주성분 분석과 동적 분류체계를 사용한 자동 이메일 분류

박선* · 김철원** · 이양원**

**호남대학교

Automatic e-mail classification using Dynamic Category Hierarchy and Principal Component Analysis

Sun Park* · Chul-Won Kim** · Yang-weon Lee

*Honam University

E-mail : sunpark@honam.ac.kr, cwkim@honam.ac.kr, ywlee@honam.ac.kr

요 약

인터넷 사용의 보편화로 이메일의 양이 급속히 증가하고 있다. 따라서 수신 메일을 효율적이면서 정확하게 분류할 필요성이 점차 증가하고 있다. 현재의 이메일 분류는 베이지안, 규칙 기반 등을 이용하여 스팸 메일을 필터링하기 위한 이원 분류가 주를 이루고 있다. 클러스터링을 이용한 다원 분류 방법은 분류의 정확도가 떨어지는 단점이 있다. 본 논문에서는 주성분 분석(PCA, Principal Component Analysis)을 기반으로 한 자동 카테고리 생성 방법과 동적 분류 체계 방법을 결합한 새로운 자동 이메일 분류 방법을 제안한다. 이 방법은 수신되는 이메일을 자동으로 분류하여 대량의 메일을 효율적으로 관리할 수 있으며, 메일을 동적으로 재분류 하여 분류 정확률을 높일 수 있다.

ABSTRACT

The amount of incoming e-mails is increasing rapidly due to the wide usage of Internet. Therefore, it is more required to classify incoming e-mails efficiently and accurately. Currently, the e-mail classification techniques are focused on two way classification to filter spam mails from normal ones based mainly on Bayesian and Rule. The clustering method has been used for the multi-way classification of e-mails. But it has a disadvantage of low accuracy of classification. In this paper, we propose a novel multi-way e-mail classification method that uses PCA for automatic category generation and dynamic category hierarchy for high accuracy of classification. It classifies a huge amount of incoming e-mails automatically, efficiently, and accurately.

e-mail classification, PCA(principal component analysis), dynamic category hierarchy

이메일 분류, 주성분 분석, 동적 분류 체계

Ⅰ. 서 론

현재 개인 및 회사에서 하루에 받는 이메일의 양은 수십여 통에서 수백 통에 이른다. 이중 스팸 메일이 대부분을 차지하고 있다. 스팸 메일을 효율적으로 차단할 수 있는 많은 도구들이 개발되어 왔으나, 대부분은 사용자가 직접 필터링 규칙을 만들거나 메일을 분류할 색인어 목록을 작성해야 한다. 이런 도구들은 색인어를 많이 포함하는 대량의 메일을 분류해야 할 경우 효율성과 정확성이 떨어지는 단점이 있다. 또한 사용자의 변화되는 요구사항에 맞추어 재분류 하거나 재 필터링할 수 없는 단점이 있다.

이메일 분류는 대부분 스팸 메일을 구분하는 이원분류가 주로 연구되었다. 스팸 분류를 위해 사용된 방법으로는 베이시안, 규칙기반 등이 있다. Androutsopoulos [3]와 Sakkis[6]은 베이시안 분류

자를 이용하였고, Cohen[9]은 텍스트 마이닝을 이용한 규칙기반 분류방법을 제안하였다. 이들의 방법은 사용자가 직접 메시지 폴더를 만들어야 하며, 학습단계가 필요한 문제가 있다.

또 다른 연구로는 수신된 메일 집합으로부터 메일 폴더를 자동으로 구성하여 이메일을 분류한다. Manco[2]는 군집 기술과 데이터마이닝 알고리즘을 이용하였으며, Mock[4]은 벡터모델을 기반으로 한 이메일의 분류시스템을 제안하였다. 이 방법들은 여러 단계의 전 처리와 다양한 특징 정보로부터 유사도를 얻기 때문에 효율성과 정확도가 떨어지는 단점이 있다.

본 논문에서는 위의 단점을 해결하기 위해 주성분 분석과 동적 분류체계 방법을 사용한 이메일 분류 방법을 제안한다. 주성분 분석을 사용하여 분류 주제를 자동으로 생성하고, 생성된 분류주제로 이메일을 자동적으로 분류한다. 분류결과의 정확도가 떨어지는 문제를 해결하기 위해 동적 분류 체계 방법을 이용하여 이메일 분류 체계를 동적으로 재구성 할 수 있게 하였다.

본 논문에서 제안한 방법은 다음과 같은 장점을 가진다. 첫째, 제시된 방법에 의해 메일의 분류 주제가 자동으로 생성됨으로써 사용자의 간섭이 필요 없다. 둘째, 동적 분류 체계 방법을 이용하여 사용자가 필요하면 언제든지 재분류 할 수 있다. 셋째, 메일분류에 대한 훈련 및 학습 과정이 필요 없어 메일을 수신 받는 즉시 분류할 수 있으므로 유동적인 이메일 환경에 적합하다.

본 논문의 구성은 다음과 같다. 2장에서는 주성분 분석 방법을 설명하고, 3장에서는 동적 분류 체계 방법을 설명한다. 4장에서는 제안된 자동 이메일 분류 방법을 보이고, 동적 분류 체계를 적용하여 분류 체계를 동적으로 재구성하는 방법을 보인다. 5장에서는 실험 및 분석결과를 보이고, 6장에서 결론을 맺는다.

II. 주성분 분석

p개의 확률특징 $X_1, X_2, \ldots, X_p$를 원소로 하는 확률특징벡터 X가 평균 벡터 $\overline{x}$와 공분산 행렬 S($p\times p$)를 갖는다고 하고, 이들을 다음의 기호로 나타내자.

$$X=\begin{vmatrix} X_1 \\ X_2 \\ \cdots \\ X_p \end{vmatrix} \quad \overline{x}=\begin{vmatrix} \overline{x_1} \\ \overline{x_2} \\ \cdots \\ \overline{x_p} \end{vmatrix}$$

$$S=\begin{vmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{vmatrix}$$

단, $s_{ik}=\frac{1}{n}\sum_{j=1}^{n}(x_{ij}-\overline{x_i})(x_{kj}-\overline{x_k})$, X_i와 X_k의 공분산, $\overline{x_i}=\frac{1}{n}\sum_{j=1}^{n}X_{ij}$, X_i의 산술평균이다. $=s_{ki}$

주성분 분석은 원래 특징벡터 X를 적절히 선형 변환 시켜 원본자료의 성질을 유지하면서 자료를 축소하고 해석하는데 사용한다. 이 선형변환은 X의 원소들 간의 상관구조관계를 나타내는 S를 분석대상으로 하며, S는 $\overline{x}$의 값의 변화에 의한 영향을 받지 않는다. 우선 S의 p개의 고유값(eigen value) λ_j 들을 크기 순으로 배열하고 각각의 고유값에 대응되는 고유벡터(eigen vector) e_j의 짝들을 $(\lambda_1, e_1), \ldots, (\lambda_j, e_j)$라 하고 λ_j들의 크기 순으로 배열하면, $Se_j=\lambda_j e_j,\quad j=1,2,\ldots,p$, $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_j$와 같은 관계가 있으며, 이를 행렬 기호를 이용하여 전체적으로 표현하면 다음과 같다.

$$SP=P\Lambda,\ S=P\Lambda P' \qquad (1)$$

여기서 P는 p개의 고유벡터 e_i들로 구성된 크기 $p\times p$ 직교행렬이고, Λ는 λ_i를 i번째 대각원소, 그리고 모든 비대각 원소가 0인 크기 $p\times p$의 대각행렬, 그리고 P'는 P의 전치행렬이다.

즉, $P=(e_1, e_2, \ldots, e_p)$, $\Lambda=Diagonal(\lambda_1, \lambda_2, \ldots \lambda_p)$ 이와 같은 P를 이용하여 다음과 같은 X의 직교변환을 생각할 때, $\Phi'=P'X$ 이 변화에 의해 새로이 창조되는 벡터 $\Phi'=(\Phi_1, \Phi_2, \ldots, \Phi_p)$를 X의 주성분이라 정의한다. 이때 j번째 고유값 λ_j에 대응하는 고유벡터 e_j의 원소들을 X와의 선형결합에서 가중계수로 사용하고 있다. 즉, Φ'의 j번째 원소 Φ_j를 X의 j번째 주성분이라고 하고 다음과 같이 나타낸다.

$e'_j=(e_{1j}, e_{2j}, \ldots, e_{pj})$, $j=1,2,\ldots,p$ 일때,

$$\Phi_j=e_jX=e_{1j}X_1+e_{2j}X_2+\ldots e_{pj}X_p=\sum_{i=1}^{p}e_{ij}X_j \qquad (2)$$

위와 같이 주성분 분석이란 전체 자료의 공분산 행렬의 구조를 파악하여 고유 값이 큰 고유벡터들의 축으로 자료의 축을 변환하여 주성분을 구하는 분석이다.[5,11]

III. 동적 분류체계 방법

동적 분류 체계 방법[1]은 검색어와 분류간의 관계를 규정하고, 분류들 간의 상호 관계를 규명하여, 분류 검색의 분류 체계를 재구성함으로써 검색 효율을 높이는 방법이다.

동적 분류 체계 방법에서 사용되는 퍼지 이론은 다음과 같다[10]. 퍼지 함의 연산자 (Fuzzy Implication Operator) 는 [0,1]×[0,1]→[0,1]로서 단위 구간의 다치 논리로 확장된 것이다. 퍼지 함의 연산자의 종류는 무수히 많으며 대표적인 Kleene-Diense 퍼지함의 연산자는 다음과 같다[1].

$$a \rightarrow b = (1 - a) \vee b = \max(1 - a, b), \quad a = 0\sim1,\ b = 0\sim1 \tag{3}$$

(정의) 퍼지 함의 연산자는 주어진 문제의 범주에 따라 달라진다. $a \in U_1$에 대한 후위집합 (afterset) aR 는 a와 연관된 $y \in U_2$로 구성된 U_2 의 퍼지 부분집합이며 그 멤버쉽 함수는 $\mu_{aR}(y) = \mu_R(a,y)$로 주어진다. $c \in U_3$ 에 대한 전위집합 (foreset) Sc는 c에 연관된 $y \in U_2$ 로 구성된 U_2 의 퍼지 부분집합이며 그 멤버쉽 함수는 $\mu_{Sc}(y) = \mu_S(y,c)$로 주어진다. aR 이 Sc 의 부분집합인 평균정도는 $y \in aR$의 멤버쉽 정도가 $y \in Sc$의 멤버쉽 정도를 함의하는 평균정도로서 다음과 같이 정의된다.

$$\pi_m(aR \subseteq Sc) = \frac{1}{N} \sum_{y \subset U_2} (\mu_{aR}(y) \rightarrow \mu_{Sc}(y)) \tag{4}$$

여기서 π_m은 평균 정도를 나타내는 함수이다 [8].

본 논문에서는 위의 식3의 Kleen-Diense 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식4의 퍼지관계곱을 적용하여 분류들 간의 퍼지함의 관계, $C_i \rightarrow C_j$를 유도할 수 있다. 그러나 C_i 에 멤버쉽 값($\mu_{Ci}(x)$)이 작은 원소 x가 많으면, $C_i \subseteq C_j$의 포함여부와 관계없이 항상 1에 가까운 값이 나오는 문제점이 있다. 따라서 다음과 같이 정의하여 두 분류 퍼지 집합의 함의 관계, $\mu_{m,\beta}(C_i \subseteq C_j)$를 계산한다.

$$\mu_{m,\beta}(C_i \subseteq C_j) = (R^T \triangle_\beta R)_{ij} = \frac{1}{|C_{i\beta}|} \sum_{K_{i\beta} \in C_{i\beta}} (R^T_{ik} \rightarrow R_{kj}) \tag{5}$$

여기서, K_k 는 k번째 검색어이고, C_i , C_j는 i번째와 j번째 분류이며, $C_{i\beta}$는 C_i의 β-제약, $\{ x \mid \mu_{Cj}(x) \geq \beta \}$이고 $|C_{i\beta}|$는 $C_{i\beta}$의 원소의 갯수이다. R은 $m \times n$행렬로서 R_{ij}는 $\mu_{Cj}(K_i)$,즉, $K_i \in C_j$인 정도이다. R^T는 행렬 R의 전치 행렬로서 $R_{ij} = R^T_{ji}$이다.

IV. 자동 이메일 분류방법

본 논문에서 제안한 이메일의 분류 과정은 다음과 같다. 첫째, 수신 메일에서 색인어를 추출한다. 둘째, 메일과 색인어의 출현 빈도를 이용하여 메일-색인어 행렬을 구성한다. 구성된 메일-색인어 행렬로부터 공분산 행렬을 만든다. 셋째, 공분산 행렬에 주성분을 사용하여 색인어의 주성분에 대한 행렬을 만든다. 색인어-주성분 행렬에서 누적 빈도가 95% 이하인 주성분을 선택한다. 선택한 주성분에서 최고 값을 가지는 색인어들을 분류 주제를 추출한다. 넷째, 제안한 방법을 이용하여 이메일을 분류 주제별로 분류한다. 마지막으로 사용자의 필요에 따라 동적 분류 체계 방법을 이용하여 분류 주제를 재구성 한다.

이메일을 자동으로 분류하는 방법은 다음과 같다. 모든 이메일 에 대해서 각각의 이메일에 대한 색인어를 추출한다. 그 후, 추출된 색인어들에 대해서 각각 주성분 분석으로 추출된 분류 주제어와 일치하는 색인어가 있는지 비교한다. 만일 분류 주제어에 포함되는 색인어가 있다면 그 이메일은 분류 주제어로 분류한다. 만일 분류 주제어와 일치하는 색인어가 하나도 없는 이메일은 기타 메일로 분류한다.

이메일을 동적으로 재구성하기 위해서는 색인어와 분류 주제 간의 관계를 규정해야 한다. 그러나 색인어와 분류 주제 간의 관계를 직접 결정할 수는 없으므로 색인어와 메일간의 관계 및 메일과 분류 주제 간의 관계에 의해서 결정한다.

메일을 색인어로 구성된 퍼지 집합으로 간주할 수 있고, 마찬가지 분류 주제를 분류된 메일들의 색인어들로 구성된 퍼지 집합으로 간주할 수 있다. 메일이 속한 두 분류 주제 간의 관계는 생성된 두 분류 주제의 퍼지 집합의 함의 정도를 식(3), (5)를 이용하여 계산하여 결정할 수 있다. 두 퍼지 집합의 함의 정도는 퍼지 함의 연산자를 이용하여 한 퍼지 집합이 다른 퍼지 집합에 포함되는 정도를 계산하여 구할 수 있고, 이를 이용하여 서로 다른 두 분류주제의 유사관계를 동적으로 생성할 수 있다.

V. 실험 및 분석

실험 자료는 2009년 2월 3일부터 2009년 3월 10일까지 수신된 메일중에서 분류 주제와는 상관없이 임의로 150개의 메일을 선택하였다. 평가는 수작업으로 분류된 메일을 제안된 방법과 비교한 정확률을 분석하였다. 이때 분류주제는 메일에 포함된 단어로 한정하였다. 수작업으로 분류하기 위한 10개의 분류주제를 선택하였다.

여러 번의 실험 결과 누적 빈도율 90% 일 때 구성된 분류 주제로 분류했을 때 메일들이 가장 폭넓게 분류될 수 있었다. 분류 주제별 분류 정확률과 평균 분류 정확률을 분석의 평가 방법으로

사용하였다. 분류정확률은 수작업으로 분류한 메일과 자동 분류한 메일을 비교하여 바르게 분류된 메일의 정확률을 계산하였다.

그림 1은 동적 분류 체계방법을 사용하여 분류주제 관계를 재구성했을 때의 평균 분류 정확률을 나타낸다. 여기서는 α값에 따른 평균 분류 정확률의 분포를 나타낸다. 실험 결과에서 $\alpha = 0.04$일때 분류 주제 체계를 재구성하면 78.99%에서 90.6%로 평균 분류 정확률이 높아짐을 알 수 있다.

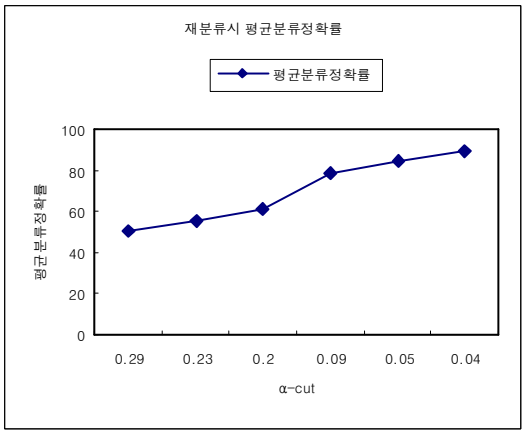

그림 1. 재분류시 평균 분류 정확률

VI. 결 론

본 논문에서는 이메일을 자동으로 분류하고 분류된 결과를 사용자의 요구사항에 맞게 재분류할 수 있는 방법을 제안하였다. 제안된 방법은 주성분 분석으로 메일에 포함된 색인어 중에서 분류 주제어를 추출하고, 분류 주제와 높은 연관도를 갖는 메일들을 분류한다. 이렇게 분류된 메일도 사용자의 요구에 따라 언제든지 동적 분류 체계 방법을 이용해서 분류주제의 체계를 재구성할 수 있다. 이러한 재구성은 사용자의 요구사항에 맞추어 조절할 수 있도록 하여 효율적으로 이메일을 관리할 수 있다. 사용자의 요구사항에 맞게 구성된 분류주제어는 사용자가 쉽게 이메일을 분류할 수 있도록 한다. 마지막으로 분류규칙에 대한 별도의 훈련 및 학습 과정이 필요 없이 이메일을 빠르게 분류함으로서 유동적인 이메일 환경을 만족시킨다.

참고문헌

[1] B.G. Choi, J. H. Lee, S. Park Dynamic Construction of Category Hierarchy Using Fuzzy Relational Products. IDEAL 2003, pp.296-302, 2003.

[2] G. Manco, E. Masciari. A Framework for Adaptive Mail Classification. In Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence. 2002.

[3] I. Androutsopoulos .... An Evaluation of Naive Bayesian Anti-Spam Filtering. In Proc. Workshop on Machine Learning in the New Information Age, 2000.

[4] K. Mock. Dynamic Email Organization via Relevance Categories. In Proceedings of the International Conference on Tools with Artificial Intelligence 1999. Chicago IL, Nov. 1999.

[5] Richard A. Johnson, Dean W. Wichern, Applied Multivariate Statistical Analysis 4th ed., Prentice hall, 1998

[6] G. Sakkis et al. Stacking classifiers for anti-spam filtering of e-mail. In Proc. 6thConf. On Empirical Methods in Natural Language Processing, 2001.

[7] S.S. Kang. Korean Information Retrieval and Morpheme analysis. HongReung Science Publishing Co., 2002.

[8] W. Bandler and L. Kohout. Semantics of Implication Operators and Fuzzy Relational Products. International Journal of Man-Machine Studies. Vol. 12, pp.89-116, 1980.

[9] W.W. Cohen. Learning Rules that classify e-mail. In Proc. AAAI Spring Symposium in Information Access, 1999.

[10] D. R. Radev, H. Jing, and M. Stys-Budzikowska. Summarization of multiple documents: clustering sentence extraction, and evaluation, In proceddings of ANLPNAACL Workshop on Automatic Summarization. 2000.

[11] 이창범,김민수,이기호,이귀상,박혁로, 주성분 분석을 이용한 문서 주제어 추출, 정보과학회논문지 : 소프트웨어 및 응용 제 29권 제 10호 (2002.10)