비음수 행렬 분해와 동적 분류체계를 사용한 이메일 분류

박선[O], 안동언
[O]전북대학교 전라북도 전략산업 혁신을 위한 전자정보고급인력양성사업단, 전자정보공학부, 전북대학교
[O]sunbak@jbnu.ac.kr, duan@chonbuk.ac.kr

Email Classification using Dynamic Category Hierarchy and Non-negative Matrix Factorization

Sun Park[O], Dong Un An
[O]Advanced Graduate Education Center of Jeonbuk for Electronics and Information Technology-BK21, Division of Electronic & Information Engineering, Chonbuk National University

요 약

이메일의 사용증가로 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 비음수 행렬 분해(NMF, Non-negative Matrix Factrazation)를 기반으로 한 자동 분류 주제 생성 방법과 동적 분류 체계(DCH, Dynamic Category Hierachy) 방법을 결합한 새로운 이메일 분류 방법을 제안한다. 이 방법은 수신되는 이메일을 자동으로 분류하여 대량의 메일을 효율적으로 관리할 수 있으며, 분류 결과 사용자의 요구사항을 만족하지 못하면 메일을 동적으로 재분류 하여 분류 정확률을 높일 수 있다.

주제어: 이메일 분류, 비음수 행렬 분해, 동적 분류 체계

1. 서 론

인터넷의 발전은 이메일 사용의 증가를 가져 왔으며, 요즘 대부분의 개인우편물부터 광고 등 다양한 분야의 우편물이 이메일로 교체되고 있다. 이러한 이메일 사용의 폭발적인 증가는 하루 동안에 수십 통에서 수천 통에 이르는 이메일을 사용자들에게 수신되게 하고 있다. 그러나 수신되는 메일의 대부분은 스팸 메일이 차지하고 있다.

이러한 이유에서 스팸 메일을 효율적으로 처리 할 수 있는 많은 도구들이 개발되었다. 그러나 대부분의 도구들은 사용자가 직접 필터링 규칙을 만들거나 메일을 분류할 색인어 목록을 작성해야 한다. 이렇게 사용자의 규칙이 들어가는 도구들은 색인어를 많이 포함하는 대량의 메일을 분류해야 할 경우 분류의 정확성이 떨어지는 단점이 있다. 또한 사용자의 변화되는 요구사항에 맞추어 재 분류 하거나 재 필터링할 수 없는 단점이 있다.

현재까지의 이메일 분류에 대한 관련 연구로는 대부분 스팸 메일을 구분하는 이원분류가 주로 연구되었다. 스팸 분류를 위해 사용된 방법으로는 베이시안, 규칙기반 등이 있다. Androutsopoulos[5]와 Sakkis[12]은 베이시안 분류자를 이용하였고, Cohen[2]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[4]는 군집 기술과 데이터마이닝 알고리즘을 이용하였으며, Mock[8]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 비음수 행렬 분해와 동적 분류체계 방법을 사용한 이메일 다원 분류 방법을 제안한다. 비음수 행렬 분해를 사용하여 자동으로 이메일의 분류 주제를 생성하고 다원분류 한다. 다원분류결과의 정확도가 떨어지는 문제를 해결하기 위해 동적 분류 체계 방법을 이용하여 이메일 분류 체계를 동적으로 재구성 할 수 있게 하였다. 즉, 사용자가 자동으로 분류된 이메일의 다원 분류 계층으로 부터 원하는 이메일을 찾을 수 없으며, 분류계층을 재분류하여서 원하는 이메일을 쉽게 찾을 수 있도록 한다.

비음수 행렬 분해는(NMF, non-negative matrix factorization) Lee와 Seung이 제안한 방법으로 인간이 객체를 인식할 때 객체의 부분정보의 조합으로 인식하는 것에 착안하여, 객체정보를 기초특징(base feature)과 부호특징(encoding feature)로 나누어 부분정보(part-base)로 표현한다. 이러한 부분정보의 조합으로 전체 객체를 표현하는 방법은 대량의 정보를 효율적으로 표현 할 수 있는 방법이다[6, 7].

동적 분류 체계 방법[3]은 최범기외 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, 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 비음수 행렬 분해를 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.

2. 비음수 행렬 분해

비음수 행렬 분해는 주어진 양의 행렬로부터 양의 인수를 찾아내는 행렬분해 알고리즘이다[6, 7]. 비음수 행렬 분해는 문서집합이 k개의 군집으로 구성된다고 가정할 때, 행렬 X를 식(2)의 목적 함수가 최소 값을 갖도록 식(1)과 같이 $m \times k$ 비음수 의미특징 행렬(NSFM) W와 $k \times n$ 비음수 의미변수 행렬(NSFM) 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 목적함수 J 값이 수렴 허용오차 보다 작아지거나 지정한 반복횟수를 초과할 때까지 식(3)과 식(4)를 반복한다[6, 7].

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

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

여기서 H^T는 H의 전치행렬이고, W^T는 W의 전치행렬이다.

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

행렬 A의 j번째 열벡터 A_{*j}는 행렬 W의 l번째 열벡터 W_{*l}와 행렬 H의 요소 H_{kj}^T가 선형조합을 이루며 식(5)과 같다.

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

예1) 다음은 식3과 식4를 이용하여 A행렬을 W와 H행렬로 분해 한 예이다. r = 2, 수렴할 반복 횟수는 50 이고, 수렴 허용오차가 0.001이다. W와 H행렬의 초기 값은 각각 0.5이다.

$$\underset{A}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}} \approx \underset{W}{\begin{bmatrix} 0.15 & 1.60 \\ 0.66 & 0.97 \\ 1.15 & 0.57 \\ 1.61 & 0.41 \end{bmatrix}} \times \underset{H}{\begin{bmatrix} 6.11 & 6.68 & 7.18 \\ 0.09 & 0.60 & 1.22 \end{bmatrix}} = \underset{\widetilde{A}}{\begin{bmatrix} 1.05 & 1.94 & 3.03 \\ 4.12 & 4.98 & 5.93 \\ 7.07 & 8.01 & 8.95 \\ 9.90 & 11.01 & 12.08 \end{bmatrix}}$$

3. 동적 분류체계 방법

동적 분류 체계 방법에서 사용되는 퍼지 이론은 다음과 같다[11]. 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[3].

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

(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.

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

여기서 π_m은 평균 정도를 나타내는 함수이다[1].

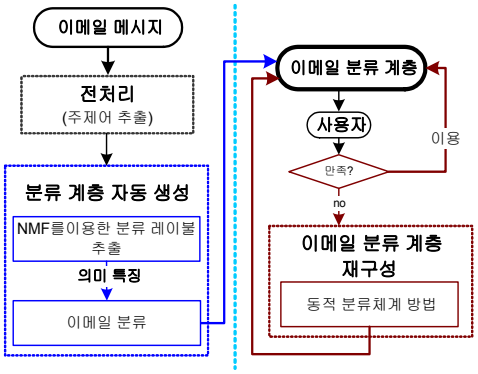
본 논문에서는 위의 식6의 Kleen-Diense 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식7의 퍼지관계곱을 적용하여 분류들 간의 퍼지함의관계 $C_i \rightarrow C_j$를 유도할 수 있다. 그러나 C_i 에 멤버쉽 값($\mu_{Ci}(x)$)이 작은 원소 x가 많으면, $C_i \subseteq C_j$의 포함여부와 관계없이 항상 1에 가까운 값이 나오는 문제점이 있다. 따라서 다음과 같이 정의하여 두 분류 퍼지 집합의 함의 관계 $\mu_{m,\beta}(C_i \subseteq C_j)$를 계산한다.

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

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.

4. 자동 이메일 분류방법

본 논문에서 제안한 이메일의 분류 과정은 다음과 같다. 첫째, 수신 메일에서 색인어를 추출한다. 둘째, 메일과 색인어의 출현 빈도를 이용하여 메일용어 행렬을 구성한다. 비음수 행렬 분해를 이용하여 구성된 메일용어 행렬로부터 이메일 분류 및 분류 주제를 생성 한다. 마지막으로 사용자의 필요에 따라 동적 분류 체계 방법을 이용하여 분류 주제를 재구성 한다. 다음 그림1은 제안 시스템으로 자동 이메일 다원 계층 분류 및 이메일 분류 계층 재구성 방법을 보여준다.



<그림1> 제안된 이메일 다원분류 시스템

이메일을 분류하는 방법으로는 본 논문에서는 Xu등이 제안한 비음수 행렬 분해를 이용한 문서 군집 방법을 사용한다. Xu등이 제안한 방법은 다음과 같다[13]. 먼저 주어진 문서 집합에서 i열로 가중치가 부여된 용어빈도 벡터 문서를 가지는 용어문서 행렬 A를 구성한다. 행렬 A에 식(3)과 식(4)를 이용하여 비음수 행렬 분해를 수행 하여서 비음수 행렬 W와 H를 얻는다.

식(9)를 이용하여 행렬 W와 H를 정규화 한다. 행렬 W를 이용하여 각 문서의 군집 레이블을 결정한다 예를 들어, 만약 x = $\arg max_j h_{ji}$ 이면 문서 d_i를 군집 x에 할당한다.

$$w_{ij} \leftarrow \frac{w_{ij}}{\sqrt{\sum_i w_{ij}^2}}, \quad h_{ij} \leftarrow h_{ij} \sqrt{\sum_i w_{ij}^2} \qquad (9)$$

예2) 다음 표1부터 표3까지는 8개의 이메일로부터 4개의 분류 주제별로 분류한 예를 나타낸다. 표1은 8개의 이메일부터 7개의 용어를 사용한 이메일용어 빈도행렬을 나타내며, 표2는 표1을 비음수 행렬 분해하여 생성된 비음수 의미 변수 행렬을 나타낸다. 표3은 표2로부터 분류 주제와 분류 주제로 이메일을 분류한 것을 나타낸 것이다.

(표1) 8개의 이메일의 이메일-용어 빈도행렬

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

(표2) 표1로부터 유도된 의미 변수 행렬

	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	1.413	0	0
C4	0	0	0	0	0.225	0	1.852	0.912

(표3) 비음수 행렬분해에 의한 이메일 분류결과

분류 주제	이메일
C1	e3, e4, e5
C2	e1, e2
C3	e6
C4	e7, e8

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.

메일을 용어로 구성된 퍼지 집합으로 간주할 수 있고, 마찬가지 분류 주제를 분류된 메일들의 용어들로 구성된 퍼지 집합으로 간주할 수 있다. 메일이 속한 두 분류 주제 간의 관계는 생성된 두 분류 주제의 퍼지 집합의 함의 정도를 식6과 식8을 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.

예3) 다음 표4부터 표7, 그림2와 그림3은 분류 주제와 용어의 관계로부터 이메일의 분류를 동적으로 재구성하는 예를 나타낸다. 표4는 분류주제와 용어의 관계를 나타내며, 표 5는 표4에 식8을 이용하여 분류 주제와 분류 주제사이의 관계를 난타 낸 것이다. 표6과 표7은 식6을 이용하여 분류 주제 간의 관계를 재구성한 것을 나타내며, 그림2와 그림3은 표6과 표7을 도식화 한 것이다.

(표4) 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

(표5) 표4에 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

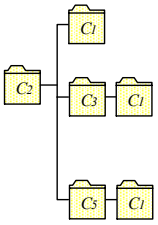
(표6) 표5에 식(6)의 알파 값이 0.94일 때 유도된 결과

	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

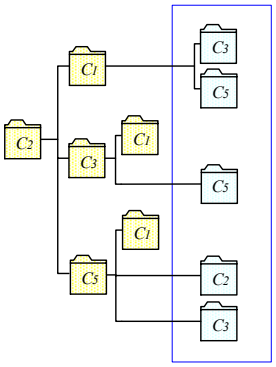
(표7) 표5에 식(6)의 알파 값이 0.76일 때 유도된 결과

	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

그림2와 그림3에서 보이는 것과 같이 표5에 식(6)의 알파 값을 조정함으로서 분류주제가 동적으로 재구성됨을 알 수 있다. 즉, 그림2에서는 알파 값이 0.94일 때의 분류 주제 간의 포함관계를 나타내고 있으며, 그림3에서는 알파 값이 0.76일 때 분류 주제 간의 포함관계를 나타내고 있다. 즉, 그림2에서 보이는 것과 알파 값이 0.94일 때 분류 주제 C_2가 최상위 분류주제를 나타내면 나머지 분류 주제들이 포함됨을 알 수 있다. 또한 그림3에서는 알파 값이 0.76일 때 그림2의 모든 분류 주제를 포함하면서 하위분류 주제에 다시 분류 주제 C_2, C_3, C_5가 확장됨을 알 수 있다.



<그림2> 표6을 도식화한 결과



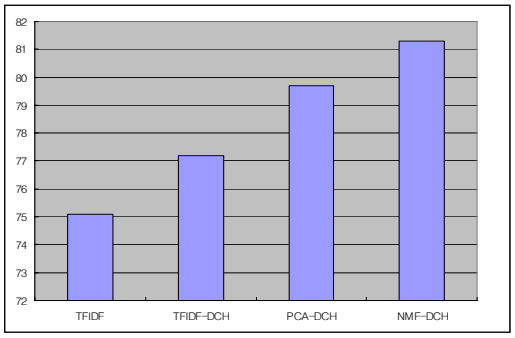
<그림3> 표7을 도식화한 결과

5. 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 분류주제는 메일에 포함된 단어로 한정하였다. 수작업으로 분류하기 위한 10개의 분류주제를 선택하였다.

본 논문에서는 평균 분류 정확률을 분석의 평가 방법으로 사용하였다. 분류정확률은 수작업으로 분류한 메일과 자동 분류한 메일을 비교하여 바르게 분류된 메일의 정확률을 계산하였다.

그림 4는 제안방법과 서로 다른3가지 방법 간의 평균 분류 정확률을 나타낸다. 여기서는 TFIDF는 이메일의 유사도를 이용하여 이메일을 분류한 방법이며, TFIDF-DCH는 이전에 제안한 방법으로 유사도와 동적 분류체계를 이용하여 제안한 방법이다[9]. PCA-DCH도 PCA와 동적 분류 체계방법을 이용하여 이전에 제안한 방법이다[10]. NMF-DCH는 본 논문에서 제안한 방법이다. 그림4에서 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 4> 평균 분류 정확률

6. 결론

본 논문에서는 이메일을 분류하고 분류된 결과를 사용자의 요구사항에 맞게 재분류할 수 있는 방법을 제안하였다. 제안된 방법은 비음수 행렬 분해를 이용하여 이메일을 분류하고 분류 주제어를 생성한다. 이렇게 분류된 이메일을 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] 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.

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

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

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

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

[6] D. D. Lee, H. S. Seung, "Learning the parts of objects by non-negative matrix factorization', Nature, vol.401, 788-791, 1999.

[7] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization", In Advances in Neural Information Processing Systems, vol.13, 556-562, 2001.

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

[9] 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)

[10] S. Park, C. W. Kim, E-mail Classification and Category Re-organization using Dynamic Category Hierarchy and PCA, In proceeding of CIKIMICS'09 (2009)

[11] 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.

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

[13] 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.