분류 주제 자동 생성 및 동적분류체계 방법을 이용한 이메일 분류

안찬민[O] 박선 박상호 최범기 이주홍
인하대학교 컴퓨터정보공학과
{ahnch1[O], sunpark, parksangho, bgchoi}@datamining.inha.ac.kr jhlee@inha.ac.kr

E-mail Classification Using Dynamic Category Hierarchy and Automatic Generation of Category Label

C.M. Ahn[O] S. Park, S.H. Park, B.K. Choi, J.H. Lee
School of Computer Science & Engineering, INHA Univ.

요 약

이메일 사용이 보편화 됨에 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 묶는 수준에 그치고 있다. 본 논문에서는 벡터모델의 유사도를 기반으로 한 분류 주제 자동 생성 알고리즘과 동적분류체계 방법을 결합하여 새로운 이메일 자동 다원분류 방법을 제안했다. 본 논문에서 제안한 방법은 이메일을 자동으로 분류하며, 분류된 결과를 색인검색과 디렉토리 검색 방법을 지원하여 대량의 메일도 효율적으로 관리할 수 있다. 또한 메시지를 동적으로 재분류 할 수 있게 함으로써 디렉토리 검색시 재현율을 높였다.

1. 서 론

이메일은 시간과 비용이 효율적인 통신방법으로 널리 사용되고 있으며, 점차 중요성이 증가하고 있다. 이메일은 일반 사용자 뿐 아니라 전자상거래, 광고, 사업등에 사용되고 있다. 기업 및 일반사용자가 받는 메일의 하루량은 수십에서 수천통에 이른다. 이렇게 수신되는 이메일이 증가하는 추세에 따라 메일 수신자들은 수신되는 대량의 이메일로부터 스팸(spam)과 스팸이 아닌 메일을 구분하고, 스팸이 아닌 메일은 다시 뉴스레터(newsletter), 메일링리스트(mailing-list), 수신자에게 중요한 메시지(important message)로 분류하고 정리하는데 많은 시간을 소비하고 있다.

이를 위해 본 논문에서는 벡터모델의 유사도를 기반으로 한 분류 주제 자동 생성 알고리즘과 동적분류체계 방법을 결합하여 새로운 이메일 자동분류 방법을 제안한다. 본 논문에서 제안한 방법은 다음과 같은 장점을 가진다. 첫째 제시된 알고리즘에 의해 메일이 분류될 메시지의 분류주제가 자동 생성됨으로 사용자의 간섭이 필요없다. 둘째, 동적분류체계 방법을 이용하여 사용자가 필요하면 언제든지 재분류 할 수 있다. 셋째, 대량의 메일을 효율적으로 관리할 수 있도록 색인검색과 디렉토리 검색방법을 지원한다. 넷째, 학습이 필요 없이 메일을 빠르게 분류함으로서 유동적인 이메일 환경에 적합하다.

본 논문의 구성은 다음과 같다. 2 장에서는 유사도 계산을 위한 벡터모델을 보인다. 3 장은 동적분류체계 방법에 대하여 알아본다. 4장은 제안된 이메일 자동분류 방법인 분류 주제 자동 생성 알고리즘과 동적인 분류체계를 구성하는 방법을 설명한다. 5절에서는 실험 및 평가결과를 보인다. 6절에서 결론을 맺는다.

2. 벡터 모델

본 장에서는 본 논문에서 유사도 계산을 위해 사용하는 벡터 모델인 *tf-idf* 기법에 대하여 알아본다. [3]

(정의 1) 검색문서 j내의 단어 i의 가중치 w_{ij}와 질의문서 q 내의 단어 i 의 가중치 w_{iq} 으로 *tf-idf* 방법으로 계산되는데, *tf-idf* 란 tf(term frequency)와 idf(inverse document frequency)를 곱한다는 의미이다. *tf*, 즉 단어빈도는 문서에서 단어가 나타나는 빈도를 의미하며 식(1)과 같이 계산된다.

$$f_{i,j} = \frac{freq_{i,j}}{\max_l \times freq_{l,j}} \qquad (1)$$

여기서, f_{ij}는 검색문서 j 내의 단어 i의 출현 회수, max_l은 검색문서 j 에서 가장 많이 출현한 단어의 출현 회수이다.

idf, 즉 역문서빈도는 식 (2)와 같이 계산되는데, 적은 개수의 문서에 걸쳐 나타난 단어의 가중치는 높이고, 많은 개수의 문서에 걸쳐 나타난 단어의 가중치는 낮추는 효과를 준다.

$$idf_i = \log \frac{N}{n_i} \quad (2)$$

질의문에서 q 내의 단어 i 의 가중치 w_{iq} 도 w_{ij} 를 구할 때와 마찬가지로 식(3)과 같이 계산한다.

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \quad (3)$$

3. 동적분류체계 방법

이전에 우리가 제안한 동적분류체계 방법[4]은 검색어와 분류 간의 관계를 규정하고, 분류들 간의 상호 관계를 규명함으로써 분류검색의 분류체계를 자동으로 동적인 체계로 재구성함으로써 검색효율을 높이는 방법이다. 분류와 검색어간의 관계는 문서에서의 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[3].

(정의 2) 퍼지 함의 연산자 (Fuzzy Implication Operator)는 크리스프 함의 연산자 (Crisp Implication Operator)를 확장하여 퍼지에 적용한다. 퍼지 함의 연산자의 종류는 무수히 많으며 대표적인 Kleene-Diense 퍼지함의 연산자의 예는 다음과 같다[2].

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

검색어와 분류간의 관계를 도출해 내는 방식은 여러 가지가 있을 수 있다. 본 논문에서는 분류에 속한 각 문서들에 대한 검색어들과의 관계를[1] 종합하여 만들어진다. 이렇게 생성된 분류들은 검색어들을 요소로 하는 퍼지 집합이 된다. 두 분류간의 관계는 생성된 두 분류의 퍼지 집합의 함의 정도를 계산하여 결정할 수 있다. 이를 이용하여 임의의 두 분류의 유사관계를 동적으로 생성할 수 있는 자동화된 시스템을 구성할 수 있다.

본 논문에서는 위의 식(4)의 Kleen-Diense 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식(5)의 퍼지관계곱을 적용하여 분류들 간의 퍼지함의관계, $C_i \rightarrow C_j$ 를 유도할 수 있다. 그러나 C_i 에 멤버쉽 값($\mu_{C_i}(x)$)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, $\pi_{m,\beta}(C_i \subseteq C_j)$, 를 계산한다.

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

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

4. 자동 이메일 분류 방법

4.1 전처리

본 논문에서는 이메일 분류 시스템의 구현상의 부담을 줄이고자 이미 개발되어 있는 한글분석 HAM 을 사용하였다. HAM 은 C 언어로 제작된 쉐어웨어(Shareware)로서 형태소 분석기를 기반으로 한 자동색인, 맞춤법 검사, 구문 분석, 복합명사 분해, 자동띄어쓰기 등의 기능들을 지원하는 한국어 분석용 커널 라이브러리이다[5].

이렇게 전처리에서 추출한 제목과 본문의 색인을 이용하여 대량의 메일 메시지로부터 색인 검색을 할 수 있다.

4.2 분류 주제 자동 생성 알고리즘

본 절에서는 분류주제 자동 생성 알고리즘을 제안하다. 이 알고리즘은 메시지에 포함된 색인어 중 유사도가 가장 높은 색인어를 대표로 지정하여 분류주제를 자동으로 생성하고 수신받는 메시지를 생성된 분류주제별로 자동 분류하는 알고리즘이다. 이 알고리즘에서 사용되는 입출력 값은 다음의 벡터집합으로 정의한다. 전처리된 e-mail 자료로부터 각각의 메시지 m를 $T = \{ d, \{S\}, \{B\} \}$로 표현한다. 여기서, d는 수신자의 이름, S는 메시지 m에 포함된 제목의 색인벡터, B는 메시지 m에 포함된 본문 색인들의 벡터이다. s는 메시지 m에 포함된 주제 색인어의 집합 $S = \{ s_1, s_2, \ldots, s_k \}$의 형태를 가지며, k는 m에 포함된 주제 색인어의 총 개수이다. B는 메시지 m에 포함된 본문의 색인어의 집합 $B = \{ b_1, b_2, \ldots, b_l \}$의 형태를 가지며, l은 m에 포함된 본문 색인어의 총 개수이다. 트랜잭션 T의 집합 M은 $M = \{ T_1, T_2, \cdots, T_i \}$이다. R는 메시지 m에 포함된 대표 색인어 r들의 집합 $R = \{ r_1, r_2, \ldots, r_n \}$, C는 분류 레이블 c들의 집합 $C = \{ c_1, c_2, \ldots, c_o \}$이다.

Algorithm. Automatic Generation of Category Label
Input : M ,
Output : R,C
for ($i = 1$; $T_i \neq \emptyset$; i++) // 대표색인어 설정
the setting of the representative index term
for ($j = 1$; $s_j \neq \emptyset$; j++) // 제목과 본문 색인간의 유사도 계산
To calculate the similarity of the index term between Subject and Body

```
for ( i = 1 ; T_i ≠ Ø ; i++ ) {
        initial category label
        if ( max >= min_threshold ) m_i ⊂ c_o
         Classify the message to category label
         // 메시지를 분류레이블로 분류
        else
         The setting of the category label from the representative
        index term
         // 대표색인에 의한 분류주제 설정
}
```

그러나, 만일 메시지의 제목이 아무런 의미도 갖지 못함은 물론 메시지의 의도 조차도 내포하지 못한다면 식 (1)을 사용한 방법은 불필요하거나 메일을 잘못 분류할 수 있다. 또한 메시지 내용이 제목과 유사한 내용이라도 중요한 의미를 담고 있는 특질(Feature)을 포함하고 있지 않다면 중요한 문장이 될 수 없으며, 반대로 제목과 유사성이 없는 내용이라도 중요한 의미를 포함한 특질등이 나타나는 내용이라면 중요하게 고려해야 한다.
본 논문에서는 이러한 문제를 해결하기 위해 분류 주제 자동 생성 알고리즘으로 분류되어진 결과를 동적분류체계 방법을 이용하여 동적으로 재분류 할 수 있게 하였다.

4.3 동적분류체계 방법에 의한 이메일 재구성

본 논문에서는 이메일을 동적분류체계로 구성하기 위해 색인어와 분류주제 간의 관계를 규정해야 한다. 그러나 색인어와 분류주제 간의 관계를 직접 결정할 수는 없으므로 색인어와 메일 간의 관계 및 메일과 분류주제 간의 관계에 의해서 결정 한다. 이러한 관계는 4.2 절의 자동 분류주제 생성 알고리즘으로 유도할 수 있다. 여기서, 메시지를 색인어로 구성된 퍼지 집합으로 간주 할 수 있고, 마찬가지 분류주제를 분류된 메시지들의 색인어들로 구성된 퍼지 집합으로 간주할 수 있다. 메시지가 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.

5 실험 결과

본 논문에서는 Visual C++6.0 과 Visual Basic 을 사용하여 프로토타입을 구현하였다. 실험은 자동 분류주제 생성 알고리즘에 의해 이메일의 분류정확률을 평가한다. 평가자료는 2 개월 동안의 수신된 267 개의 메일을 대상으로 하였다. 실험은 제안된 알고리즘의 분류경계값 (min_threshold)의 변화에 따라 자동으로 생성되는 분류주제와의 분류정확률을 평가한다.

분류정확률은 메시지가 관련된 분류주제에 속하는지를 분류전문가에 의해 평가하였다. 그림 1은 분류경계값이 작아질수록 분류주제는 많아지면서 정확률도 높아진다는 것을 보여준다.

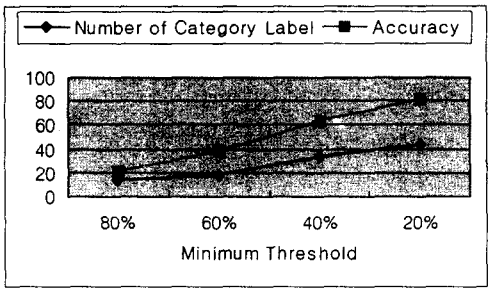

그림 1. 분류경계값에 따른 카테고리의 수와 정확률

6 결론

이 논문에서 우리는 이메일 메시지를 자동으로 분류할 수 있는 방법을 설명 및 구현하였다. 우리의 방법은 자동 분류 주제 생성 알고리즘과 동적분류계층 기술을 기반으로 이메일을 분류하였으며, 다양한 실험을 통하여 메시지 분류 및 검색 시 높은 유연성과 효율성을 보였다. 본 논문에서 제안한 방법은 다음과 같은 장점을 가진다.
1) 제시된 알고리즘에 의해 메일이 분류될 메시지의 분류 주제가 자동 생성됨으로 사용자의 간섭이 필요 없다.
2) 우리의 방법은 동적분류체계 방법에 의해 사용자가 이메일의 재분류와 디렉토리 검색을 쉽게 할 수 있다.

참고문헌

[1] Alrashid, T. M., Barker, J. A., Christian, B. S., Cox, S.C., Rabne, M. W., Slotta, E. A. and Upthegrove,. L. R., " *Safeguarding Copyrighted Contents, Digital Libraries and Intellectual Property Management, CWRU' s Rights Management System.*" D-Lib Magazine, April 1998.

[2] Ogawa Y., Morita T., and Kobayashi K., " *A fuzzy document retrieval system using the keyword connection matrix and a learning method.*" Fuzzy Sets and System, pp. 163-179, 1991.

[3] Baeza-Yates R. and Ribero-Neto B. " *Modern Information Retrieval*" Addison Wesley, 1999.

[4] Bumghi C., Ju-Hong L., Sun P. " *Dynamic Construction of Category Hierarchy Using Fuzzy Relational Products.*" In proc. of the 4th International Conference On Intelligent Data Engineering and Automated Learing. Hong Kong, China, pp.296-302, 2003.

[5] Kang S.S., Korean Information Retrieval and Morpheme analysis. Hong-Reung Science Publishing Co., 2002.