Email Categorization using Inherent Features and Fuzzy Theory

Sun Park[1], Jun-Woo Shin[2], JangWoo Kwon[3], Min A Jeong[4], Yeonwoo Lee[5], Seong Ro Lee[6]

[1]Institute Research of Information Science and Engineering, Mokpo National University, South Korea
[2]National IT Industry Promotion Agency, South Korea
[3]Department of Computer & Information Engineering, INHA University, South Korea
[4]Department of Computer Engineering, Mokpo National University, South Korea
[5]Department of Information Communication Engineering, Mokpo National University, South Korea
[6]Department of Information Science and Electronic, Mokpo National University, South Korea
[1,4,5,6]{sunpark, majung, ylee, srlee}@mokpo.ac.kr, [2]sjw@nipa.kr, [3]jwkwon@inha.ac.kr

Abstract. In this paper, we propose an email categorization method using fuzzy theory and inherent feature of messages set of email. The proposed method can automatically classify emails into category labels, which supports keyword search and directory search method to efficiently manage the classified result with relation to a large volume of emails. In addition, it can reorganize email category hierarchy regarding user's viewpoint which enhances the efficiency of directory search for the recall rate.

Keywords: E-mail classification, category organization, fuzzy relational product, inherent feature.

1 Introduction

Generally, email classification method can be classified into supervised approaches and unsupervised approaches. The supervised approaches have posed attention to solve the problem of detecting spam messages. Among the approaches used for classifying such spam messages, which are based on classification techniques such as support vector machines (SVM) [1], Bayesian classifiers [2], rule-based classifiers [3], etc. These approaches can well classify emails in connection with labels (i.e., classes) of category which includes representing the inherent features of emails. However, in the approaches, the preparation work has to be preceded before emails are classified.

Other approaches are based on unsupervised classification technique using clustering methods [4, 5, 6] and data mining algorithms [7], which automatically creates a folder from a set of incoming messages for mail classification. But, these approaches achieve less performance of email classification than the supervised

approaches. In addition, the created labels cannot occasionally reflect the property of email set with relation to its class.

In order to resolve the above limitations of supervised and unsupervised approaches, this paper proposes an automatic email categorization method using an inherent feature and fuzzy theory. The proposed method has the following advantages. First, it is the automatic email categorization to classify emails and automatically creating email labels by using fuzzy association and NMF (non-negative matrix factorization)[8]. Thus, it can improve the quality of email classification since the clustered emails helps us to remove redundant information easily and to avoid the biased inherent semantics of emails with regard to email category labels. Second, the proposed method is unsupervised classification method which classifies emails without training and manual labels. So it can classify emails more quickly than supervised approaches. Third, a large number of emails are managed efficiently to support keyword search and directory search method. Final, the classified email set can be reorganized by using fuzzy product whenever a user is dissatisfied with the results of email categorization.

2 Emails Categorization Method

The proposed email categorization method consists of three phases: preprocessing phase, email category construction phase, and reorganizing email category phase.

2.1 Preprocessing

In this section, keywords are extracted from a Body and Subject of emails. The Rijsbergen's stop words list is used to remove all stop words, and word stemming is performed using the Porter's stemming algorithm [9]. Then, the email keyword frequency matrix E is constructed from the email messages set.

2.2 Email Category Construction

The email category construction phase generates category labels by using semantic features of NMF, and then emails are classified to category labels by using fuzzy association. This phase is described as follows.

First, the keywords are extracted from the received emails, and then the email-keyword frequency matrix is constructed. Second, the email-keyword frequency matrix E is decomposed into the non-negative semantic feature matrix W and the non-negative semantic variable matrix H by using NMF. The number of category label is set by the number of semantic feature r. Email category labels are selected using semantic feature matrix W, which the label is decided by keyword corresponding to an element having the highest value in column vector W_{*j}. Final, we exploit the Haruechaiyasak's [10] document classification method to classify emails into category label.

2.3 Reorganizing email category

In the reorganizing email category phase, email category hierarchy is reorganized by fuzzy relational product, whenever a user is dissatisfied with the results of classifying emails by the proposed email category construction method. Here, the relationship between category and keyword can be constructed based on the frequency of keywords in the corresponding email category label. This relationship enables a category to be regarded as fuzzy set comprising keywords and their membership degrees as members. The relationship of two category labels can be defined using the similarity of two category labels, and their similarity can be calculated in the inclusion degree of two fuzzy sets. Therefore, a similar relation of two different category labels can be created so as to reorganize a category hierarchy of email messages by using fuzzy relational product [11].

Reorganizing email category hierarchy is reconstructed from a similar relation of two different category labels by using Equation (1) and (2) with respect to the control of α value. The reorganizing root category includes all the categories and email messages. Emails in the category labels are reorganized by the inclusion of the email message between the child category labels. We use the implication operator defined for reorganizing email category as follows [11].

$$a \rightarrow b = (1 - a) \vee b = \max(1 - a, b),\ a = 0 \sim 1,\ b = 0 \sim 1 \tag{1}$$

$$\pi_m(aR \subseteq Sc) = \frac{1}{N_{U_2}} \sum_{y \in U_2} (\mu_{aR}(y) \rightarrow \mu_{Sc}(y)) \tag{2}$$

here, π_m is a function to calculate the mean degree, the after-set aR for $a \in U_1$ is a fuzzy subset of U_2 such that y is related to a, for $y \in U_2$. Its membership function is denoted by $\mu_{aR}(y) = \mu_R(a,y)$.

3 Performance Evaluations

We have conducted performance evaluation by testing proposed method and comparing it with 5 other representative data clustering method using the same data corpora. We implemented 6 email classification methods: TFIDF, MINING, TFIDF-DCH, PCA-DCH, NMF-DCH, and NMF-FA. NMF-FA denotes our proposed method. TFIDF denotes Mock's method using vector model [4]. MINING denotes Manco and Masciari's method using data mining algorithm [7]. TFIDF-DCH denotes our previous method using vector model and DCH [5]. PCA-DCH and NMF-DHC denote our previous method using DHC based on PCA and NMF [6]. The normalized metric $\overline{MI}$ evaluation results of NMF-FA is approximately 10.45% higher than that of TFIDF, 9.89% higher than that of MINING, 8.27% higher than that of TFIDF-DCH, 6.39% higher than that of PCA-DCH, and 3.63% higher than that of NMF-DCH.

4 Conclusions

In this paper we propose automatic email categorization method, and the architecture of a system to implement it. The proposed method uses inherent feature of emails and fuzzy theory to construct email category and reorganizing email category hierarchy. The method was tested in experiment which shows a high degree of flexibility, efficiency and effectiveness in the email categorization and category hierarchy reorganization.

Acknowledgements

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0028295). "This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency)"(NIPA-2011-C1090-1121-0007).

REFERENCES

1. Drucker, H. Wu, D. and Vapnik, V. N. Support Vector Machines for Spam Categorization. IEEE Transactions on Neural network, 10(5), (1999).
2. Androutsopoulos, I. et al. An Evaluation of Naïve Bayesian Anti-Spam Filtering. In Proc. Workshop on Machine Learning in the New Information Age, (2000).
3. Cohen. W.W. Learning Rules that classify E-mail. In Proc. AAAI Spring Symposium in Information Access, (1999).
4. Mock, K. Dynamic Email Organization via Relevance Categories. In Proceedings of the International Conference on Tools with Artificial Intelligence 1999. Chicago IL, Nov. (1999).
5. Park, S. Park, S. H. Lee, J. H. Lee, J. S. E-mail Classification Agent Using Category Generation and Dynamic Category Hierarchy. LNAI 3397, (2005), 207-214.
6. Park, S. An, D. U. Automatic E-mail Classification Using Dynamic Category Hierarchy and Semantic Features, IETE Technical Review, vol. 27, no. 6, (2010), 478-492.
7. Manco, G. Masciari, E. A Framework for Adaptive Mail Classification. In Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence, (2002).
8. Lee, D. D. Seung, H. S. Learning the parts of objects by non-negative matrix factorization. Nature, 401, (1999), 788-791.
9. Ricardo, B. Y., Berthier, R. N.: Moden Information Retrieval: The Concepts and Technology behind Search (2nd Edition), ACM Press (2011)
10. Haruechaiyasak, C. Shyu, M. L. Chen, S. C. Web Document Classification Based on Fuzzy Association, In proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC'02), Oxford, England, Aug. (2002)
11. Ogawa, Y. Morita, T. and Kobayashi, K. A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy Sets and System, (1991), 163-179.