E-mail Classification and Category Re-organization using Dynamic Category Hierarchy and PCA

Sun Park, Chul-Won Kim*, Dong-Un An, *Member, KIMICS*

Abstract—The amount of incoming e-mails is increasing rapidly due to the wide usage of Internet. We often group emails into categories for maintaining e-mail efficiently. However reading the email messages and classifying them is still tedious task. Moreover, the number of e-mails and manual classifying is increasing everyday. So, automatic e-mail classification is important techniques. In this paper, we propose a multi-way e-mail classification method that uses PCA for automatic category generation and dynamic category hierarchy for re-organizing e-mail categories. It classifies a huge amount of receiving e-mail messages automatically, efficiently, and accurately.

Index Terms—E-mail Classification, Category Re-organization, Dynamic Category Hierarchy, PCA.

I. INTRODUCTION

E-mail has become increasingly important and widespread using method of communication due to reduce time and cost efficiently. It is used not only for personal contacts but also for business, advertising and electronic commerce. So, e-mail users receive tens or even hundreds of e-mail messages everyday, it has need of e-mails management efficiently.

Many tools for e-mail management have been developed. Such tools, however, are mainly human-centered, in the sense that the user is required to manually describe rules and keyword list that can be used to recognize the relevant features of messages. Such an approach has the following disadvantages: Fist, it is incongruent to a large number of e-mails, which many contain too many distinct and possibly overlapping concepts; Second, it is not applied to requirement re-organize their mail, which usually occurs when users need to periodically re-organize their messages and filters [8].

PCA (principal components analysis) are orthogonal projections that together explain the maximum amount of variation in set of data. The principle components can be found by computing the singular value decomposition on the correlation matrix of the set of data [15].

Dynamic category hierarchy method (DCHM) is to enhance search efficiency. The relationship between category and keyword can be constructed based on the frequency of keywords in the corresponding category. This relationship enables a category to be regarded as fuzzy set comprising keywords and their membership degrees as members. The relationship of two categories can be defined using the similarity of two categories, and their similarity can be calculated in the inclusion degree of two fuzzy sets. Therefore, a similar relation of two different categories can be created so as to automatically construct a dynamic category hierarchy [16, 17].

In this paper, we propose an e-mail classification method combining automatic construction of category by PCA and dynamic construction of category hierarchy method. The proposed method in this paper has the following advantages: First, there is unsupervised classification method, which is the automatic construction of subject-based category labels from a set of incoming e-mail messages; Second, we assume a dynamic construction of category hierarchy method of interaction in which the user re-organizes all his e-mail messages; Third, because our method is no use a learning, it classify e-mail more quickly then conventional methods. So it is suitable to the e-mail environment of flexibility.

This paper is organized as follows. In Section 2, we describe the related works regarding e-mail classification methods. Section 3, the dynamic category hierarchy method is described. Section 4, we propose a

Manuscript received April 25, 2009 ; Revised June 3, 2009. Sun Park is with the Advanced Graduate Education Center of Jeonbuk for Electronics and Information Technology-BK21, Chonbuk National University, Jeonju, Korea (Email: sunbak@jbnu.ac.kr), *Corresponding Author: Chul-Won Kim is with the Department of Computer Engineering, Gwangju, Korea(Email: cwkim@honam.ac.kr), Dong Un An is with the Division of Electronic & Information Engineering, Chonbuk National University, Jeonju, Korea (Email: duan@chonbuk.ac.kr)

e-mail classification method using automatic construction of category label based on PCA and dynamic category hierarchy method by the fuzzy relational production. Section 5, some experimental results are presented to show efficiency of the proposed method. Finally, conclusion is made in Section 6.

II. RELATED WORKS

The current study has posed attention to the problem of detecting spam messages. Among the techniques used for classifying such spam messages, which are based on Support-Vector Machines, Bayesian classifiers, rule-based classifiers.

Cohen [1] describes two rule-based systems exploiting text mining techniques for the classification of e-mail correspondence. These approaches differ mainly in the preprocessing phase using simple Boolean or frequency-based vector model.

Androutsopoulos [2] and Sakkis [3] describe Bayesian classifiers for anti-spam filtering. These approaches have been shown more accurate then rule-based classifiers.

Drucker [4] study the use of support vector machines (SVM's) in classifying e-mail as spam or non-spam by comparing it to three other classification algorithms: Ripper, Rocchio, and boosting decision trees. SVM's are the best case.

Kunlun [5] proposes a new method for spam categorization based on support vector machines using active learning strategy.

Woitaszek [6] utilize a simple support vector machine to identify commercial e-mail. This classification system was subsequently implemented as an add-in for Microsoft Outlook XP, providing sorting and grouping capabilities using Outlook's interface to the typical desktop e-mail user.

The above approaches provide a supervised classification method: message folders pre-exist and the main objective is to detect the most likely folder for an incoming message. Also, this methods need learning before classifying e-mails, and it takes some measure time training and test.

Other approaches are based on unsupervised classification technique, which create a folder from a set of incoming messages for mail classification.

Mock [7] proposed classification systems that automatically classify email based upon an inverted index with integrated tf-idf values.

Manco and Masciari [8] uses a techniques based on data mining algorithms for classifying incoming messages, as a basis for an overall architecture for maintenance and management of email messages.

Park et al. propose automatic e-mail classification agent using category generation and dynamic category hierarchy [17].

III. DYNAMIC CATEGORY HIERARCHY METHOD

We define dynamic restructuring of the hierarchy of searched categories [13, 16, 17]. In this section, we give a brief introduction to the Fuzzy Relational Products that is used in Dynamic Construction of Category Hierarchy. The fuzzy set is defined as follows.

Definition 1. α-cut of a fuzzy set A, denoted by A_α, is a set that contains all elements whose membership degrees are equal to or greater than α. $A_\alpha = \{ x \in X \mid \mu_A(x) \geq \alpha \}$.

A fuzzy implication operator is an extended crisp implication operator to be applied in the fuzzy theory. A crisp implication operator is defined as $\{0,1\}\times\{0,1\} \rightarrow \{0,1\}$, while a fuzzy implication operator is defined as $[0,1]\times[0,1] \rightarrow [0,1]$ to be extended in multi-valued logic. We use the implication operator defined as follows [10].

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

Definition 2. The fuzzy implication operators vary in the environments of given problems. The afterset 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 $\mu_{aR}(y) = \mu_R(a,y)$. The foreset Sc for $c \in U_3$ is a fuzzy subset of U_2 such that y is related to c, for $y \in U_2$. Its membership function is denoted by $\mu_{Sc}(y) = \mu_S(y,c)$ for $y \in U_2$. The mean degree that aR is a subset of Sc is meant by the mean degree such that the membership degree of $y \in aR$ implies the membership degree of $y \in Sc$, so it is defined as follows:

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

Here, π_m is a function to calculate the mean degree ♦

IV. PROPOSED METHOD

The proposed e-mail classification method consists of a preprocessing, automatic category construction, and re-organized e-mail category. The preprocessing

performs the task of extracting the keywords, such as representative subjects and contents from e-mail messages. The automatic category construction generates category labels by using PCA, and then e-mail messages are classified to category labels. Re-organized e-mail category labels and messages reconstruct the category hierarchy of e-mail messages using a dynamic category hierarchy method.

A. Preprocessing

It is not suitable a property of e-mail that extract keywords from the preprocessing all items of e-mail messages. This is the reason why e-mail is constricted feature in the ranges of the Subject, Sender, and Body of the text. The preprocessing is stage extracts keywords from a message Body and Subject fields.

In this paper is reduced implementation cost of e-mail classification method using developed already Korea language analysis HAM (Hangul Analysis Module). HAM is shareware that is made in C language. It is supported function that automatic indexing, spelling checker, construction analysis, compound noun disjointing, automatic word spacing are based on Morpheme analyzer. This is kernel library for Korean analysis [14]. Thus, the keywords (index terms) of Subject and Body are extracted in the preprocessing.

B. Automatic Category Construction by PCA

The automatic category construction method use PCA to extract the category labels for e-mail classification. The proposed method is described as follows. We extract keywords from the received messages, and then construct email-keyword frequency matrix.

Table 1 Email-keyword frequency matrix

keyword / email	*t1*	*t2*	*t3*	*t4*	*t5*	*t6*	*t7*	*t8*
e1	0	0	0	0	2	1	0	0
e2	0	0	0	0	2	0	0	0
e3	0	0	0	0	4	4	0	0
e4	5	0	0	0	0	0	0	0
e5	4	2	0	0	0	0	0	0
e6	1	0	0	0	0	0	0	0
e7	0	0	0	0	0	0	1	0
e8	0	1	0	0	0	0	0	0
e9	0	0	1	1	0	0	0	1
e10	0	0	0	1	0	0	0	0

Table 1 shows the email-keyword frequency matrix with respect to 10 e-mail messages. We compute the singular value decomposition on the correlation matrix of the email-keyword frequency matrix by PCA. Table 2 shows the results of PCA from Table 1. We select the largest singular values from under 96.26% accumulation rate in the columns of Table 2 for category labels. Table 3 shows the result of email classification by proposed method.

Table 2 Results of PCA from Table 1

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
t1	**0.741**	-0.50	0.149	-0.41	-0.00	-0.00	-0.00	-0.00
t2	-0.19	0.004	**0.907**	-0.00	-0.26	0.004	-0.18	-0.18
t3	0	0	0	0	0	0	-0.70	0.707
t4	0.040	**0.530**	0.177	-0.52	0.213	-0.47	0.270	0.270
t5	0.637	0.523	0.068	**0.533**	-0.13	-0.06	-0.06	-0.06
t6	-0.00	-0.32	0.307	0.494	0.481	-0.14	0.387	0.387
t7	0.032	0.139	0.074	-0.09	-0.39	0.631	0.451	0.451
t8	-0.06	-0.24	-0.13	0.144	-0.69	-0.59	0.179	0.179
singular	4.129	2.251	0.335	0.264	0.165	0.069	0.034	0
accumulation%	56.95	88.01	92.64	**96.29**	98.56	99.52	100	100

Table 3 Result of email classification by proposed method

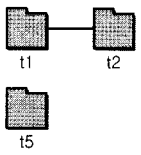
Category labels	Automatic classified e-mail
t1	*e4, e5, e6*
t6	*e9, e10*
t2	*e5, e8*
t5	*e1, e2, e3*
t7	*e7*

C. Re-organized e-mail category by DCHM

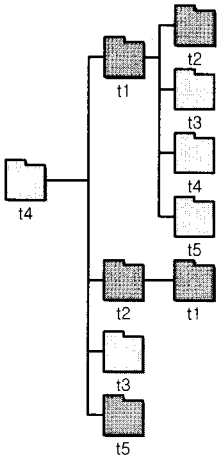
In this section, we use dynamic category hierarchy method for re-organized e-mail category. Here, the relationship between keywords and categories can be decided by normalized keyword frequency values between 0 and 1 (meaning a fuzzy degree of membership). That is, a category can be regarded as a fuzzy set comprising the keywords appearing in the e-mail messages pertaining to the corresponding category. Actually, in the domain of categories, we must consider the hierarchical inclusion between categories as well as the similarity between categories. That is why we rely on the fuzzy relational products, which is able to describe the inclusive relation of two objects. The relationship of two categories can be decided by calculating the average degree of inclusion of a fuzzy set to another one using the fuzzy implication operator. An average degree of fuzzy set inclusion can be used for making a similarity relationship between two categories. By using this, similarity relations of

categories can be obtained dynamically [13, 16, 17].

Figure 1 shows e-mail classification relationship with respect to α-cut values in Table 3 by using dynamic category hierarchy method.



(a) Category classification relation with respect to α-cut value be 0.2



(b) Category classification relation with respect to α-cut value be 0.05

Fig. 1. Result of dynamic category hierarchy

Figure 1 (a) shows the diagram regarding the fuzzy hierarchical relations of categories in the case of $\alpha = 0.2$, t2 is a subcategory, and t5 is an independence category. Figure 1 (b) shows the diagram regarding the fuzzy hierarchical relations of category in the case of $\alpha = 0.05$, t4 is located at the highest of the hierarchy, while t1, t2, t3, t4, t5 are at the lowest. It can be indicated that the case in (b) has a broader diagram than the case of (a), including all relations in case of (a).

V. EXPERIMENTAL RESULTS

In this paper, we implemented the prototype using Visual Basic 6.0. We performed two experiments on the prototype. Experiment 1 evaluated the accuracy of e-mail classification using automatic category construction by PCA. Experiment 2 evaluated the accuracy of re-organize e-mail using dynamic category hierarchy method.

Experiment 1. Our test data is 150 e-mail messages that are received in the 1 month period. Experiment evaluated the accuracy of e-mail classification with respect to the constructed category label. The accuracy of classification was tested by manual decision whether messages belongs to category label or not. The average accuracy of e-mail classification is 78.99%.

Experiment 2. We evaluated the accuracy of re-organize according to α value. The test data is the result of the Experiment 1. We can know that re-organize email' accuracy improves to 78.99% to 90.6% with respect to α value be 0.04.

VI. CONCLUSIONS

In this paper we describe a method automatic classification of e-mail messages, and the architecture of a system that implement it. Our method, based on automatic construction of category label method and dynamic construction of category hierarchy method for e-mail classification and re-organization, was tested in two experiments, showing a high degree of flexibility, efficiency and effectiveness in the message classification and category re-organization. The proposed method in this paper has the following advantages:

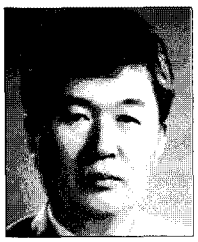
1) There is unsupervised classification method that the automatic construction of category labels and e-mail classification from a set of incoming messages.
2) Our method make user easy re-organizes all his e-mail messages and directory search by dynamic construction of category hierarchy.

REFERENCES

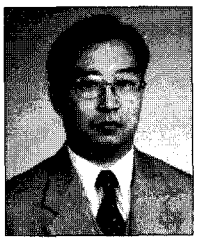
[1] W.W. Cohen. Learning Rules that classify E-mail. In Proc. AAAI Spring Symposium in Information Access, 1999.
[2] I. Androutsopoulos et al. An Evaluation of Naïve Bayesian Anti-Spam Filtering. In Proc. Workshop on Machine Learning in the New Information Age, 2000.
[3] G. Sakkis et al. Stacking classifiers for anti-spam filtering of e-mail. In Proc. 6th Conf. On Empirical Methods in Natural Language Processing, 2001.
[4] H. Drucker, D. Wu, and V. N. Vapnik, Support Vector Machines for Spam Categorization. IEEE Transactions on Neural network, 10(5), 1999.

[5] L. Kun-Lun, Li, Kai, H, Hou-Kuan, T. Sheng-Feng, Active Learning with Simplified SVMS for SPAM Categorization. In Proc. First Conf. On Machine Learning and Cybernetics, Beijing, 4-5, November, 2002.
[6] M. Woitaszek, M. shaaban. Identifying Junk Electronic Mail in Microsoft Outlook with a Support Vector Machine. In Proc. 2003 Symposium. On Application and the Internet. 2003.
[7] K. Mock. Dynamic Email Organization via Relevance Categories. In Proceedings of the International Conference on Tools with Artificial Intelligence 1999. Chicago IL, Nov. 1999.
[8] G. Manco, E. Masciari. A Framework for Adaptive Mail Classification. In Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence. 2002.
[9] T. M. Alrashid, J. A. Barker, B. S. Christian, S.C. Cox, M. W. Rabne, E. A. Slotta and L. R. Upthegrove. Safeguarding Copyrighted Contents, Digital Libraries and Intellectual Property Management, CWRU's Rights Management System. D-Lib Magazine, April 1998. http://www.dlib.org/dlib/april98/04barker.html.
[10] Y. Ogawa, T. Morita, and K. Kobayashi. A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy Sets and System, pp. 163-179, 1991.
[11] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 465-480, 1988.
[12] R. B. Y and B. R. N. Modern Information Retrieval. Addison Wesley, 1999.
[13] C. Bumghi, L. Ju-Hong, P. Sun. Dynamic Construction of Category Hierarchy Using Fuzzy Relational Products. In proceedings of the 4th International Conference On Intelligent Data Engineering and Automated Learing. Hong Kong, China, pp.296-302, 2003.
[14] S.S. Kang. Korean Information Retrieval and Morpheme analysis. HongReung Science Publishing Co., 2002.
[15] D. Zhang, Y. Dong. Semantic, hierarchical, online clusteing of web search results. In proceedings Asia Pacific Web Conference (APWEB), Hangzhou, China, pp. 67-78, 2004.
[16] B. G. Choi, T. S. Park, J. H. Lee, S. Park. Web Search Model for Dynamic and Fuzzy Directory Search. LNCS 3878. pp. 406-409. 2006.
[17] S. Park, S. H. Park, J. H. Lee, J. S. Lee. E-mail Classification Agent Using Category Generatoin and Daynamic Category Hierarchy. LNAI 3397. pp. 207-214. 2005.

Sun Park
Member KIMICS. Received the Ph.D degree in Computer & Information Engineering, Inha University in 2007. Since In 2009, he has been a post doctor in Division of Electronic & Information Engineering, Chonbuk National University, Korea. Prior to becoming a researcher at Chonbuk National University, he has worked for professor in Dept. of Computer Engineering, Honam University, Korea. His research interests include Data Mining, Information Retrieval, and Information Summarization.



Chul-Won Kim
Member KIMICS. Received the Ph.D degree in Computer Engineering, Kwangwoon University in 1997. He is a Professor at Honam University, . His research interests include XML Retrieval, Multimedia Information Retrieval, and Multimedia Processing.



Dong-Un An
Member KIMICS. He is a professor at Chonbuk National University, Korea. He received the Ph.D degree in Computer Engineering from KAIST in 1995, the M.S. degree in Computer Engineering from KAIST in 1987, and the B.S. degree in Electronic Engineering from Hanyang University in 1981. His research interests include Natural Language Processing, Information Retrieval, and Machine Translation.