비음수 행렬 분해와 동적 분류 체계를 사용한 자동 이메일 다원 분류

(Automatic Email Multi-category Classification Using Dynamic Category Hierarchy and Non-negative Matrix Factorization)

박 선 † 안 동 언 ††
(Sun Park) (Dong Un An)

요 약 이메일 사용의 증가로 수신 메일을 효율적이면서 정확하게 분류할 필요성이 점차 늘고 있다. 현재의 이메일 분류는 SVM, 베이지안 분류자, 규칙 기반 분류자 등을 이용하여 스팸 메일을 필터링하기 위한 이원 분류가 주를 이루고 있다. 그러나 이러한 지도 학습 방법들은 적합한 이메일을 인식하기 위하여서 사용자가 규칙이나 색인어 목록을 작성해야 한다. 비지도 학습 방법으로 군집을 이용한 다원 분류 방법은 메일의 분류 주제를 설정해주어야 한다. 본 논문에서는 비음수 행렬 분해(NMF, Non-negative Matrix Factorization)를 기반으로 한 자동 분류 주제 생성 방법과, 동적 분류 체계(DCH, Dynamic Category Hierarchy) 방법을 이용한 분류 주제 내에 이메일을 재구성하는 방법을 결합한 새로운 이메일 다원 분류 방법을 제안한다. 이 방법은 수신되는 이메일을 자동으로 다원 분류하여 대량의 메일을 효율적으로 관리할 수 있으며, 사용자가 분류 결과를 만족하지 못하면 분류 주제 내의 이메일을 동적으로 재구성하여 분류의 정확률을 높인다.

키워드 : 이메일 다원 분류, 비음수 행렬 분해, 동적 분류 체계

Abstract The explosive increase in the use of email has made to need email classification efficiently and accurately. Current work on the email classification method have mainly been focused on a binary classification that filters out spam-mails. This methods are based on Support Vector Machines, Bayesian classifiers, rule-based classifiers. Such supervised methods, in the sense that the user is required to manually describe the rules and keyword list that is used to recognize the relevant email. Other unsupervised method using clustering techniques for the multi-category classification is created a category labels from a set of incoming messages. In this paper, we propose a new automatic email multi-category classification method using NMF for automatic category label construction method and dynamic category hierarchy method for the reorganization of email messages in the category labels. The proposed method in this paper, a large number of emails are managed efficiently by classifying multi-category email automatically, email messages in their category are reorganized for enhancing accuracy whenever users want to classify all their email messages.

Key words : email multi-category classification, NMF, dynamic category hierarchy

· 이 논문은 2009 한글 및 한국어 정보처리 학술대회에서 '비음수 행렬 분해와 동적 분류체계를 사용한 이메일 분류'의 제목으로 발표된 논문을 확장한 것임

† 정 회 원 : 전북대학교 전자정보고급인력양성사업단 포닥
sunbak@jbnu.ac.kr
†† 종신회원 : 전북대학교 전자정보공학부 교수
duan@jbnu.ac.kr
(Corresponding author임)

논문접수 : 2009년 10월 26일
심사완료 : 2010년 3월 3일

정보과학회논문지: 소프트웨어 및 응용 제37권 제5호(2010.5)

1. 서 론

인터넷을 기반으로 한 이메일 사용의 증가는, 개인 우편물로부터 기업의 업무 및 광고 등 다양한 분야의 우편물을 이메일로 교체하였다. 이러한 이메일 사용의 폭발적인 증가로 인하여서 개인들은 수 통에서 수십 통의 이메일을, 기업들은 수백에서 수천 통의 이메일을 매일 받고 있다. 그러나 수신되는 메일의 대부분은 스팸 메일이 차지하거나 일정 기간 동안에 지속적으로 중복되는 내용을 포함하고 있다.

이 때문에 수신되는 메일을 효율적으로 관리할 수 있는 많은 도구들이 개발되었다. 그러나 대부분의 도구들

은 사용자가 직접 필터링 규칙을 설정하거나, 메일을 분류할 색인어 목록을 작성해야 한다. 이렇게 사용자의 규칙이 들어가는 도구들은, 색인어를 많이 포함하는 대량의 메일을 분류해야 할 경우 분류의 정확성이 낮아지거나, 사용자의 변화되는 요구 사항에 맞추어 재분류하거나 재 필터링할 수 없는 경우도 있다.

지금까지의 이메일 분류에 대한 연구로는 지도 학습 방법을 이용한 이원 분류로 스팸 메일을 구분하는 방법과 비지도 학습 방법으로 군집 기법을 이용한 다원 분류가 연구되고 있다.

지도 학습 방법으로는 SVM[1-3], 베이지안 규칙[4,5], 이진 결정 트리[6], 신경망[7], 규칙 기반[8]등을 이용한 방법이 연구되고 있으며, 이들 방법은 주로 스팸 메일을 분류하는데 사용된다. 지도 학습 방법들은 사용자가 원하는 필터링 규칙을 미리 설정하거나 학습 단계를 거쳐야하는 단점을 가지고 있다.

비지도 학습 방법은 데이터 마이닝[9,10]이나 군집[11]등을 이용한 다중 분류 방법이 연구되고 있다. 이들 방법은 유사한 특성을 가진 이메일들로 군집하기 때문에 군집의 특성을 내타내는 분류 주제를 파악할 수 없다. 이 때문에 사용자들이 수신되는 이메일을 효율적으로 관리할 수 없는 문제를 가지고 있다. 또한, 분류 주제를 자동으로 생성하더라도, 이것 역시 준지도 방법을 이용하기 때문에 분류에 대한 학습과 평가가 필요하다.

본 논문에서는 위의 단점을 해결하기 위해 비음수 행렬 분해와 동적 분류 체계 방법을 사용하여서 이메일을 다원 분류하고 재분류하는 방법을 제안한다. 비음수 행렬 분해(NMF, non-negative matrix factorization)는 Lee와 Seung이 제안한 방법으로 원본 자료를 두 개의 의미 특징 행렬로 분해하여 부분 정보의 선형 조합으로 원본자료를 표현할 수 있는 방법이다. 이 방법은 정보 구조 안에 포함하고 있는 의미 특징을 쉽게 파악할 수 있고, 차원 축소를 통하여서 대량의 정보를 효율적으로 표현할 수 있게 한다[12,13]. 동적 분류 체계 방법[14]은 최범기 외 저자가 제안한 방법으로 검색어와 분류간의 관계를 규정하고, 분류들 간의 상호 관계를 규명하여, 분류 검색의 분류 체계를 재구성함으로써 검색 효율을 높이는 방법이다.

본 논문에서 제안한 방법은 다음과 같은 장점을 가진다. 첫째, 비음수 행렬 분해에 의해서 분해된 의미 특징을 이용하여 이메일의 분류 주제를 생성하고, 분류 주제에 따라서 자동으로 이메일을 다원 분류 한다. 이 때문에, 사용자의 간섭 및 학습 과정이 없어서 메일을 수신받는 즉시 분류할 수 있으므로 유동적인 이메일 환경에 적합하다. 둘째, 동적 분류 체계 방법을 이용하여 사용자가 필요하면 언제든지 이메일을 동적으로 재분류할 수 있게 하여서 다원 분류 결과 정확도가 떨어지는 문제를 해결하였다. 즉, 사용자가 자동으로 분류된 이메일의 다원 분류 주제로부터 원하는 이메일을 찾을 수 없으면, 분류 주제 내의 이메일을 재구성하여서 원하는 이메일을 쉽게 찾을 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 기존의 이메일 분류에 대한 연구 및 비음수 행렬 분해와 동적 분류 체계 방법을 설명한다. 3장에서는 비음수 행렬 분해를 이용한 자동 이메일 분류 방법과 동적 분류 체계방법을 이용하여 다원 분류된 이메일을 동적으로 재구성하는 방법을 보인다. 4장에서는 실험 및 분석 결과를 보이고, 5장에서 결론을 맺는다.

2. 관련연구

2.1 이메일 분류에 대한 기존 연구

이메일 분류에 대한 연구는 지도 학습 방법을 이용한 스팸 메일 분류와 비지도 학습 방법을 이용한 이메일 다원 분류로 구분할 수 있다.

지도 학습 분류(supervised classification) 방법을 이용한 이메일 분류에 대한 연구는 다음과 같다. Drucker[1]는 SVM을 이용하여 이메일의 스팸과 스팸이 아닌 것을 분류하는데 Ripper, Rocchio, boosting decision trees 등의 세 알고리즘과 비교하여 SVM이 가장 성능이 좋은 것을 보였다. Kunlun[2]는 스팸을 분류하기 위해 활성 학습 정책을 이용하는 SVM 기반의 새로운 방법을 제안하였다. Woitaszek[3]는 simple SVM을 이용하여 상업적 이메일 분류 시스템을 만들었다. Androutsopoulos[4]와 Sakkis[5]는 안티 스팸 필터링을 하기 위해 베이지안 분류자를 이용하였다. 그들의 접근 방법은 규칙기반 분류자를 사용하는 것에 비해 좀 더 좋은 정확성을 보였다. Xia[6]등은 이진 결정 트리를 이용하여 메일을 분류하는 방법을 제안하였다. Yu[7]등은 신경망과 의미 특징을 이용한 스팸 이메일을 분류하는 방법을 제안하였다. Cohen[8]은 이메일을 분류하기 하기 위해 텍스트 마이닝 기법을 이용한 두개의 규칙기반 시스템을 제안하였다. 이러한 위의 지도 학습 분류 방법은 수신된 메시지가 들어갈 비슷한 폴더를 찾을 수 있도록 사용자가 직접 메시지 폴더를 설정해야 하고, 분류하기 이전에 일정량 이상의 학습이 필요하다. 또한, 분류를 위한 학습과 평가를 위해서 어느 정도 시간이 걸리는 단점이 있다.

비지도 학습 분류(unsupervised classification) 방법들로는 다음과 같다. Mock[9]는 벡터 공간 모델의 역색인 방법에 기반을 둔 이메일 자동 분류 시스템을 제안하였다. Manco[10]는 이메일 관리 및 유지하기 위하여 데이터 마이닝 알고리즘에 기반을 두고 수신 메일을 분

류하였다. Cutting[11]은 군집 방법과 베이지안 분류자를 이용하여 분류 주제를 생성하고, 주제별로 메일을 분류하는 방법을 제안하였으나, 분류에 대한 학습과 평가가 필요하다. 박선 외 저자들은 벡터 공간 모델[15,16], PCA[17,18], NMF[19]를 이용하여 분류 주제를 자동으로 생성하여 분류하는 방법과, 동적 분류 체계를 이용하여 분류 결과를 재구성하는 방법을 제안하였다.

2.2 비음수 행렬 분해

본 논문에서 행렬 X의 j번째 열벡터는 X_{*j}로, i번째 행벡터는 X_{i*}로, i번째 행과 j번째 열의 원소는 X_{ij} 표시한다.

비음수 행렬 분해는 주어진 양의 행렬로부터 양의 인수를 찾아내는 행렬 분해 알고리즘이다[12,13]. 비음수 행렬 분해는 문서집합이 k개의 군집으로 구성된다고 가정할 때, 행렬 X를 식 (2)의 목적 함수가 최소값을 갖도록 식 (1)과 같이 $m\times k$ 비음수 의미 특징 행렬(NSFM, non-negative semantic feature matrix) W와 $k\times n$ 비음수 의미 변수 행렬(NSVM, non-negative semantic variable matrix) H로 분해한다.

$$X \approx WH \tag{1}$$

$$J = \frac{1}{2}\| X - WH \| \tag{2}$$

여기서 $W = [w_{ij}]$이고 $H = [h_{ij}]$ 이며 $W = [W_1, W_2, \cdots, W_k]$이다. W와 H의 원소 값을 갱신하기 위하여 목적 함수 J 값이 수렴 허용오차 보다 작아지거나 지정한 반복 횟수를 초과할 때까지 식 (3)과 식 (4)를 반복한다[12,13].

$$w_{ij} \leftarrow w_{ij} \frac{(XH^T)_{ij}}{(WHH^T)_{ij}} \tag{3}$$

$$h_{ij} \leftarrow h_{ij} \frac{(W^T X)_{ij}}{(W^T WH)_{ij}} \tag{4}$$

여기서 H^T는 H의 전치행렬이고, W^T는 W의 전치 행렬이다.

예 1. 다음은 식 (3)과 식 (4)를 이용하여 A행렬을 W와 H행렬로 분해 한 예이다. 다음의 예는 MATLAB 7.8.0의 *nnmf*( )함수를 이용하였고, 의미 특징의 개수 r은 3으로 설정하였다.

$$\underbrace{\begin{bmatrix} 1 & 0 & 3 & 20 & 1 \\ 0 & 0 & 2 & 0 & 30 \\ 1 & 3 & 5 & 1 & 9 \\ 0 & 0 & 1 & 1 & 2 \\ 1 & 2 & 2 & 0 & 0 \end{bmatrix}}_{A} \approx \underbrace{\begin{bmatrix} 0.06 & 20.27 & 0 \\ 30.07 & 0 & 0 \\ 8.98 & 1.03 & 5.27 \\ 1.977 & 1.05 & 0.54 \\ 0 & 0 & 2.96 \end{bmatrix}}_{W} \times \underbrace{\begin{bmatrix} 0 & 0 & 0.07 & 0 & 0.99 \\ 0.05 & 0 & 0.15 & 0.99 & 0.05 \\ 0.22 & 0.59 & 0.78 & 0 & 0 \end{bmatrix}}_{H}$$

비음수 행렬 분해된 의미 특징들에 의하여 원본 행렬을 다음과 같이 표현할 수 있다. 즉, 행렬 A의 j번째 열벡터 A_{*j}는 행렬 W의 l번째 열벡터 W_{*l}와 행렬 H의 요소 H_{kj}^T가 선형 조합을 이루며 식 (5)과 같다.

$$A_{*j} = \sum_{l=1}^{k} H_{jk}^T W_{*l} \tag{5}$$

예 2. 다음은 예 1의 A행렬의 첫 번째 열벡터를 식 (5)와 같이 의미 특징과 의미 변수의 곱을 선형 합으로 표현한 예이다.

$$\underbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}}_{A_{*1}} \approx 0 \times \underbrace{\begin{bmatrix} 0.06 \\ 30.07 \\ 8.98 \\ 1.977 \\ 0 \end{bmatrix}}_{W_{*1}} + 0.05 \times \underbrace{\begin{bmatrix} 20.27 \\ 0 \\ 1.03 \\ 1.05 \\ 0 \end{bmatrix}}_{W_{*2}} + 0.22 \times \underbrace{\begin{bmatrix} 0 \\ 0 \\ 5.27 \\ 0.54 \\ 2.96 \end{bmatrix}}_{W_{*3}}$$
$$H_{11} \qquad H_{21} \qquad H_{31}$$

2.3 동적 분류 체계 방법

동적 분류 체계 방법[3]에서 사용되는 퍼지 이론[1]은 다음과 같다. 퍼지 함의 연산자(Fuzzy Implication Operator) 는 $[0,1]\times[0,1]\rightarrow[0,1]$로서 단위 구간의 다치 논리로 확장된 것이다. 퍼지 함의 연산자의 종류는 무수히 많으며 대표적인 Kleene-Diense 퍼지함의 연산자는 다음과 같다[20].

$$a \rightarrow b = (1-a) \vee b = \max(1-a, b),$$
$$a = 0\sim1 \ , \ b = 0\sim1 \tag{6}$$

정의. 퍼지 함의 연산자는 주어진 문제의 범주에 따라 달라진다. $a \in U_1$에 대한 후위 집합(afterset) aR는 a와 연관된 $y \in U_2$로 구성된 U_2의 퍼지 부분 집합이며 그 멤버십 함수는 $\mu_{aR}(y) = \mu_R(a,y)$로 주어진다. $c \in U_3$에 대한 전위 집합 (foreset) Sc는 c에 연관된 $y \in U_2$로 구성된 U_2의 퍼지 부분 집합이며 그 멤버십 함수는 $\mu_{Sc}(y) = \mu_S(y,c)$로 주어진다. aR과 Sc의 부분 집합인 평균 정도는 $y \in aR$의 멤버십 정도가 $y \in Sc$의 멤버십 정도를 함의하는 평균 정도로서 다음과 같이 정의된다.

$$\pi_m(aR \subseteq Sc) = \frac{1}{N} \sum_{y \subset U_2} (\mu_{aR}(y) \rightarrow \mu_{Sc}(y)) \tag{7}$$

여기서 π_m은 평균 정도를 나타내는 함수이다[1].

본 논문에서는 위의 식 (6)의 Kleen-Diense 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식 (7)의 퍼지 관계곱을 적용하여 분류들 간의 퍼지함의 관계, $C_i \rightarrow C_j$를 유도할 수 있다. 그러나 C_i에 멤버십 값($\mu_{Ci}(x)$)이 작은 원소 x가 많으면, $C_i \subseteq C_j$의 포함 여부와 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.

$$\mu_{m,\beta}(C_i \subseteq C_j) = (R^T \triangle_\beta R)_{ij} \tag{8}$$
$$= \frac{1}{|C_{i\beta}|} \sum_{K_{ik} \in C_{i\beta}} (R_{ik}^T \rightarrow R_{kj})$$

여기서, K_k는 k번째 검색어이고, C_i, C_j는 i번째와 j번째 분류이며, $C_{i\beta}$는 C_i의 β-제약, $\{x|\mu_{Cj}(x) \geq \beta\}$이고 $|C_{i\beta}|$는 $C_{i\beta}$의 원소의 개수다. 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.

3. 자동 이메일 분류 방법

본 논문에서 제안한 이메일의 분류 과정은 다음과 같다. 첫째, 수신 메일에서 색인어를 추출한다. 둘째, 메일과 색인어의 출현 빈도를 이용하여 용어-이메일 빈도 행렬을 구성한다. 비음수 행렬 분해를 이용하여 구성된 용어-이메일 빈도 행렬로부터 이메일 분류 주제 생성 및 다원 분류를 한다. 마지막으로 사용자의 필요에 따라 동적 분류 체계 방법을 이용하여 분류 주제 내의 이메일을 재구성한다. 다음 그림 1은 제안 시스템으로 자동 이메일 다원 계층 분류 및 이메일 분류 계층 재구성 방법을 보여준다[19].

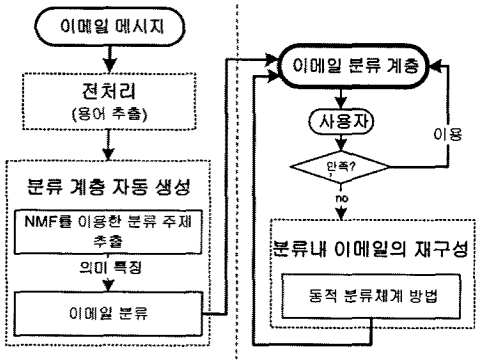


그림 1 제안된 이메일 다원 분류 시스템

3.1 비음수 행렬 분해에 의한 자동 이메일 분류 방법

본 논문에서는 이메일의 분류 주제 추출을 위해서는 표 2(a)와 같이 비음수 행렬 분해에 의해서 생성된 의미 특징을 이용하고, 분류 주제에 따라서 이메일을 자동으로 분류하는 방법으로 표 2(b)와 같이 Xu등[21]이 제안한 비음수 행렬 분해의 의미 변수 행렬 H를 이용한 문서 군집 방법을 사용한다.

예 3. 다음 표 1부터 표 3까지는 8개의 이메일로부터 4개의 분류 주제로 분류한 예를 나타낸다. 표 1은 8개의 이메일부터 7개의 용어를 사용한 용어-이메일 빈도 행렬을 나타내며, 표 2는 표 1에 MATLAB 7.8.0의 *nnmf*()함수를 이용 하여 비음수 행렬 분해된 비음수 특징 행렬 W와 비음수 의미 변수 행렬 H를 나타낸다. 예에서는 의미 특징 (분류 주제) r의 개수를 4로 설정하였다. 표 3은 표 2로부터 분류 주제와 분류 주제로 이메일

표 1 8개의 이메일의 용어-이메일 빈도행렬 A

용어 \ 이메일	*e1*	*e2*	*e3*	*e4*	*e5*	*e6*	*e7*	*e8*
t1	0	0	5	4	1	0	0	0
t2	0	0	0	2	0	1	0	0
t3	0	0	0	0	0	0	1	0
t4	0	0	0	0	0	0	1	1
t5	2	4	0	0	0	0	0	0
t6	1	4	0	0	0	0	0	0
t7	0	0	0	0	0	0	1	0

을 분류한 결과를 나타낸 것이다.

제안된 방법은 다음과 같다. 첫째, 분류 주제를 추출하는 방법으로는, 표 2(a)의 비음수 의미 특징 행렬 W를 이용하여, 각각의 군집 안에 포함된 의미 특징의 값이 가장 큰 의미 특징과 대응되는 용어를 분류의 주제로 선택한다. 의미 특징 값이 크다는 것은 대응되는 용어가 중요하다는 것을 의미한다. 즉, 표 2(a)에서 군집 *C1*에 포함된 가장 큰 의미 특징 값 6.487과 대응되는 용어 *t1*이 분류 주제로 선택된다.

둘째, Xu등이 제안한 군집 방법[14]을 이용한 자동 이메일 다원 분류 방법은 다음과 같다. 표 1의 용어-이메일 빈도 행렬 A에 식 (3)과 식 (4)를 이용하여 비음수 행렬 분해를 수행하여서 비음수 의미 특징 행렬 W와 비음수 의미 변수 행렬 H를 얻는다. 식 (9)를 이용하여 행렬 W와 H를 정규화한다. 행렬 W를 이용하여 이메일을 분류한다. 예를 들어, 만약 $x = argmax_j h_{ji}$ 이면 이메일 e_i를 분류 x에 할당한다. 즉, 표 2(b)에서 분류 *C1*에 포함되는 이메일은 *e3, e4, e5*가 된다.

$$w_{ij} \leftarrow \frac{w_{ij}}{\sqrt{\sum_i w_{ij}^2}}, \; h_{ij} \leftarrow h_{ij} \sqrt{\sum_i w_{ij}^2} \tag{9}$$

표 2 표 1로부터 유도된 행렬 W와 H

(a) 의미 특징 행렬 W

	C1	*C2*	*C3*	*C4*
t1	6.487	0	0	0
t2	0	0	2.224	0
t3	0	0	0	0.924
t4	0	0	0	1.307
t5	0	4.449	0	0
t6	0	4.094	0	0
t7	0	0	0	

(b) 의미 변수 행렬 H

	e1	*e2*	*e3*	*e4*	*e5*	*e6*	*e7*	*e8*
C1	0	0	1.947	1.678	0.789	0	0	0
C2	1.240	2.147	0	0	0	0	0	0
C3	0	0	0	0.563	0	15413	0	0
C4	0	0	0	0	0.225	0	1.852	0.912

표 3 비음수 행렬 분해에 의한 분류 주제 및 이메일 분류 결과

분류 주제	이메일
t1	*e3, e4, e5*
t5	*e1, e2*
t2	*e6*
t4	*e7, e8*

3.2 동적 분류 체계방법에 의한 분류 내 이메일의 재구성 방법

이메일을 동적으로 재구성하기 위해서는 용어와 분류 주제 간의 관계를 규정해야 한다. 그러나 용어와 분류 주제 간의 관계를 직접 결정할 수는 없으므로 용어와 메일간의 관계 및 이메일과 분류 주제 간의 관계에 의해서 결정한다. 즉, 이메일을 용어로 구성된 퍼지 집합으로 간주할 수 있고, 마찬가지 분류 주제를 분류된 이메일들의 용어들로 구성된 퍼지 집합으로 간주할 수 있다. 이메일이 속한 두 분류 주제 간의 관계는 생성된 두 분류 주제의 퍼지 집합의 함의 정도를 식 (6)과 식 (8)을 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 분류 주제의 유사 관계를 동적으로 생성할 수 있다[14-19]. 다음 예 4는 동적 분류 체계 방법을 예를 이용하여서 설명한다.

예 4. 다음 표 4부터 표 7, 그림 2와 그림 3은 분류 주제와 용어의 관계로부터 이메일의 분류를 동적으로 재구성하는 예를 나타낸다. 표 4는 분류 주제와 용어의 관계를 나타내며, 표 5는 표 4에 식 (8)을 이용하여 분류 주제와 분류 주제사이의 관계를 난타 낸 것이다. 표 6과 표 7은 식 (6)을 이용하여 분류 주제 간의 관계를 재구성한 것을 나타내며, 그림 2와 그림 3은 표 6과 표 7을 도식화한 것이다.

표 4 분류 주제와 용어와의 관계표

	t_1	t_2	t_3	t_4	t_5
C_1	0.9	1.0	1.0	1.0	1.0
C_2	0.0	1.0	0.1	0.0	1.0
C_3	1.0	0.8	0.0	1.0	1.0
C_4	0.0	0.0	1.0	0.0	0.1
C_5	0.0	1.0	1.0	0.8	1.0

표 5 표 4에 식 (8)을 이용하여 유도된 결과

	C_1	C_2	C_3	C_4	C_5
C_1	0.98	0.44	0.76	0.24	0.78
C_2	1.00	1.00	0.90	0.05	1.00
C_3	0.97	0.33	1.00	0.03	0.60
C_4	1.00	0.10	0.00	1.00	1.00
C_5	1.00	0.70	0.60	0.37	1.00

표 6 표 5에 식 (6)의 알파 값이 0.94일 때 유도된 결과

	C_1	C_2	C_3	C_4	C_5
C_1	1	0	0	0	0
C_2	1	1	1	0	1
C_3	1	0	1	0	0
C_4	1	0	0	1	1
C_5	1	0	0	0	1

표 7 표 5에 식 (6)의 알파 값이 0.76일 때 유도된 결과

	C_1	C_2	C_3	C_4	C_5
C_1	1	0	1	0	1
C_2	1	1	1	0	1
C_3	1	0	1	0	1
C_4	1	0	0	1	1
C_5	1	1	1	0	1

그림 2와 그림 3에서 보이는 것과 같이 표 5에 식 (6)의 알파 값을 조정함으로서 분류 주제가 동적으로 재구성됨을 알 수 있다. 즉, 그림 2에서는 알파 값이 0.94일 때의 분류 주제 간의 포함 관계를 나타내고 있으며, 그림 3에서는 알파 값이 0.76일 때 분류 주제 간의 포함 관계를 나타내고 있다. 즉, 그림 2에서 보이는 것과 알파 값이 0.94일 때 분류 주제 C_2가 최상위 분류 주제를 나타내면 나머지 분류 주제들이 포함됨을 알 수 있다. 또한 그림 3에서는 알파 값이 0.76일 때 그림 2의 모든 분류 주제를 포함하면서 하위 분류 주제에 다시 분류 주제 C_2, C_3, C_5가 확장됨을 알 수 있다.

제안 방법은 다음과 같다. 위의 동적 분류 체계 방법을 이용하여 이메일 분류 주제를 그림 2나 그림 3과 같이 재구성한다. 재구성된 분류 주제에서 C_2같이 최상위 루트의 분류 주제는 재구성되기 전의 이메일을 분류 주제에 포함시킨다. 이것은 그림 2나 그림 3에서 보이는 것과 같이 C_2는 모든 분류 주제를 포함하기 때문이다. 그림 3의 C_1과 같이 C_3과 C_5의 자손 노드를 분류 주제로 갖고 있는 경우, 분류 주제는 부모 노드인 C_1을 사용하고, C_3와 C_5의 이메일은 C_1의 분류 주제 내의 이메일에 중복되는 않는 이메일을 C_1의 분류 주제에 포함시켜서 분류 주제 내의 이메일을 재구성한다.

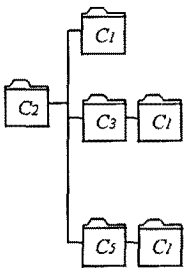


그림 2 표 6을 도식화한 결과

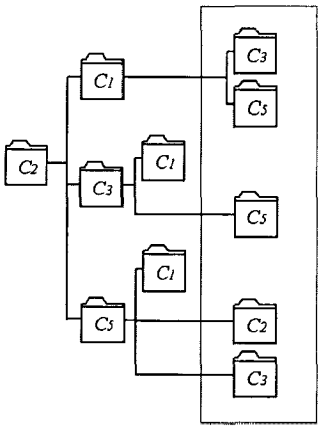


그림 3 표 7을 도식화한 결과

3.3 이메일 자료를 사용한 자동 이메일 분류 및 재구성의 예

다음의 예제에서 사용된 자동 이메일 다원 분류 시스템은 비주얼 베이직 6.0으로 구현하였다. 예에 사용된 이메일 자료는 2009년 8월에서 2009년 9월 사이의 110개의 메일을 대상으로 하였다. 그림 4는 이메일 다원 분류 시스템에 대한 인터페이스이다. 인터페이스는 파일, 전처리, 이메일 다원 분류, 동적 계층 분류 데이터베이스, 정보 메뉴 등으로 구성된다. 파일 메뉴는 시스템을 종료하며, 전처리 메뉴에서는 메일 서버로부터 메일을 가져와 용어-메일 빈도 행렬의 데이터베이스를 만든다.

이메일 다원 분류 메뉴에서는 전처리된 자료를 이용하여 분류 주제를 추출하고, 추출된 분류 주제를 이용하여 그림 5와 같이 메일을 분류한다. 동적 계층 분류 메뉴에서는 사용자가 분류된 결과를 만족하지 못하면 그림 6과 같이 a 값을 조정하여 이메일을 다시 재분류한다. 데이터베이스 메뉴에서는 각 단계의 메뉴에서 만들어진 데이터베이스들을 볼 수 있다. 정보 메뉴는 시스템 버전을 나타낸다.

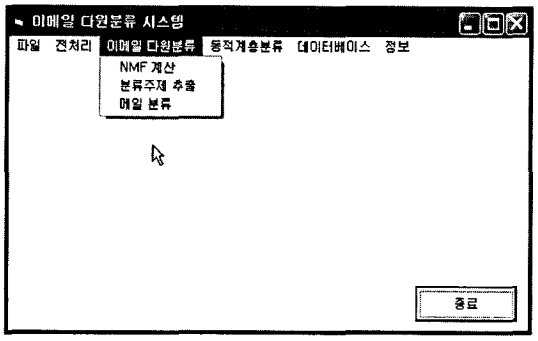


그림 4 이메일 다원 분류 시스템의 인터페이스

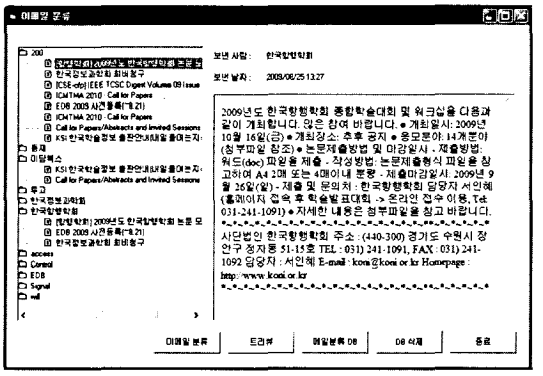
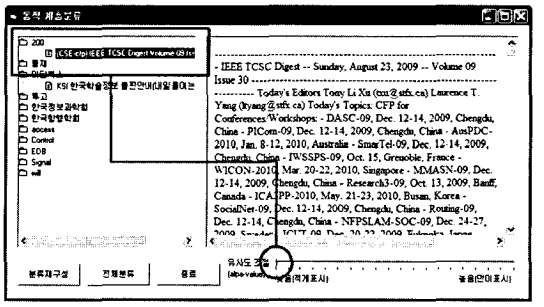
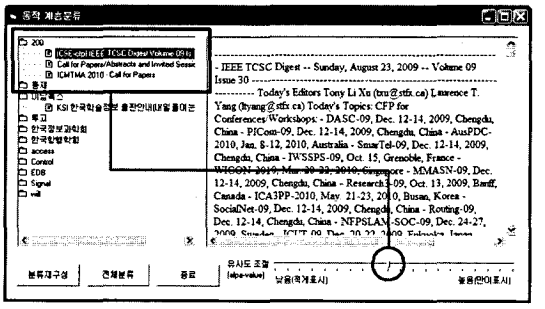


그림 5 이메일 다원 분류 결과



(a) a 값이 0.05일 때 이메일 재분류 결과



(b) a 값이 0.55일 때 이메일 재분류 결과

그림 6 a 값 변화에 따른 분류 주제 내의 이메일 재구성 결과

4. 실험 및 분석

실험 자료는 2009년 6월 1일부터 2009년 6월 30일까지 수신된 메일 중에서 분류 주제와는 상관없이 임의로 200개의 메일을 선택하였다. 평가는 수작업으로 분류된 메일을 제안된 방법과 비교한 정확률을 분석하였다. 이때 분류 주제는 메일에 포함된 단어로 한정하였다. 수작업으로 분류하기 위한 10개의 분류 주제를 선택하였다.

본 논문에서는 평균 분류 정확률을 분석의 평가 방법으로 사용하였다. 분류정확률은 수작업으로 분류한 메일

표 8 분류의 적합성

분류		수작업에 의한 분류	
		correct	incorrect
분류 방법에 의한 분류	correct	a	b
	incorrect	c	d

과 자동 분류한 메일을 비교하여 바르게 분류된 메일의 정확률을 다음 식 (10)과 같이 계산하였다.

$$\text{정확률 } (p) = \frac{a}{a+b} \tag{10}$$

그림 7은 제안 방법과 서로 다른 3가지 방법 간의 자동 이메일 다원 분류 결과에 대한 평균 분류 정확률을 나타낸 것이다. 여기서는 TFIDF는 이메일의 유사도를 이용하여 이메일을 분류한 방법이며[7], TFIDF-DCH는 이전에 제안한 방법으로 유사도와 동적 분류 체계를 이용하여 제안한 방법이다[9]. PCA-DCH도 PCA와 동적 분류 체계방법을 이용하여 이전에 제안한 방법이다[10]. NMF-DCH는 본 논문에서 제안한 방법이다. 그림 7에서 제안 방법의 평균 분류 정확률이 TFIDE에 비하여서는 7.6%가, TFIDF-DCH에 비해서는 5.1%가, PCA-DCH에 비해서는 2%가 더 높은 것을 알 수 있다.

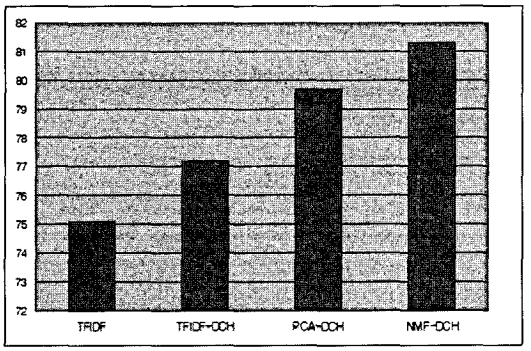


그림 7 평균 분류 정확률

또한, 분류 주제 내의 이메일을 재구성할 때에 a 값에 따라서 81.3%에서 95.6%까지 평균 분류 정확률이 높은 것을 알 수 있다.

5. 결 론

본 논문에서는 이메일을 자동으로 다원 분류하고 분류된 결과를 사용자의 요구 사항에 맞게 재분류할 수 있는 방법을 제안하였다. 제안된 방법은 비음수 행렬 분해를 이용하여 이메일의 분류 주제를 생성하고 다원 분류한다. 이렇게 분류된 이메일을 사용자의 요구에 따라서 언제든지 동적 분류 체계 방법을 이용해서 분류 주제 내의 이메일을 재구성할 수 있다. 이러한 분류 주제 내의 이메일의 재구성 정도는 사용자가 조절할 수 있도록 하여 효율적으로 이메일을 관리할 수 있다. 마지막으로 분류 규칙에 대한 별도의 훈련 및 학습 과정이 필요 없이 이메일을 빠르게 분류함으로서 유동적인 이메일 환경에 적합하다.

참 고 문 헌

[1] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE Transactions on Neural network*, 10(5), 1999.

[2] 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, pp.4-5, November, 2002.

[3] 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.

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

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

[6] Y. Q. Xia, J. X. Wang, F. Zheng, Y. Liu, "A Binarization Approach to Email Categorization using Binary Decision Tree," *in the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, Aug. pp.19-22, 2007.

[7] B. Yu, D. H. Zhu, "Combining neural networks and semantic feature space for email clasification," *Knowledge-Based Systems*, 22, pp.376-381, 2009.

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

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

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

[11] D. Cutting, K. David, J. Pedersen, J. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collection," In *the ACM SIGIR International Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, Jun, pp.318-329, 1992.

[12] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization,"

Nature, vol.401, pp.788-791, 1999.
[13] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," In *Advances in Neural Information Processing Systems*, vol.13, pp.556-562, 2001.
[14] B. G. Choi, J. H. Lee, S. Park. "Dynamic Construction of Category Hierarchy Using Fuzzy Relational Products," *IDEAL 2003*, pp.296-302, 2003.
[15] S. Park, S. H. Park, J. H. Lee, J. S. Lee, "E-mail Classification Agent Using Category Generation and Daynamic Category Hierarchy," *LNAI 3397*, pp.207-214. 2005.
[16] 박선, 안찬민, 박상호, 이주홍, 최범기, "자동 카테코리 생성과 동적 분류 체계를 사용한 이메일 분류", *한국지능정보시스템학회논문지*, 제10권 제2호, pp.79-89, 2004.
[17] S. Park, C. W. Kim, "E-mail Classification and Category Re-organization using Dynamic Category Hierarchy and PCA," In *proceeding of CIKI-MICS'09*, 2009.
[18] 박선, "PCA와 동적 분류 체계를 사용한 자동 이메일 계층 분류", *한국항행학회논문지*, 제13권 제3호, pp.419-425, 2009.
[19] 박선, 안동언, "비음수 행렬 분해와 동적 분류 체계를 사용한 이메일 분류", *2009년도 제21회 한글 및 한국어 정보처리 학술대회*, 35-39, 2009.
[20] 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.
[21] W. Xu, X. Liu, Y. Gon, "Document Clustering Based On Non-negative Matrix Factorization," *Proceeding of Special Interest Group on Information Retrieval (SIGIR)*, 267-274, 2003.

박 선

1996년 2월 전주대학교 전자계산학과(이학사). 2001년 8월 한남대학교 정보산업대학원 정보통신학과(공학석사). 2007년 8월 인하대학교 컴퓨터정보공학과(공학박사). 2008년~2009년 8월 호남대학교 컴퓨터공학과 전임강사. 2009년 9월~현재 전북대학교 전기전자정보인력양성사업단 박사후과정. 관심분야는 정보검색, 데이터 마이닝, 데이터베이스

안 동 언

1981년 한양대학교 전자공학과(공학사) 1987년 KAIST 컴퓨터공학과(공학석사) 1995년 KAIST 컴퓨터공학과(공학박사) 현재 전북대학교 전기정보공학부 교수 관심분야는 자연어처리, 정보검색, 기계번역