E-MAIL CLASSIFICATIONS USING SUPPORT VECTOR MACHINE

Mariah Mohd Daud, Ali Selamat

Department of Graphics and Multimedia
Faculty of Computer Science & Information System
Universiti Teknologi Malaysia
(mariah_daud@hotmail.com, aselamat@fsksm.utm.my)

Extended Abstract

Text categorizations contain classification process of text documents into a fixed number of pre-defined categories by user. While many researches have been focused on image classification, text classification on web page, we propose a method to help manage the inbox by automatically classifying email based on category using Support Vector Machine technique. The processes involved are as follows: read input data email from subject and body, feature extraction, feature selection and classify data using Support Vector Machine (SVM) algorithm. The expected result will show that Support Vector Machine is very effective in email classifying process.

Keywords : email classification, support vector machine (SVM), feature selection, text categorization

THE RESEARCH APPROACH

Electronic mail and information overload has become a significant problem over the last several years. Time magazine estimated that 776 billion email message were sent in 1994, 2.6 trillion sent in 1997 and 6.6 trillion email message will be sent in 2000 [Gwynne, S, 1997]. To address this problem, many researchers have designed systems to automatically classify email. The email is typically classified into folders. Figure 1 show an example email that will be classified based on content subject and body. The folder hierarchy is usually flat and distinct, an example message cannot belong to two folders and the content of a folder is independent from the content email.
Existing research has focused on a variety of learning algorithms to classify email into folders. The best choosing an algorithm can make a system to classify incoming email with very effective. For this research, email classification has been designed using Support Vector Machine (SVM) technique. A commonly deployed email classification learning algorithm is based on vectors of term-frequency / inverse-document-frequency (tf-idf) values. These values are used to create a vector that represents both email messages and the contents of a folder. Email vectors and folder vectors can be compared to one another through the cosine metric or a dot product. Email message is classified into folder whose vector mostly matches the vector for the message [4].

Support Vector Machine is the best choosing for classifying process. Support vector machines are based on the Structural Risk Minimization principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis $h$ for which we can guarantee the lowest true error [Joachims, 1998]. Support Vector Machine can solve linear and non-linear problem. Using a binary classifier for each such category, this leads to the following type of supervised learning problem. Example is training sample, $S_n$

$$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \quad (1)$$

$x_i$ represents the document content and $y_i \in \{-1, +1\}$ indicates the class. Support Vector Machine learn linear decision rules $h(x) = \text{sign}(w \cdot x + b)$ described by a weight vector $w$ and a threshold $b$. For a given training sample $S_n$, the SVM finds the hyperplane with maximum soft-margin. Figure 3 shows that email classifying process begin from incoming email until the email is categorized.
Figure 2: Flow Diagram email classifying process
Figure 3: Email has been classified into category

Figure 3 show that email has been classified into category based on email classifying process. This project has been implemented using Microsoft Outlook. Outlook’s interface already supports a view that groups email by category [Kenrick Mock, 2001].

CONCLUSION

In this paper we have presented a new approach of email classification process using Support Vector Machine. This method has been applied to classify email in Outlook. From our research, we conclude that Support Vector Machine is the best choosing compared with other method.

ACKNOWLEDGEMENT

Thank to the reviewers for giving any suggestion.

References


[5] Gwynne, S and Dickerson, J Lost In The E-mail. Time Magazine, April 21, 1997