Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Soft Computing 6 (2005) 62–71
Applied Soft Computing
www.elsevier.com/locate/asoc

# An e-mail analysis method based on text mining techniques

S. Sakurai*, A. Suyama

*Corporate Research and Development Center, Toshiba Corporation, 1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan*

Received 1 October 2003; received in revised form 9 August 2004; accepted 1 October 2004

**Abstract**

This paper proposes a method employing text mining techniques to analyze e-mails collected at a customer center. The method uses two kinds of domain-dependent knowledge. One is a key concept dictionary manually provided by human experts. The other is a concept relation dictionary automatically acquired by a fuzzy inductive learning algorithm. The method inputs the subject and the body of an e-mail and decides a text class for the e-mail. Also, the method extracts key concepts from e-mails and presents their statistical information. This paper applies the method to three kinds of analysis tasks: a product analysis task, a contents analysis task, and an address analysis task. The results of numerical experiments indicate that acquired concept relation dictionaries correspond to the intuition of operators in the customer center and give highly precise ratios in the classification.


*Keywords:* Text mining; Fuzzy inductive learning; E-mail; Customer center

## 1. Introduction

Recently, the decisive importance of delivering the highest possible level of customer satisfaction has been widely recognized. Consequently, customer centers dealing with the requests and complaints of customers are assuming a more important role. At the same time, the number of inquiries made to customer centers via e-mail is rapidly increasing. There are two reasons for this increase: e-mail is increasingly selected as the inquiry medium for companies and it makes it easier for customers to make inquires. It is becoming difficult for customer centers to process inquiries and analyze them. Customer centers need a method that classifies e-mails, facilitates their analysis, and transfers them to the appropriate departments.

An e-mail is composed of date, e-mail address, subject, body of the e-mail, and so on. It is possible for the body to include pictures, sounds, and programs, but the body is mainly composed of textual data. Thus, it is possible to use text mining techniques [1,4,7,15] in order to analyze e-mails [10,17]. One paper [17] proposed a method that uses the number of words, the number of lines, and the frequency of important keywords as characteristic values and identifies the person who wrote the e-mail. Another paper [10]

* Corresponding author. Tel.: +81 44 549 2398; fax: +81 44 520 1308.

1568-4946/$ – see front matter 
doi:10.1016/j.asoc.2004.10.007

proposed a method that evaluates each word by using the *age* of its word, deletes unimportant words from a word vector, and classifies e-mails based on the word vector. Here, the age of a word is calculated based on the arrival times of e-mails including the word. These methods are not sufficiently able to analyze e-mails collected in a customer center, because the e-mails are sent from many people and include various expressions. That is, even if two e-mails have the same meaning, they may not include common keywords. Also, even if these methods classify e-mails appropriately, it is difficult to understand the validity of classified results by referring to a simple word vector.

On the other hand, some papers [8,11] uses Support Vector Machine (SVM) [2,16] to classify the textual data and show that SVM is an appropriate classifier. Also, Latent Semantic Indexing (LSI) [3], and Probabilistic Latent Semantic Indexing [5] (PLSI) are used to characterize textual data in the field of natural language processing. These methods may be helpful for the analysis of e-mails. However, it is difficult for users to judge whether a classification model given by SVM is valid or not, because SVM acquires hyperplanes, which discriminate classes of items of the textual data in high dimensional space. These indexing methods generate characteristic vectors by integrating words and phrases included in items of textual data into some values. It is difficult for users to intuitively understand relationships between words and phrases, and the items, because the relationships are buried in the characteristic vectors.

Thus, this paper proposes a method to analyze e-mails based on text mining techniques [7,12]. The method generates training examples from training e-mails and their classes based on a key concept dictionary. The method acquires a concept relation dictionary from the training examples by using a fuzzy inductive learning algorithm. The acquired concept relation dictionary is described in the format of a fuzzy decision tree. Users are able to understand the concept relation dictionary easily. The method also classifies new e-mails to be evaluated by using the acquired dictionaries and presents statistical information related to each class. The paper demonstrates the effectiveness of the method by applying it to three kinds of analysis task: a product analysis task, a contents analysis task, and an address analysis task.

## 2. A text mining method

### 2.1. A text mining flow

The text mining method [7,12] classifies textual data into some text classes by using a flow shown in Fig. 1. The method uses both the lexical analysis and two kinds of domain-dependent knowledge dictionaries: a key concept dictionary and a concept relation dictionary. The key concept dictionary is a kind of thesaurus and is composed of three layers: a concept class, a key concept, and an expression. Each concept class shows a set of concepts that have a common feature, each key concept shows a set of expressions that have the same meaning, and each expression shows important words and phrases for a target problem. It is possible for the dictionary to deal with different expressions based on their meaning. On the other hand, the concept relation dictionary is composed of relations, which have a condition part with some key concepts, and a result part with a text class. The relations are described in the format of a fuzzy decision tree. Here, the tree is composed of two kinds of nodes and branches connecting an upper node to a lower node. One kind of node is called a branch node and has an attribute. The other kind of node is called a leaf node and has classes with degrees of certainty. Each branch has a fuzzy class item corresponding to an attribute of an upper node. The fuzzy class item is composed of a key concept and its membership function. In the tree, a concept relation is expressed by a path from the top node (root node) to a leaf node. Each relation describes complicated meaning created by the combination of key concepts. The complicated meaning shows a viewpoint of text analysis and the viewpoint is corresponding to a text class.

First, the method applies the lexical analysis to the item and decomposes the item into words, because a language, such as Japanese, is described without word segments. Each word is checked to ascertain whether the word is registered in a key concept dictionary. If a word is registered in the dictionary, a key concept corresponding to the word is assigned to the item.