08
Headset	9.92	29.01
Fare	9.16	38.17
Video	8.65	46.82
Saving money	7.89	54.71
Voice quality	5.85	60.56
International call	5.09	65.65
MSN	4.07	69.72
Availability	3.56	73.28
Phone number assignment	3.31	76.59
Cluster 2 Enterprise/company customers		
Skype	27.69	27.69
Voice quality	20.92	48.62
Saving money	6.46	55.08
Fare	6.15	61.23
MSN	5.54	66.77
Headset	4.92	71.69
Video	4.62	76.31
Cluster 3 Economic/public customers		
Skype	18.92	18.92
Free	11.64	30.56
Saving Money	11.02	41.58
Fare	9.98	51.56
Voice quality	8.73	60.29
International call	7.28	67.57
Internet call	4.57	72.14
Video	3.74	75.88

Customers in cluster 2 are interested mainly in "voice quality," "saving money," and "fare." Among them, "voice quality" owns very high priority. Customers in this cluster might be from enterprises/companies. For this kind of user, they hope to enjoy high voice quality to make sure smooth/effective communication and avoid losing any messages from suppliers/customers. Meanwhile, they care about fare. Therefore, they may prefer reliable VoIP phone products, which can provide high quality voice as well. This type of customer belongs to "enterprise/company customers."

Cluster 3: Economic/public customers

Customers in cluster 3 focus on "free," "fare," and "saving money." Cost is their first priority. They want to spend little money and enjoy lots of services. This type of customer might come from the public, such as personal or economic users. Price might be a major concern for them.

Customer Classification. This study uses three popular supervised learning methods for customer classification. The 5-fold cross-validation experiments have been implemented. It means we divide 200 examples into five equal partitions (see Table 10). Then each partition is in turn taken as test examples and the rest are used as training instances. For supervised learning, we need the class label to be used as the learning target. The class labels are obtained from the following procedure. First, we cluster the training set and then label the test set according to the distance between test sample and cluster center of constructed clusters which are obtained from training sets. Finally, the test set is tested based on the classifier trained by using the training set.

This study employs three famous classification tools—SVM, BPN, and C4.5. The BPN with one hidden layer is adopted and implemented in a Matlab 6.1 (The MathWorks Inc., Natick, MA, USA) environment. The optimal structure and parameter settings are obtained from trial-and-error experiments whose results can be found in Table 9. Besides, See5 (C4.5 commercial version) software is utilized to construct a decision tree. In See5, there are two parameters that can be tuned during the pruning phase: the minimal number of examples represented at any branch of any feature-value test; and the confidence level of pruning. To avoid the occurrence of overfitting and generating a simple tree, 2 is set as the minimum number of instances at each leaf, and the confidence level for pruning is set to 25%. For SVM, when using RBF kernel functions, two parameters, C and γ, needs to be determined. Fortunately,

Table 9. The parameter settings of SVM and BPN

Method / Data set	SVM		BPN		
	C	γ	Structure	Learning rate	Iteration
Fold 1	32	0.03125	35-15-1	0.1	200
Fold 2	128	0.03125	35-20-1	0.1	100
Fold 3	512	0.0078125	35-10-1	0.15	200
Fold 4	32	0.0078125	35-15-1	0.1	150
Fold 5	512	0.03125	35-25-1	0.01	200

Table 10. Summary of 5-fold experiments

	SVM (%)	BPN (%)	C4.5 (%)
Fold 1	82.5	75.0	77.5
Fold 2	80.0	70.0	77.5
Fold 3	57.5	62.5	60.0
Fold 4	67.5	57.5	60.0
Fold 5	65.0	55.0	63.5
Average	70.50	64.00	67.70
StDev	10.52	8.40	9.06
Max. Accuracy	82.50	75.00	77.50
Min. Accuracy	57.50	55.00	60.00
Range	25.00	20.00	17.50

these two parameters can be found automatically by LIBSVM. The settings of C and γ can be found in Table 9. The inputs and outputs of these three methods are 35 attributes and one defined class label, respectively. The summary of experimental results can be found in Table 10.

From Table 10, the average classification accuracy of SVM (70.50%) is better than those of C4.5 (67.70%) and BPN (64%). Considering maximal accuracy, SVM still owns the better performance (82.5%) compared to the other two methods (BPN: 82.5%; C4.5: 77.5%). However, the results of SVM are more uncertain (standard deviation = 10.52%; range = 25%) than those of BPN (standard deviation = 8.40%; range = 20%) and decision tree (standard deviation = 9.06%; range = 17.50%). In 5-fold cross-validation experiments, SVM has a better performance for classifying customer type, except the fold 3 experiment. If we do not consider the result of the fold 3 experiment, the result could be more clearly illustrated. In this situation, SVM owns the best accuracy (73.75%) and the lowest standard deviation (8.78%), compared to those of BPN (average accuracy = 64.38%; standard deviation = 9.66%) and C4.5 (average accuracy = 69.63%; standard deviation = 9.20%).

CONCLUSION AND FUTURE WORKS

The blog is becoming the new media of the Internet generation. Because most blogs on the Internet are personal or journalistic, the main purpose of this work is to find what out they are interested in and to provide useful knowledge to companies for customer segmentation and classification. This study proposed a BM model for extracting knowledge from blogs.

A real data of VoIP phone products is provided to evaluate the effectiveness of our method. In the proposed model, by using SOM, several customer segments have been established. Subsequently, based on the constructed segments, we use three learning algorithms to classify customers. Experimental results show the superiority of the proposed procedure. We can segment customers and build a classifier to predict the characteristics of unknown customers. This study also found that SVM indeed has a better ability for classifying sparse data than BPN and C4.5.

There are some issues of blog mining needed to be investigated in the future. The first one is the variable definition problem. In addition to the polysemy and synonym problem, there are many blogger-created netspeaks including funny words and interesting icons in blogs. It is a very tough task to define them clearly. For further works, researchers should pay much attention to understanding what their true meanings are. The second issue is the sparse data problem in blogs. These sparse data not only consume the saving space, but also possibly degrade the performance of classifiers. Additionally, due to the population of Youtube, the content of blog mining should contain multimedia, especially video data, not limited to text.

DECLARATION OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Asano, E. 2007. A public outreach in epilepsy surgery using a serial novel on BLOG: A preliminary report. *Brain & Development* 29: 102–104.

Borzemski, L. 2006. The use of data mining to predict web performance. *Cybernetics and Systems* 37: 587–608.

Chakrabarti, S. 2000. Data mining for hypertext: A tutorial survey. *SIGKDD Exploration* 1: 1–11.

Chang, C.-C. and Lin, C.-J. 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm.

Chau, M. and Xu, J. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human – Computer Studies* 65: 57–70.

Chen, H. and Chau, M. 2004. Web mining: Machine learning for web applications. *Annual Review of Information Science and Technology (ARIST)* 38: 289–329.

Cho, Y. H., Kim, J. K., and Kim, S. H. 2002. A personalized recommender system based on web usage mining and decision tree induction. *Expert Systems with Applications* 23: 329–342.

Cohen, E. and Krishnamurthy, B. 2006. A short walk in the Blogistan. *Computer Networks* 50: 615–630.

Cristianini, N. and Shawe-Taylor, J. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Du, H. S. and Wagner, C. 2006. Weblog success: Exploring the role of technology. *International Journal of Human-Computer Studies* 64: 789–798.

Epstein, M., Friedl, M., and Yuthas, K. 2009. Managing customer value. *CMA Management*: 28–31.

Facca, F. M. and Lanzi, P. L. 2005. Mining interesting knowledge from weblogs: A survey. *Data & Knowledge Engineering* 53: 225–241.

Freeman, A. J. and Skapura, M. D. 1992. *Neural networks: Algorithms, applications, and programming techniques*. Reading, MA: Addison Wesley.

Gil-Saura, I. and Ruiz-Molina, M. 2009. Customer segmentation based on commitment and ICT use. *Industrial Management & Data Systems* 109: 206–223.

HaCohen-Kerner, Y., Stern, I., Korkus, D., and Fredj, E. 2007. Automatic machine learning of keyphrase extraction from short HTML documents written in Hebrew. *Cybernetics and Systems* 38: 1–21.

Hof, R. 2005. Blogs on ice: Signs of a business model? *Business Week Online – The Tech Beat*, June 2, 2005. Available at http://www.businessweek.com/the_thread/techbeat/archives/2005/06/blogs_on_ice_si.html (accessed May 19, 2009).

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. 2009. A practical guide to support vector classification. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html (accessed May 19, 2009).

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE* 78: 1464–1480.

Kolari, P. and Joshi, A. 2004. Web mining: research and practice. *Computing in Science & Engineering* 6: 49–53.

Kosala, R. and Blockeel, H. 2000. Web mining research: A survey. *ACM SIGKDD Explorations* 2: 1–15.

Kukolj, D., Atlagic, B., and Petrov, M. 2006. Data clustering using a reorganizing neural network. *Cybernetics and Systems* 37(7): 779–790.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. 1999. Trawling the web for emerging cyber-communities. *Computer Networks* 31: 1481–1493.

Lappas, G. 2007. An overview of web mining in social benefit areas. *The 9th IEEE International Conference on E-Commerce Technology and 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Service*, July 23–26, 683–690.

Larkin, M. 2005. Blogs: New way to communicate about infectious diseases. *The Lancet Infectious Diseases* 5: 748.

Lin, Y.-S. and Huang, J.-Y. 2006. Internet blogs as a tourism marketing medium: A case study. *Journal of Business Research* 59: 1201–1205.

Lu, Z., Lin, F., and Ying, H. 2007. Design of decision tree via kernelized hierarchical clustering for multiclass support vector machines. *Cybernetics and Systems* 38: 187–202.

Mathieu, J. 2007. Blogs, podcasts, and wikis: The new names in information dissemination. *Journal of the American Dietetic Association*: 553–555.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., and Swartz, L. 2004. Why we blog. *Communications of the ACM* 47: 41–46.

Olafsson, S., Li, X., and Wu, S. 2008. Operations research and data mining. *European Journal of Operational Research* 187: 1429–1448.

Oransky, I. 2005. Cancer blogs. *Lancet Oncology* 6: 838–839.

Philip, D. W. 1992. *Neural computing: Theory and practice.* New York: Van Nostrand Reinhold.

Porter, L. V., Trammell, K. D., Chung, D., and Kim, E. 2007. Blog power: Examining the effects of practitioner blog use on power in public relations. *Public Relations Review* 33: 92–95.

Qin, J., Zhou, Y., Reid, E., Lai, G., and Chen, H. 2006. Unraveling internet terrorist groups' exploitation of the web: technical sophistication, media richness, and web interactivity. *Proceedings of the Workshop on Intelligence and Security Informatics (WISI 06)*, Singapore.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1: 81–106.

Quinlan, J. R. 1993. *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Rahman, M. K. M., Pi, Y., Wang, C., Tommy, W. S., and Wu, S. 2007. A flexible multi-layer self-organizing map for generic processing of tree-structured data. *Pattern Recognition* 40(5): 1406–1424.

Rosenbloom, A. 2004. The blogosphere. *Communications of the ACM* 47: 31–33.

Rumelhart, D. E. and McClelland, J. L. 1986. *Parallel distributed processing,* Vol. 1. Cambridge, MA: MIT Press.

Salas, R., Moreno, S., Allende, H., and Moraga, C. 2007. A robust and flexible model of hierarchical self-organizing maps for non-stationary environments. *Neurocomputing* 70: 2744–2757.

Seiffert, U. and Jain, L. 2002. *Self-Organizing Neural Networks: Recent Advances and Applications. Studies in Fuzziness and Soft Computing*, 78. Berlin: Springer.

Simkin, L. 2008. Achieving market segmentation from B2B sectorisation. *Journal of Business & Industrial Marketing* 23: 464–474.

Su, C.-T. and Shiue, Y.-R. 2003. Intelligent scheduling controller for shop floor control systems: A hybrid genetic algorithm/decision tree learning approach. *International Journal of Production Research* 12: 2619–2641.

Todoroki, S., Konishi, T., and Inoue, S. 2006. Blog-based research notebook: personal informatics workbench for high-throughput experimentation. *Applied Surface Science* 252: 2640–2645.

Trammell, K. D. 2006. Blog offensive: an exploratory analysis of attacks published on campaign blog posts from a political public relations perspective. *Public Relations Review* 32: 402–406.

Vapnik, V., Golowich, S., and Smola, A. J. 1996. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Wang, B. and Liu, Z. 2003. Web mining research. *Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2003)*: 84–89.

Wang, C., Lu, J., and Zhang, G. 2007. Mining key information of web pages: A method and its application. *Expert Systems with Applications* 33: 425–433.

Wood-Black, F. and Pasquarelli, T. 2007. Blogs. *Journal of Chemical Health & Safety* 14: 37.

Wu, G. and Chang, E. Y. 2005. KBA: Kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering* 17: 786–795.

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D., and Steinberg, D. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14: 1–37.

Zhou, Y., Reid, E., Qin, J., Chen, H., and Lai, G. 2005. US domestic extremist groups on the web: Link and content analysis. *IEEE Intelligent Systems* 20: 44–51.