한국지능정보시스템학회논문지 제10권 제2호 2004년 11월(pp.79~89)

자동 카테고리 생성과 동적 분류 체계를 사용한 이메일 분류

박 선
인하대학교 컴퓨터공학과
(sunpark@datamining.inha.ac.kr)

안찬민
인하대학교 컴퓨터공학과
(ahnch1@datamining.inha.ac.kr)

박상호
인하대학교 컴퓨터공학과
(parksangho@datamining.inha.ac.kr)

이주홍
인하대학교 컴퓨터공학과
(juhong@inha.ac.kr)

최범기
쿼크(주)
(bumghichoi@yahoo.co.kr)

이메일 사용이 보편화됨에 따라 점차 수신되는 메일의 량이 증가하고 있다. 이러한 메일 량의 증가는 사용자로 하여금 이메일을 좀더 효율적으로 분류할 수 있는 방법을 필요하게 한다. 그러나 현재의 이메일 분류는 규칙기반, 베이시안, SVM 등을 이용하여 스팸메일을 필터링 하는 이원분류가 주로 연구되고 있다. 이외에도 다원분류에 대한 연구로는 클러스터링을 이용한 방법이 있으나, 이는 단순히 유사도에 의해 메일을 그룹화 하는 수준이다. 본 논문에서는 벡터모델의 유사도를 기반으로 한 자동 카테고리 생성 방법과 동적분류체계 방법을 결합하여 새로운 이메일 자동 분류 방법을 제안했다. 본 논문에서 제안한 방법은 이메일을 자동으로 다원분류하며 대량의 메일도 효율적으로 관리할 수 있다. 또한 메일을 동적으로 재분류 할 수 있게 함으로써 정확율을 높였다.

논문접수일 : 2004년 5월 게재확정일 : 2004년 11월 교신저자 : 박 선

1. 서론

이메일 사용의 증가는 수신자에게 메일을 분류하고 정리하는데 많은 시간을 요구한다. 이러한 문제를 해결하기 위해 많은 도구들이 개발되었으나 대부분은 사전에 사용자가 직접 필터링 규칙이나 메일이 분류될 수 있도록 색인어 목록을 작성해야 하며, 메일이 많은 동음이의어나 동의어를 포함하는 경우 주제가 불분명해져서 오분류가 증가된다. 또한 시간이 지나 사용자의 변화되는 요구사항에 맞추어 재분류하거나 재필터링 할 수 없는 단점이 있다.

본 논문에서는 벡터모델의 유사도를 기반으로 한 자동 카테고리 생성 방법과 동적분류체계 방법을 결합하여 새로운 이메일 다원분류 방법을 제안한다. 본 논문에서 제안한 방법은 다음과 같은 장점을 가진다. 첫째, 자동 카테고리 생성 방법에 의해 메일의 분류주제가 자동 생성됨으로 사용자의 간섭 필요 없다. 둘째, 동적분류체계 방법을 이용하여 사용자가 필요하면 언제든지 재분류할 수 있다. 셋째, 수신되는 메일을 다원분류함으로 대량의 메일을 효율적으로 관리할 수 있다. 넷째, 학습이 필요 없기 때문에 유동적인 이메일 환경에 적합하다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를, 3장에서는 유사도 계산을 위한 벡터모델을 보인다. 4장은 동적분류체계 방법에 대하여 알아본다. 5장은 자동 카테고리 생성 방법과 동적인 분류 체계를 구성하는 방법을 설명한다. 6장에서는 실험 및 분석결과를 보인다. 7장에서 결론을 맺는다.

2. 관련연구

이메일 분류는 대부분 스팸메일을 찾는 이원분류가 주로 연구되었다. 적용된 방법으로는 규칙기반 분류(Rule-based Classifiers), 베이시안 분류(Bayesian Classifiers), SVM(Support-Vector Machines)등이 있다. Cohen은 텍스트 마이닝 기법과 전처리시 불리안과 벡터모델을 이용한 두개의 규칙기반 시스템을 제안하였다(Cohen, 1999). Androutsopoulos(Androutsopoulos, 2000)와 Sakkis(Sakkis, 2001)은 안티스팸 필터링을 하기 위해 베이시안 분류자를 이용하였다. 그들의 접근방법은 규칙기반 분류자를 사용하는 것에 비해 좋은 정확성을 보였다. Drucker(Drucker et al., 1999)는 SVM을 이용한 스팸 메일 분류를 제안하였다. Kunlun는 스팸을 분류하기 위해 활성 학습 정책을 이용하는 SVM 기반의 새로운 방법을 제안하였다(Kun-Lun et al., 2002). Woitaszek는 단순 SVM을 이용하여 상업적 이메일 분류 시스템을 만들었다(Woitaszek and Shaaban, 2003). SVM을 이용한 이들의 방법은 규칙기반이나 베이시안 분류에 비하여 좋은 성능을 보였다. 위의 규칙기반, 베이시안, SVM 접근방법은 관리자 분류 방법으로 수신된 메시지가 들어갈 비슷한 폴더를 찾을 수 있도록 사용자가 직접 메시지 폴더를 만들어야 한다. 또한, 메일을 분류하기 이전에 일정량 이상의 학습이 필요하고, 학습과 테스트에 시간이 걸리는 문제가 있다.

다원분류에 대한 연구로는 비관리자 분류 방법으로 수신된 메일 집합으로부터 메일 폴더를 자동으로 구성하여 이메일을 분류한다. Mock는 벡터모델에 의한 역색인방법으로 이메일 자동분류 시스템을 제안하였다(Mock, 1999). 그러나 Mock의 방법은 사용자의 필요에 따라 메일을 재분류할 수 없다. Manco는 이메일 메일을 관리 및 유지하기 위하여 데이터마이닝에 기반으로 한 k-NN을 이용하여 메일을 분류한다(Manco and Masciari, 2002). Manco의 방법은 사용자의 필요에 따라 재분류 할 수 있으나, 여러 단계의 전처리와 다양한 추출 정보에 의하여 유사도를 얻기 때문에 계산이 복잡하여 분류나 재분류시 많은 시간이 필요하다.

3. 유사도 계산을 위한 벡터 모델

본 장에서는 본 논문에서 유사도 계산을 위해 사용하는 벡터 모델인 $tf \cdot idf$ 기법에 대하여 알아본다. 벡터 모델은 불리안 모델의 0 또는 1의 가중치의 한계를 극복하고 질의문서와 검색문서 간의 부분일치를 가능하게 한다. 즉, 질의문서와 검색문서의 단어들에 연속형 수치의 가중치를 부여하고, 이 가중치들을 이용하여 유사도를 계산한 후, 상위의 유사도를 갖는 문서들을 검색해오는 방법인데, 불리안 모델에 의한 방법보다 검색효율이 좋아 현재 많이 사용되고 있다(Ricardo and Berthier, 1999).

벡터 모델은 하나의 문서를 t개의 정규화된 단어로 구성된 t-차원의 벡터로 표현한다. 즉, w_{ij} 는

검색문서 j의 단어 i의 가중치 ($w_{ij} \geq 0$), w_{iq} 는 질의문서 또는 질의어 q의 단어 i의 가중치 ($w_{iq} \geq 0$), t는 검색문서와 질의문서 내의 단어들의 개수 일때, 질의문서의 벡터 $\overrightarrow{q}$는 식(1)과 같이 나타 낼 수 있다.

$$\overrightarrow{q} = (w_{1q}, w_{2q}, \cdots, w_{tq}) \quad (1)$$

또한, 검색문서의 벡터 $\overrightarrow{d}$는 식(2)와 같이 나타낼 수 있다.

$$\overrightarrow{d_j} = (w_{1j}, w_{2j}, \cdots, w_{tj}) \quad (2)$$

벡터 모델에서 유사도 산출 공식은 식(3)과 같다.

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (3)$$

4. 동적분류체계 방법

이전에 우리가 제안한 동적분류체계 방법은 검색어와 분류 간의 관계를 규정하고, 분류들 간의 상호 관계를 규명함으로써 분류검색의 분류체계를 재구성함으로써 검색효율을 높이는 방법이다 (Choi et al., 2003). 분류와 검색어간의 관계는 문서에서 검색어의 중요도와 문서의 분류에서의 중요도 등의 관계를 구하여 설정할 수 있다. 이러한 관계는 분류를 검색어로 구성된 퍼지 집합으로 간주할 수 있게 한다. 두 분류 간의 관계는 유사도를 계산함으로써 규정할 수 있는데, 유사도는 한 퍼지 집합이 다른 퍼지 집합을 포함하는 정도로써 계산할 수 있다. 이것을 이용하면 서로 다른 분류의 유사관계를 동적으로 생성할 수 있다.

동적분류체계 방법에서 사용되는 퍼지 이론은 다음과 같다(Radev et al., 2000). 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(Choi et al., 2003).

$$a \rightarrow b = (1 - a) \vee b = \max(1 - a, b), \quad a = 0 \sim 1, \ b = 0 \sim 1 \quad (4)$$

본 논문에서는 위의 식(4)의 Kleen-Diense 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식(5)의 퍼지관계곱을 적용하여 분류들 간의 퍼지 함의관계, $C_i \rightarrow C_j$ 를 유도할 수 있다.

$$\pi_{m\beta}(C_i \subseteq C_j) = (R^T \triangleleft_{\beta} R)_{ij} = \frac{1}{|C_{i\beta}|} \sum_{K_k \in C_{i\beta}} (R_{ki}^T \rightarrow R_{kj}) \quad (5)$$

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

5. 자동 이메일 분류 방법

일반적으로 이메일을 분류하기 위해서는 메일의 구조를 분석하여 관련성이 있는 특질을 추출하고, 추출된 특질간의 유사도를 계산하여 관련 있는 메일을 분류한다. 그러나 메일의 기본적인 구조를 보면, 보낸사람, 보낸날짜, 받는사람, 참조, 제목, 본문, 첨부파일등 여러 항목으로 구성되어

있다. 이러한 기본구조를 모두 분석하여 유사도에 따라 분류하는 것은 계산이 복잡하여 처리시간이 길어진다. 또한, 한번에 모든 문서를 모아서 처리하는 문서분류와는 달리 메일분류는 분류할 메일들이 어느 순간에 어느 정도의 양을 처리할지 정확히 알 수 없다. 그러므로 메일을 수신 할 때 마다 유사도를 계산하여 분류하는 것이 가장 효율적이다.

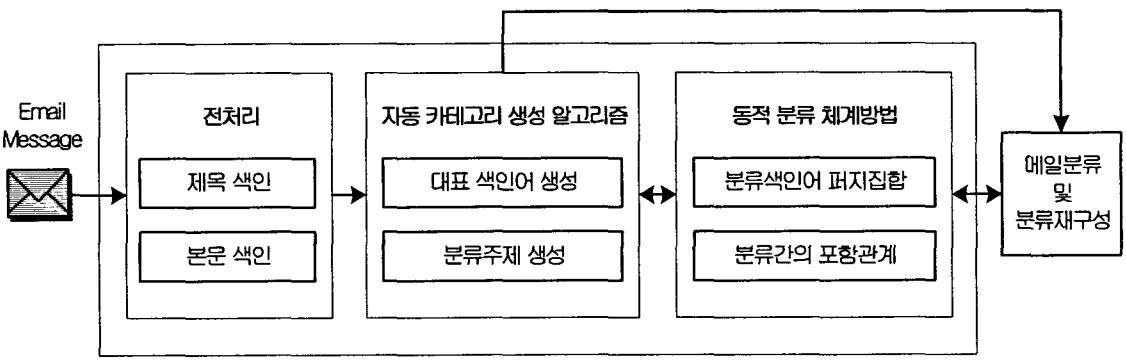
본 논문에서 이메일을 분류하는 과정은 다음 세단계로 이루어진다. 첫 단계는 수신 이메일로부터 제목과 내용을 전처리하여 각각의 색인어를 추출하다. 두번째 단계는 색인어간의 유사도를 계산하여 카테고리를 자동으로 생성하고, 카테고리별로 이메일을 자동분류 한다. 세번째 단계는 두번째 단계에서 얻은 분류주제, 분류주제에 포한된 메일, 각 메일의 색인어 관계를 동적분류체계 방법에 적용하여 사용자가 원할 때면 언제든지 재분류 및 재구성할 수 있다. [그림 1]은 메일분류의 전체 과정을 나타낸 것이다.

5.1 전처리

메일의 모든 항목을 전처리하여 유사도를 계산하는 것은 유동적인 이메일 특성상 맞지 않다. 때문에 본 논문에서는 메일의 제목, 본문, 보낸사람만으로 제한한다. 전처리는 수신된 이메일의 제목과 본문으로 부터 색인어를 추출하는 단계이다. 그러나 색인어 추출을 위한 불용어의 정의나 스테밍방법에 대한 연구가 많이 되어있는 영어 문서와는 달리, 한글 문서는 그 색인어의 추출 방법이 상대적으로 까다롭다. 특히 정보 검색과 관련해 문제가 되는 것은 복합명사와 고유명사 혹은 신조어의 존재인데, 띄어쓰기에 대한 명확한 규정이 없는 한글의 복합명사는 띄어쓰기의 방법에 따라 추출되는 색인어의 형태가 다르다. 본 논문에서는 이메일 분류 시스템의 구현상의 부담을 줄이고자 이미 개발되어 있는 한글분석 HAM을 사용하여 색인어를 추출하였다(Kang, 2002).

5.2 자동 카테고리 생성 방법

본 절에서는 분류주제의 자동 카테고리 생성 방법을 제안하다. 제안된 방법은 전처리된 색인어들 간에 유사도를 계산하고, 유사도가 가장 높은 색인어를 대표색인어로 설정한다. 다음으로 대표색인어를 첫번째 분류주제로 지정한다. 두번째 수신 메일도 마찬가지로 대표색인어를 설정하고, 설정된 대표색인어를 이전 메일의 색인어간



[그림 1] 이메일의 자동분류 및 재분류과정

유사도를 식(3)으로 계산한다. 계산된 유사도가 사용자가 지정한 분류 경계값보다 높으면 첫번째 분류주제에 포함시키고, 이보다 낮으면 두번째 메일의 대표색인어를 두번째 분류주제로 지정한 후, 분류주제에 메일을 포함시킨다. 이러한 과정을 마지막 수신 메일까지 반복하여 메일을 분류한다.

제안방법은 메일에 포함된 색인어 중 유사도가 가장 높은 색인어를 대표로 지정하여 분류주제를 자동으로 생성하고, 수신받는 메일을 생성된 분류주제별로 자동 분류하는 방법이다. 그러나 메일의 제목이 아무런 의미도 갖지 못함은 물론 메일의 의도도 내포하지 못한다면 식(3)를 사용한 자동 카테고리 생성 방법은 불필요하거나 메일 분류를 왜곡시킬 수 있다. 또한 메일 내용이 제목과 유사한 내용이라도 중요한 의미를 담고 있는 특질을 포함하고 있지 않다면 중요한 문장이 될 수 없으며, 반대로 제목과 유사성이 없는 내용이라도 중요한 의미를 포함한 특질이 나타나는 내용이라면 중요하게 고려해야 한다.

본 논문에서는 이러한 문제를 해결하기 위해 사용자의 필요에 따라 동적분류체계 방법을 이용하여 동적으로 재분류할 수 있게 하였다.

5.3 동적분류체계 방법에 의한 이메일 재분류

본 논문에서는 이메일을 동적분류체계로 구성하기 위해 색인어와 분류주제 간의 관계를 규정해야 한다. 그러나 색인어와 분류주제 간의 관계를 직접 결정할 수 는 없으므로 색인어와 메일 간의 관계 및 메일과 분류주제 간의 관계에 의해서 결정 한다. 이러한 관계는 5.2절의 자동 카테고리 생성 방법으로 유도할 수 있다. 여기서, 메일을 색인어로 구성된 퍼지 집합으로 간주 할 수 있고, 마찬가지 분류주제를 분류된 메일들의 색인어들로 구성된 퍼지 집합으로 간주할 수 있다. 메일이 속한 두 분류주제 간의 관계는 생성된 두 분류주제의 퍼지 집합의 함의 정도를 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.

퍼지 함의 연산자는 각 응용의 필요성에 맞게 제시되어야 하는데 본 논문에서는 식(4)의 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식(5)의 퍼지 관계곱을 적용하여 분류주제들 간의 퍼지 함의 관계, $C_i \rightarrow C_j$ 를 유도할 수 있다. 이렇게 유도된 $C_i \rightarrow C_j$ 는 $C_i \subseteq C_j$ 의 포함 정도를 나타낸다. 다음에는 분류간의 포함정도를 α-cut 하여 크리스프 값으로 바꾸면 최종 결과로서 각 분류간의 관계를 얻을 수 있다. 여기서 α 값을 조정하여 분류주제와 분류주제의 포함 관계를 동적으로 축소하던지 확장할 수 있다. 다음은 동적분류체계 방법을 적용한 예이다.

예1) $\beta = 0.9$ 일때 $\pi_{m,\beta}(C_5 \subseteq C_2)$는 $(R^T \triangleleft_\beta R)_{52} = 0.82$ 이고 $\pi_{m,\beta}(C_3 \subseteq C_4)$는 $(R^T \triangleleft_\beta R)_{34} = 0.42$이다. 각 분류간 함의 관계는 β-제약 퍼지 관계곱에 의해 <표 1>의 (a), (b)와 같이 설정될 수 있다.

다음에 $(R^T \triangleleft_\beta R)$ 를 α-cut 하여 크리스프 값으로 바꾼다. <표 2>의 (a)는 $(R^T \triangleleft_\beta R) = 0.82$ 로 α-cut 한 최종 결과이다. 즉 0.82 미만의 값은 0이 되고 원래 0.82 이상인 값은 1이 된다. <표 2>의 (b)는 $(R^T \triangleleft_\beta R) = 0.72$ 로 α-cut한 최종결과이다.

[그림 2]는 <표 2>에 의하여 얻어진 최종 결과로서 각 분류주제간 관계를 보여준다. α = 0.82일 때는 (a)와 같으며, α = 0.76 일 때는 (b)와 같다. [그림 2]의 (a)에서 α = 0.82 일 때 분류주제 간의 함의관계를 살펴보면, C_4 분류주제 항목은 모든 분류주제 항목의 상위분류이고, C_1, C_2, C_5 는 하위 분류주제이다. 또한, C_1은 C_2, C_5에, C_2은 C_5에 대해서 동시에 하위분류주제로 구성된다. 이것은 일반 고정분류체계의 배타적 개념 대신 공유개념과 다중계승의 개념이 도입된 것이다. 즉 분류주제 C_1에 의해 분류되는 메일들은 상위 분류주제로서 C_2, C_4, C_5 를 공유하게 된다. (b)에서는 α = 0.76 일 때 분류간의 함의관계를 살펴보면, C_4가 최상위 분류에 위치하며, α = 0.82 일 때의 분류주제관계를 모두 포함하면서 하위분류주제 C_1, C_2, C_3, C_5로 확장된 것을 알 수 있다.

[그림 2]와 같은 동적인 분류관계를 생성하면, 분류주제 안에 원하는 메일이 없을 때는 유사한 하위분류주제로 재구성 하여 검색할 수 있다. 그러나 위의 재분류는 분류내의 색인어들이 중복되어 나타나기 때문에 재현율은 높아지나 정확율이 낮아지는 문제가 있다. 이러한 문제를 해결하기 위해 분류 관계에 식(3)으로 유사도를 계산하여 유사도가 낮은 분류는 제거하고 유사도가 높은 분류는 분류간 합병하여 정확율을 높인다.

6. 실험 및 분석

본 논문에서는 펜티엄 III 1.2GHz, 640Mb RAM상의 윈도우 XP환경에서 Visual C++6.0을 사용하여 제시된 방법을 구현하여 실험하였다. 실

<표 1> 분류와 색인어의 β-제약 퍼지 관계곱

	B_1	B_2	B_3	B_4	B_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

(a) R^T

	C_1	C_2	C_3	C_4	C_5
C_1	0.98	0.46	0.76	0.24	0.78
C_2	1.00	1.00	0.96	0.62	1.00
C_3	0.98	0.64	1.00	0.42	0.76
C_4	1.00	0.82	0.80	1.00	1.00
C_5	1.00	0.82	0.76	0.62	1.00

(b) ($R^T \triangleleft_\beta R$)

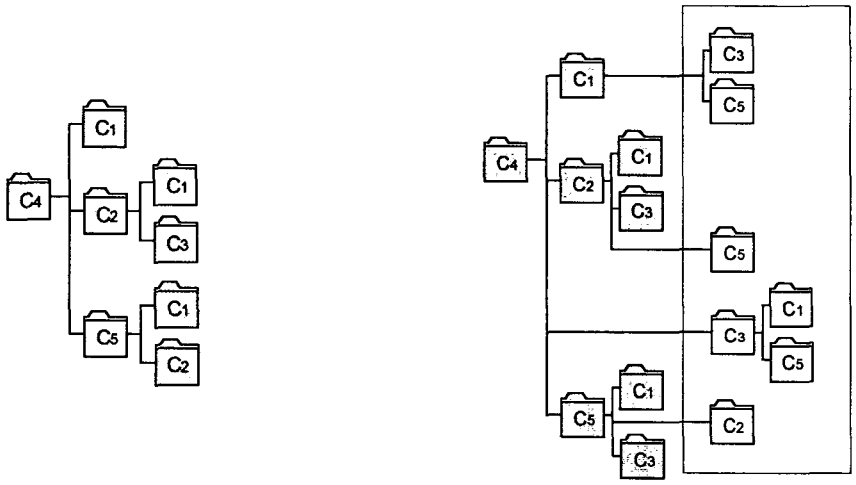
<표 2>

	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	1	0	1	1
C_5	1	1	0	0	1

(a) α = 0.82

	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	1	1	1	1
C_5	1	1	1	0	1

(b) α = 0.76



(a) a = 0.82 일 때의 분류 관계 (b) a = 0.76 일 때의 분류 관계

[그림 2] 최종결과의 분류 관계도

험 자료는 2004년 4월 27일~5월 22일의 4주 동안의 수신된 227개의 메일을 대상으로 하였다. 실험은 이메일 다원분류에 이용된 k-NN 분류방법과 동적분류체계방법을 비교하여 성능평가 하였다. 성능평가 방법은 문서분류에서 이용되는 재현율, 정확률, $F1$척도를 사용하였으며 식은 다음과 같다(Ricardo and Berthier, 1999).

<표 3> 분류의 적합성

분 류		전문가에 의한 분류	
		Correct	Incorrect
분류방법에 의한 분류	Correct	a	b
	Incorrect	c	d

재현율 $$(r)=\frac{a}{a+c} \quad (6)$$

정확률 $$(p)=\frac{a}{a+b} \quad (7)$$

$F1$척도 $$(F1(r,p))=\frac{2\times p\times r}{p+r} \quad (8)$$

실험은 동적분류체계 방법의 a 값 변화에 따른 이메일의 재분류에 대한 성능평가를 하였으며, k-NN에 의한 이메일 분류는 이웃하는 메일의 수인 k를 이용하였다. 제안방법과 k-NN에 의한 평가 결과는 <표 4>와 같다.

<표 4>에서 보는 바와 같이 k-NN을 이용하여 이메일 분류시 k값이 작을 수 록 재현율과 정확률이 높아지나, 재현율에 비하여 정확률은 상당히 낮다. 제안방법 역시 a 값 작을 수 록 재현율과 정확률이 높아지며, k-NN 방법에 비하여 정확률이 높아지는 것을 알 수 있다.

자동 카테고리 생성방법에서는 생성된 카테고리 수를 k로 설정하여 k-NN을 이용한 성능 평가를 비교하면 결과가 같기 때문에 자동 카테고리 생성방법과는 비교 하지 않았다. 그러나 자동 카테크고 생성 방법을 이용하여 이메일 분류후 동적분류체계방법을 이용하여 재분류하는 결과가 k-NN방법에서 k값을 조절하면서 재분류하는 것보다 더욱 향상된 성능을 보이는 것을 <표 4>의 비교 결과로 알 수 있다.

<표 4> 제안방법과 k-NN을 이용한 이메일 분류 결과

실험방법	k 및 a 값	재현율	정확률	$F1$척도	평균$F1$척도
k-NN	17	81.2	73.3	77.05	75.16
	18	79.4	72.1	75.63	
	19	78.57	70.6	74.37	
	20	78.32	69.4	73.59	
제안방법 (a 값)	20	86.2	84.5	85.34	83.55
	40	86.1	83.23	84.64	
	60	85.72	81.1	83.35	
	80	82.4	79.43	80.88	

7. 결론

본 논문에서는 이메일을 자동으로 분류하고 분류된 결과를 사용자의 요구사항에 맞게 재분류할 수 있는 방법을 제안하였다. 제안된 방법은 벡터 모델의 유사도를 기반으로 한 자동 카테고리 설정 알고리즘으로 분류될 주제 그룹을 자동 생성하며, 추출된 색인어의 유사도에 의해 메일을 자동 분류하였다. 또한 동적분류체계 방법을 이용하여 사용자가 필요하면 언제든지 재분류할 수 있다. 이러한 재분류는 사용자의 요구사항에 맞추어 조정할 수 있게 하여 효율적으로 이메일을 관리할 수 있고, 검색시 정확률을 높였다. 마지막으로 학습이 필요 없이 메일을 빠르게 분류함으로 유동적인 이메일 환경을 만족시킨다.

Acknowledgement

본 연구는 대학 IT연구센터 육성 지원사업의 연구결과로 수행되었음.

참고문헌

Androutsopoulos, I., "An Evaluation of NaveBayesian Anti-Spam Filtering", *In Proc. Workshop on Machine Learning in the New Information Age*,(2000).

Bandler, W., and Kohout, L., "Semantics of Implication Operators and Fuzzy Relational Products", *International Journal of Man-Machine Studies*. Vol. 12(1980) 89-116.

Choi, B.G., J.H. Lee and S. Park, "Dynamic Construction of Category Hierarchy Using Fuzzy Relational Products", *In proceddings of the 4th International Conference On Intelligent Data Engineering and Automated Learing. Hong Kong, China*, (2003), 296-302.

Cohen, W. W, "Learning Rules that classify E-mail", *In Proc. AAAI Spring Symposium in Information Access*, (1999).

Drucker, H., Wu, D., and Vapnik, V. N., "Support Vector Machines for Spam Categorization", *IEEE Transactions on Neural network*, Vol. 10 No. 5(1999).

Kang, S. S., *Korean Information Retrieval and Morpheme analysis*, HongReung Science

Publishing Co., 2002.

Korfhage, and Robert. R., *Information Storage and Retrieval*, WILEY, 1997.

Kun-Lun, L., Kai, Li., Hou-Kuan, H., and Sheng-Feng, T., "Active Learning with Simplified SVMS for SPAM Categorization", *In Proc. First Conf. On Machine Learning and Cybernetics, Beijing*, (2002), 4-5.

Manco, G. and Masciari, E., "A Framework for Adaptive Mail Classification", In Proceedings of the 14thIEEE International Conference on Tools with Artificial Intelligence, (2002).

Mock, K., "Dynamic Email Organization via Relevance Categories", *In Proceedings of the International Conference on Tools with Artificial Intelligence* (1999).

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

Ricardo , B. Y., and R. N. Berthier, *Modern Information Retrieval*, Addison Wesley, 1999.

Sakkis, G., "Stacking classifiers for anti-spam filtering of e-mail", *In Proc. 6thConf. On Empirical Methods in Natural Language Processing*,(2001).

Woitaszek, M., and shaaban, M., "Identifying Junk Electronic Mail in Microsoft Outlook with a Support Vector Machine", *In Proc. 2003 Symposium. On Application and the Internet*, (2003).

Abstract

Classification of e-mail Using Dynamic Category Hierarchy and Automatic category generation

Sun Park* · Chan Min Ahn* · Sang Ho Park* · Ju-Hong Lee* · Bum-Ghi Choi**

Since the amount of E-mail messages has increased , we need a new technique for efficient e-mail classification. E-mail classifications are grouped into two classes: binary classification, multi-classification. The current binary classification methods are mostly spam mail classification methods which are based on rule driven, bayesian, SVM, etc. The current multi- classification methods are based on clustering which groups e-mails by similarity. In this paper, we propose a novel method for e-mail classification. It combines the automatic category generation method based on the vector model and the dynamic category hierarchy construction method. This method can multi-classify e-mail automatically and manage a large amount of e-mail efficiently. In addition, this method increases the search accuracy by dynamic reclassification of e-mails.

Key words : E-mail classification, dynamic category hierarchy, automatic classification

* School of Computer Information Science and Engineering, In-Ha University
** Quark Co., Ltd.