Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 603–610
Expert Systems with Applications
www.elsevier.com/locate/eswa

Web service quality control based on text mining using support vector machine

Shuchuan Lo *

Department of Industrial Engineering and Management, National Taipei University of Technology 1, Section 3, Chung-Shiao East Road, Taipei, Taiwan 106, Taiwan, ROC

Abstract

Popular websites can see hundreds of messages posted per day. The key messages for customer service department are customer complaints, including technical problems and non-satisfactory reports. An auto mechanism to classify customer messages based on the techniques of text mining and support vector machine (SVM) is proposed in this study. The proposed mechanism can filter the messages into the complaints automatically and appropriately to enhance service department productivity and customer satisfaction. This study employs the *p*-control chart to control the complaining rate under the expected service quality level for the website execution. This study adopts a community website as an example. The experimental results demonstrated that namely the ability of the SVM to correctly recognize defective messages exceeded 83% with an average of 89% for the classifying mechanism, and the *p*-control chart was capable of reflecting unusual changes of service quality timely.

Keywords: Text mining; Classification; Support vector machine (SVM); Compensating *p* control chart; Web quality control

1. Introduction

Community websites generally have multiple functions, which combine contents, members and commerce to attract web users and achieve the beneficial purpose of web execution. Community websites face severe challenges because too many similar community websites share a limited market. To increase member numbers and income, the managers of community websites regularly update the web contents and functions to enhance survival conditions in this highly competitive environment.

Most websites provide a message board function in customer service department to gather complaints and requests from customers. Popular community websites experience hundreds of thousands of messages entering their databases every day, leading to the customer department facing a ''data explosion''. Websites require an auto mechanism to filter the useful messages and even transfer them into customer knowledge. This study proposes an auto mechanism known as Web-complaint Quality Control (WebQC), which can recognize the complaint message and issue a warning signal when the number of complaints exceeds the usual level. In the WebQC, this study first uses text mining to extract the keywords based on the weight calculation of TFIDF (Term-Frequency Inverse-Document-Frequency) (Salton & McGill, 1983) and uses SVM (Support Vector Machine) (Cortes & Vapnik, 1995) to classify the messages into four categories, including Non-Chinese message or disorder code, technical problem, report of dissatisfaction, and others. This study regards the messages involving technical problems and expressing dissatisfaction as complaints. This study uses the *p*-control chart to control the complaint rate. If the complaint rate exceeds the upper control limit, then WebQC issues a warning signal to demonstrate the decline in service quality.

The rest of this paper is organized as follows. In Section 2, we review some related techniques of text mining. Section 3 describes the auto mechanism structure of WebQC and the

* Tel.: +886 2 27712171x2368; fax: +886 2 87739603.
E-mail address: sclo@ntut.edu.tw

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.09.026

methods used in this research, SVM and *p*-control chart. The experimental results employing WebQC on a community website for eight months are presented in Section 4. Finally, the paper is concluded in Section 5.

2. Related work of text mining

2.1. Text classification

Text classification (or text categorization) is to label the natural language texts with thematic categories from a predefined set (Sebastiani, 2002). Text classification dates back to 1960s, but it became a major subfield in the early 1990s. In the late 1980s, the most popular method of text classification was knowledge engineering, which classified documents under the given categories according to the rules encoded by expert knowledge. However, in the 1990s, the general approach in text classification was machine learning building an automatic text classifier by learning from a set of pre-classified documents, the characteristics of the categories of interest.

The text categorization is to assign a Boolean value to each pair $(d_i, c_j) \in \mathbf{D} \times \mathbf{C}$, where $\mathbf{D}$ is a domain of documents and $\mathbf{C}$ is a set of predefined categories. A value T assigned to (d_i, c_j) indicates a document d_i under category c_j, while a value F represents that d_i is not under c_j. It can be formally described by Φ: $\mathbf{D} \times \mathbf{C} \rightarrow \{T, F\}$, where Φ is an unknown function. Classifier ϕ is an estimator of Φ. In this research, the classification is not overlapped. That means that for any document d_j in $\mathbf{D}$, if there is a classifier $\phi_i(d_j) = T$, then all other classifiers $\phi_k(d_j) = F$, for $i = 1, \ldots, |\mathbf{C}|$ and $k \neq i$, where $\mathbf{C} = \{c_1, \ldots, c_{|C|}\}$ consists of $|\mathbf{C}|$ independent categories and a classifier for c_i is then a function $\phi_i : \boldsymbol{D} \rightarrow \{T, F\}$ that approximates an unknown target function $\Phi_i : \mathbf{D} \rightarrow \{T, F\}$.

2.2. Segmentation of chinese text

Text segmentation that divides texts into linguistic units and common words is a prerequisite for text classification. The words segmented from customers' complaints (Chinese text) can be used as indexing terms. We retrieve the information (semantics) about the complaints and build the rules for each complaint category by those keywords. Chinese texts do not have obvious word boundaries as English texts with natural boundaries as spaces. A Chinese text consists of a linear sequence of non-spaced or equally spaced indo-graphic characters, which are similar to the morphemes in English. There are three main methods for Chinese text segmentation and text retrieval, which are word identification (Case & Zeng, 1991), statistical word identification (Fan & Tsai, 1988) (Sproat & Shih, 1990) and hybrid word identification (Nie, Briscbois, & Ren, 1996). Word identification is based on a word/phrase dictionary or thesaurus. The maximum-matching (or longest matching) algorithm is often used to select the word sequence, which contains the longest words. This method is intuitive and easy to implement. Statistical approaches rely on statistical information such as word and character occurrence frequencies in the training data – often a set of manually segmented texts. A simple statistical approach is to calculate the probability of a character string S to be a word as $P(S) = \frac{CW(S)}{C(S)}$, where $CW(S)$ is the number of occurrences of S being segmented as a word in the training set and $C(S)$ is the number of occurrences of S in the training set. Given an input string to be segmented, the best solution is composed of a sequence of potential words S_i such that $\prod_i P(S_i)$ is the highest. The hybrid approach combines the word based approach and statistical approach. It uses word dictionary to segment all kinds of words and then the statistical probability to decide the best segmentation. In this research, we use a hybrid based system, AutoTag, (CKIP, 2006) developed by CKIP group of Academia Sinica, Taiwan, to be our Chinese segmentation tool. The thesaurus of AutoTag has about 100,000 chinese terms.

2.3. Information retrieval

Information retrieval or text retrieval is the extraction from texts of certain meaningful content and bearing units. Basically every text is composed of many keywords. If we use a vector to represent the text, then the text is a vector composed of the projecting length from all keywords. Keyword retrieval is mainly to recognize the meaningful and representative words or phrases in texts.

Text segments have all kinds of words include noun, verb, pronoun, article, conjunction and preposition, which are not necessary for the information retrieval. The meaning of the text would not be twisted if we deleted the words of pronoun, article, conjunction and preposition. Nevertheless, the process of keywords retrieval is done after text segmentation.

There are three general approaches in keyword retrieval, which are dictionary approach, linguistic approach and statistical approach. Dictionary approach is to use the pre-defined phrase dictionary to retrieve keywords. Linguistic approach is to retrieve all the phrases of noun, verb, pronoun or preposition by linguistics and use some rules to filter meaningless phrases. Statistical approach is to retrieve keywords by the frequency weight of a term in a text and text set. Salton and McGill (1983) proposed that in order to decide the importance and representation of a term in a document, the term frequency (TF) in this document and the frequency of this term that appears in other documents can be calculated, which is called inverse document frequency (IDF) (Salton & McGill, 1983). The value TF of a term "higher" means it is more important in this text. The purpose of IDF is to find the terms different from other texts. The value IDF of a term "lower" means it can distinguish "better" from other texts. The statistical information retrieval combines these two indexes as TFIDF which is defined as: