# 이메일 추천 시스템의 분류 향상을 위한 3단계 전처리 알고리즘

論 文
54D-4-6

# A Three-Step Preprocessing Algorithm for Enhanced Classification of E-Mail Recommendation System

鄭 玉 蘭† · 趙 東 燮*
(Ok-Ran Jeong · Dong-Sub Cho)

**Abstract** - Automatic document classification may differ significantly according to the characteristics of documents that are subject to classification, as well as classifier's performance. This research identifies e-mail document's characteristics to apply a three-step preprocessing algorithm that can minimize e-mail document's atypical characteristics. In the first stage, uncertainty based sampling algorithm that used Mean Absolute Deviation(MAD), is used to address the question of selection learning document for the rule generation at the time of classification. In the subsequent stage, Weighted vlaue assigning method by attribute is applied to increase the discriminating capability of the terms that appear on the title on the e-mail document characteristic level. in the third and last stage, accuracy level during classification by each category is increased by using Naive Bayesian Presumptive Algorithm's Dynamic Threshold. And, we implemented an E-Mail Recommendtion System using a three-step preprocessing algorithm the enable users for direct and optimal classification with the recommendation of the applicable category when a mail arrives.

**Key Words** : Automatic Document Classification, Weighted Value Assigning Mmethod, Naive Bayesian Presumptive Algorithm, Dynamic Threshold, E-Mail Recommendation System

## 1. 서 론

현재 보편적으로 사용되고 있는 통신상의 문서들은 비정형적인 특성을 많이 포함하고 있다. 따라서 일반적인 문서 자동분류에서는 정제되고 정형화된 문서를 대상으로 분류를 수행한다. 특히, 이메일 문서와 같은 속어 및 약어의 빈번한 사용과 자유로운 문체로 인하여 비정형성이 어느 다른 문서들 보다도 크게 나타나고 있다. 따라서 이러한 문서들을 대상으로 하는 분류의 경우, 분류의 정확도가 떨어지게 된다. 본 연구에서는 이러한 이메일 문서의 특성을 파악하여 이메일 문서의 비정형적인 특성을 최소화함으로써 전체적인 분류 성능의 향상을 위한 세가지 전처리 알고리즘을 제안한다. 문서자동 분류의 과정은 크게 전처리 단계, 특징 추출 단계, 문서 분류 단계로 구분된다. 그러나 본 연구에서는 실제 분류 규칙을 적용하여 분류하는 전단계인 전처리 단계, 특징 추출 단계를 모두 합쳐서 전처리 단계라 정의한다.

본 연구에서는 이메일 문서의 비정형적인 특성을 최소화하여 사용자에 맞게 이메일시스템의 카테고리별 정확한 자동 분류을 위하여 3단계 전처리 알고리즘을 단계별로 제안하였다.

첫단계에서는, 특징추출 단계에서 문서분류을 위한 학습문서를 선택하게 되는데, 임의의 학습 문서집합을 있는 그대로 이용하지 않고 지능적으로 재구성하는 평균절대편차값(MAD:Mean Absolute Deviation)을 이용하는 불확실성 기반 샘플링 알고리즘을 적용한다. 이는 물론 기존의 방법을 이용한것이나, 다음단계의 전처리 알고리즘을 적용하기 전단계로 이용했을 때 성능 향상에 도움이 되었다.

두 번째단계에서는 이메일 문서의 편중성을 고려하여 속성별 가중치를 부여하여 특징추출을 한다. 이메일 문서는 제목과 본문으로 구성되어 있는데, 확장된 Naive Bayesian Classifier를 적용하여 제목 부분에 본문 부분보다 가중치를 더 주는 방식이다.

세 번째단계에서는 문서 분류의 정확도를 결정하는 것은 규칙형성하는 추정알고리즘을 들 수 있다. 학습문서집합 구성방법에 의해 채택된 학습문서집합을 이용하여 최종적으로 규칙을 형성하는 것이 이 알고리즘의 역할인데, 기존의 고정 임계치를 개선시킨 동적 임계치를 이용한 Naive Bayesian 알고리즘[1]를 적용하였다.

일반적인 문서분류시 이용되는 추정알고리즘에는 의사결정트리(decision tree)알고리즘, k-최단 인접기법(k-nearest neighbor)알고리즘, 신경망(neural network)알고리즘, Naive Bayes알고리즘, 그리고 지지벡터기계(support vector machine)들이 있다[2,3] .

† 교신저자, 正會員 : 이화여자대학교 공대 컴퓨터학과 박사과정
E-mail : orchung@ewhain.net
** 正 會 員 : 이화여자대학교 공대 컴퓨터학과 정교수
接受日字 : 2004年 9月 13日
最終完了 : 2005年 2月 18日

Naive Bayes 문서분류 알고리즘을 이용한 문서 분류 시스템은 일반적으로 다른 알고리즘들에 비해 문서분류의 정확도가 상대적으로 높다고 알려져 있다[3]. 또한 이 알고리즘에 따르면 분류를 위한 규칙형성이 간단하며, 문서분류의 속도가 상대적으로 빠르기 때문에 문서분류시스템을 구축하는데 매우 빈번히 이용된다. 따라서 나이브 베이지안 알고리즘을 토대로하여 이메일 문서 분류의 정확도를 높이는 작업은 큰 의미가 있다고 할 수 있다. 기존의 알고리즘은 고정 임계치(threshold)를 사용하였는데, 본 연구에서는 임계치를 동적 임계치[1]로 개선하여 문서분류의 정확도를 높이고자 하였다.

본문은 다음과 같이 구성되어 졌다. 2장에서는 이메일 문서의 특성에 따른 전처리 알고리즘을 설명하였으며, 3장에서는 문서 분류의 정확도를 향상시키기 위해 제안된 3단계 전처리 알고리즘를 기술한다. 4장에서는 제안된 전처리 알고리즘을 적용하여 구현된 이메일 추천 시스템을 보여주고, 이를 기반으로 실험 및 결과 분석을 한다. 5장에서 결론을 맺는다.

## 2. 이메일 문서의 전처리 과정

특징추출을 위한 전처리 과정은 우선 문서의 정보를 표현하는데 있어 불필요한 요소인 특수기호를 제거한다. 그리고, 제목과 본문으로 구분하여 토큰을 생성하고 이를 정렬한다. 이단계에서 1바이트 기호들과 태그가 제거된다. 다음 단계로 전처리기는 약어, 속어 및 특정 집합에 속하는 표현들을 표준화하는 표현인 대표어에 대한 사용자 사전을 로딩하여 정렬된 상태로 유지하게 된다.

다음 단계에서는 불필요한 기호 및 태그정보를 제거하여 만들어진 벡터화된 이메일 문서와 로딩된 사용자 사전을 매핑하여 이메일문서에 나타나는 비정형화된 단어들을 정형화된 단어로 변환한다. 이러한 특징추출과정에서 본 연구에서는 제목부분의 토큰을 따로 색인어로 추출하여 본 연구에서 제안한 속성별 가중치 부여 방식으로 가중치를 부여한다.

이 단계에서는 불필요한 기호 및 태그 정보를 제거하여 만들어진 벡터화 된 이메일문서와 로딩된 사용자 사전을 매핑하여 이메일문서에 나타나는 비정형화된 단어들을 정형화된 단어로 변환한다. 특징추출단계에서는 이렇게 만들어진 학습무서집합을 수치화된 벡터로 표현한다. 이는 추출된 색인어가 문서에 나타나는 빈도수와 색인어가 나타난 전체 문서의 개수를 기반으로 계산되는 $tf-idf$값에 의하여 구해진다. $tf-idf$는 키워드 집합을 수치화된 벡터로 표현하는 모듈로 대부분의 정보검색 시스템이나 문서 분류 시스템에서 사용되는 방법이다.

문서 $d_i$의 $j$번째 키워드의 가중치 $w_{ij}$는 다음과 같이 표현된다[3,4].

$$w_{ij} = tf_{ij} \log(\frac{N}{df_i}) \quad (1)$$

여기서 $tf_{ij}$는 $j$번째 키워드의 문서 $i$에서의 빈도수이며, $df_j$는 키워드 $j$가 전체 문서집합에서 나타나는 문서의 개수이다. $N$은 전체 문서의 개수를 말한다.

이렇게 얻어진 문서의 수치벡터는 색인어 사전에 존재하는 총 색인어 개수 만큼의 차원을 갖는다. 한편 벡터의 차원이 크면 학습의 속도가 늦어지고, 문서분류 성능 또한 나빠지게 되는데, 이러한 문제를 극복하기 위해 수치벡터의 차원을 적당한 크기로 축약하는 과정이 필요하다.

## 3. 가중치와 임계치를 이용한 3단계 전처리 알고리즘

문서 자동 분류는 분류기 자체의 성능 뿐만 아니라 분류의 대상이 되는 문서의 특성에 의해 그 결과가 크게 좌우될 수 있다. 따라서 일반적인 문서 자동 분류에서는 정제되고 정형화된 문서를 대상으로 분류를 수행한다. 그러나 현재 보편적으로 사용되고 있는 통신상의 문서들은 비정형적인 특성을 많이 포함하고 있다. 특히 이메일문서와 같은 경우 속어 및 약어의 빈번한 사용과 자유로운 문체로 인하여 그 비정형성이 어느 다른 문서들 보다도 크게 나타나고 있다. 따라서 이러한 문서들을 대상으로 하는 분류의 경우, 분류의 정확도가 떨어지게 된다.

본 논문은 이러한 이메일문서의 특성을 파악하여 이메일 문서의 비정형적인 특성들을 최소화하기 위하여 단계별 3가지 전처리 알고리즘을 제안함으로써, 전체적인 분류성능의 향상을 꾀하는데 목적이 있다. 먼저 학습단계에서는 기존의 불확실성기반 샘플링 알고리즘을 적용하여 규칙형성시 좋은 학습 집단을 선정한다.

다음으로는 본 연구에서 제안한 두가지 방법의 전처리 알고리즘인 속성별 가중치 부여 기법과 동적 임계치를 이용한 문서 분류 알고리즘이 차례로 적용되는 것이다. 기존의 속성별 가중치 부여 기법은 이메일 특성상 제목부분에 두배나 아님 서버 분석을 한 후 서버관리자가 수동으로 가중치를 부여하는 방법들이 제안되어 왔다. 이를 이메일 시스템에서 자동으로 분석한 후 가중치를 계산하여 부여하는 것이다.

마지막 3단계에서는 베이지안 분류 알고리즘은 문서 분류에서 기본적으로 사용되고 있고 여러 가지 장점에 인해 본 연구에서 기본적 베이지안 분류 알고리즘을 적용하였으나, 알고리즘 자체에서 임계치를 정하는 부분을 문서의 확률값의 분포에 따라 동적으로 변화시키는 것이다[8].

본 연구의 전단계에서 우리는 기존의 전처리 알고리즘을 이메일 카테고리별 분류에 적용하여 그 특성을 파악한 후, 각각 가중치 부여 기법과 동적 임계치를 적용한 기법을 제안하여 독립적으로 연구 진행을 해왔다[1,9].

그러나, 이 방법들을 단계별로 연계하여 적용한다면 좀 더 나은 결과를 볼 수 있다는 점을 착안하여 3단계를 하나의 전처리 알고리즘의 개념으로 설정하고 실험 및 결과 분석하였다.

### 3.1 불확실성 기반 샘플링 알고리즘 (Uncertainty based Sampling Algorithm)

능동적 학습 알고리즘은 전체문서집합으로부터 정보량이 큰 문서를 선택하여 이를 규칙 형성을 위한 학습문서집합에 추가한다. 잘 된 규칙은 분류시 정확도를 높이는데 가장 큰 역할을 하므로, 규칙형성을 위한 학습문서집합은 매우 중요하다고 할 수 있으며, 이때 불확실성 기반 샘플링 알고리즘을 적용하는 것이다.

즉 능동적 학습 알고리즘에서 정보량이 큰 문서를 판단하는 기준은 여러 가지로 다양한데 그중 대표적인 것이 불확실성 개념을 이용하는것이다[5,6,7]. 능동적학습 알고리즘 중에서 이 개념을 이용하는 알고리즘을 특히 불확실성 기반 샘플링 알고리즘이라고 한다. 이에 따르면 정보량이 큰 문서는 현재의 문서분류함수가 분류하기 어려운 문서이다. 문서분류함수가 분류하기 어려운 문서란, 문서분류함수가 입력으로 주어진 문서의 카테고리를 판단할 때 확신이 작은 문서를 의미한다. 이러한 문서는 카테고리와 카테고리를 나누는 경계부근에 위치하고 있기 때문에 확실히 어떠한 카테고리에 할당되어야 하는지 판단하기 어려운 성질을 지닌다.

불확실성이란 특정 문서가 문서분류함수에 의해 분류되는 경우, 얼마나 불명확하게 분류되는가를 측정한 수치를 말한다[6]. 따라서 불확실성이 클수록 문서분류함수가 해당 문서를 카테고리로 분류하는 확신이 작다. 위에서 살펴보았듯이 불확실성 기반 샘플링 알고리즘은 라벨이 없는 문서집합 내의 모든 문서의 불확실성을 측정한 후에 가장 불확실성이 큰 문서를 골라서 이를 학습문서로 채택한다. 불확실성은 문서의 카테고리를 예측하고, 이 예측에 대한 확신을 수치로 나타낼 수 있는 문서분류 알고리즘에서는 모두 정의가 가능하다. 여기서는 Naive Bayes 문서분류 알고리즘에서 불확실성을 특정하는 두가지 측정치를 소개한다.

첫 번째로 신뢰도(confidence) 측정치를 살펴보자. 문서 $x$가 카테고리 $c_i$로 할당되는 경우의 신뢰도는 다음과 같이 정의된다[7].

$$U_{confidence}(x) = \frac{P(c_i \mid x) - P(c_j \mid x)}{P(c_i \mid x)} \quad (2)$$

이 식에서 문서 $x$가 카테고리 $c_i$에 속할 확률을 의미한다. 이는 다음과 같은 성질을 갖는다.

$$P(c_1 \mid x) + P(c_2 \mid x) + \cdots + P(C_{|d|} \mid x) = 1,$$
$$\text{모든 } c \in C\text{에 대하여}, 0 \le P(c \mid x) \le 1 \quad (3)$$

식(2)에서 $c_i$는 문서 $x$와 가장 가까운 카테고리이다. 즉, $c_i$는 $P(c \mid x)$ 값을 가장 크게 하는 카테고리이다. 또한 $c_j$는 문서 $x$와 두 번째로 가까운 카테고리이다. $P(c_i \mid x) - P(c_j \mid x)$ 값이 클수록 현재 문서분류함수가 분류 결과에 대한 확신을 크게 갖고 있다는 의미이므로, 식(2)의 신뢰도 값은 커진다. 이와 같이 신뢰도와 불확실성은 반비례의 관계에 있다. 신뢰도가 크다는 것은 문서분류함수가 문서를 정확하게 분류할 가능성이 크다는 것을 의미하기 때문이다. 그러므로 신뢰도의 역수를 이용하여 불확실성을 측정할 수 있다.

두 번째로 평균절대편차(MAD, Mean Absolute Deviation)를 이용한 불확실성 측정치를 알아보자[6]. 이 측정치에서는 앞에서 정의한 $P(c \mid x)$의 값들이 그 값들의 평균($\mu$)과 얼마나 떨어져 있는지를 이용하여 불확실성을 측정한다. 이는 다음과 같이 정의된다.

$$U_{MAD}(x) = \frac{1}{|C|}\sum_{i=1}^{|C|}|(P(c_i \mid x) - \mu)| \quad (4)$$

$$\mu = \frac{1}{|C|}\sum_{i=1}^{|C|} P(c_i \mid x) \quad (5)$$

$U_{MAD}(x)$는 문서 $x$가 각 카테고리에 속할 확률 또는 소속값들이 그 값들의 평균으로부터 떨어진 평균거리를 의미하므로 $U_{MAD}(x)$가 작을수록 불확실성이 크다고 할 수 있고, 클수록 불확실성이 작다고 할 수 있다.

다음장의 실험에서 볼 수 있듯이 불확실성 기반 샘플링 알고리즘을 Naive Bayes 문서 분류 알고리즘에 적용하여 문서분류함수의 정확도를 크게 향상시켰다. 불확실성의 측정치로서 앞에서 설명한 평균절대편차(MAD) 측정치를 이용했을 경우의 문서분류 정확도가 신뢰도 측정치를 이용한 경우보다 더 컸다.

### 3.2 이메일 문서의 특성을 고려한 속성별 가중치 기법

앞절의 알고리즘을 이용하여 학습문서를 선택한 후 다음 단계로 문서분류시 기준이 되는 규칙형성(Rule Generation)시 이메일 문서의 특성을 고려하여 제목부분에 가중치를 부여하고자 하였다. 다른 문서와는 달리 메일은 제목부분을 보게 되면 그 내용과 중요도를 거의 알수있는게 대부분이다. 즉 속성별 가중치부여기법을 이용하여 제목에 본문보다 몇 배의 가중치를 부여할 것인지를 결정하여 규칙형성단계에서 제목에 가중치가 부여되는 것이다.

전처리 단계에서는 특수기호 및 태그의 제거, 속어 및 약어에 대한 표준어 변환, 분류 대상 문서의 성격에 따른 대표어 변환과 불용어 제거 등을 처리하여 비정형화된 문서를 정제하고, 제목에 있는 단어들에 대하여 높은 가중치 부여로 분류의 정확도를 높이고자 한다. 기존의 이메일 문서 분류에 적용된 경우에는 메일의 제목에 변별력을 높여 가중치를 부여하기 위해 단순히 이중화(duplication)하는 방법[8]을 이용하였다. 본 연구에서는 제목에 본문보다 몇배의 가중치를

부여할 것인지를 다음 식(7)의 결과로 제목에 있는 키워드에 가중치가 부여되는 것이다.

가중치를 부여하는 방법으로는 Naive Bayesian Classifier를 기반으로 모든 속성값을 독립이며 분류에 동등한 영향력을 끼칠 것으로 가정하고 응용하였다. 즉 제목부분에 있는 속성별로 서로 다른 가중치를 줄 수 있도록 기존의 Naive Bayesian Classifier를 다음과 같이 확장하였다. 단, 모든 속성값들이 독립이라는 가정은 같다.

카테고리 $C_j$에 속하는 메일 문서중에서 제목에 있는 키워드 $a_k$에 해당하는 속성값 $V$는 $Doc(V)$로 파싱되어 개개의 속성값에 대한 키워드 $v_i$로 변경하여 사용한다. 다음은 문서분류에 응용되고 있는 Classifier model이다.

$$V_{NB} = \arg\max_{c_j \in C} \{ P(c_j) \prod_i P((a_k, v_i)|c_j) \} \quad (6)$$

$sa_k$는 해당$v_i$가 속한 속성값을 의미하며 제목 속성에 가중치를 부여하여 규칙형성시 다음 식(7)을 적용하였다.

$$P(v_i|\, c_j, a_k) P(a_k) \leftarrow \frac{n_{(a_k, l)} + 1}{n_{a_k} + |Doc_{sa_k}|} \times w_{sa_k} \quad (7)$$

$n_{a_k}$는 학습 데이터 내 전체 메일문서에서, 카테고리 $c_j$에 포함된 메일문서의 속성 $a_k$에 해당하는 모든 속성값 키워드의 총 개수를 의미한다. $Doc_{sa_k}$는 속성 $a_k$에 대하여 분류를 대표하기 위해 메일문서에서 선출한 모든 속성값 키워드의 수를 의미하며, $n_{(a_k, l)}$는 카테고리 $c_j$ 내에서 속성 $a_k$에 속한 속성값 키워드 $v_i$ 의 출현 빈도수(frequency)를, $w_{a_k}$는 제목속성 $sa_k$ 에 대한 개별 가중치를 나타낸다.

즉 이메일문서의 제목에 나타난 단어들의 변별력을 높이기 위해 제목의 키워드와 유사한 단어가 본문에 나타나는 빈도수만큼 가중치를 부여하는 방법이다. 실험결과는 4장에서 각각 전처리 단계의 성능을 비교 분석해보았다.

### 3.3 동적 임계치를 이용한 베이지안 알고리즘

나이브 베이지안 알고리즘에서 사용되는 기존의 고정된 임계치를 동적으로 개선하여 필터링의 정확도를 향상시켰다[9]. 본 연구에서는 문서 분류를 위한 대표적인 교사학습 알고리즘인 베이지안 학습 기법을 통하여 이메일 문서 분류가 이루어진다[3,10].

$C$ 를 (8)과 같이 전체 카테고리 집합이라고 하고, $C_0$는 분류가 불가능한 경우이다. 이메일 문서들의 전체 집합을 $D$ 로 한다면 (9)와 같이 정의할 수 있다.

$$Category\ Set\ C = \{c_1, c_2, \cdots, c_k\},\quad C_0 = unknown\ \ category \quad (8)$$

$$E-Mail\ \ Set\ \ D = \{d_1, d_2, \cdots, d_n\} \quad (9)$$

나이브 베이지안 분류법에서 한문서 $d_i$의 각 카테고리 $c_j$에 대한 조건부 확률을 식(10)과 같이 구해준다.

$$Re(d_i) = \{p(d_i|\, c_1), p(d_i|\, c_2), p(d_i|\, c_3), \cdots, p(d_i|\, c_k)\} \quad (10)$$

대부분 시스템에서는 분류 대상 문서에 대해서 식(11)와 같이 가장 높은 확률값을 가지는 카테고리로 분류하게 된다. 기존의 베이지안 알고리즘에서는 임계치를 사용자가 값을 일정하게 숫자로 정할 수 있게 하여 사용하였다.

그러나 본 연구에서는 이와 같이 이용되었던 고정 임계치 $T$ 를 식 (12)에 의해서 동적임계치$T'$로 변환하였다. 본 시스템의 성능 평가를 위해 동적임계치$T'$를 적용했을 때 향상된 정확도 결과를 보여주었다[9].

$$P_{max}(d_i) = \max\{p(d_i|\, C_t)\} \quad , \ t = 1,\ldots,k \quad (11)$$

$$C_{best}(d_i) = \left\{ \begin{array}{ll} \{c_j | P(d_i|\, c_j)\} = P_{max}(d_i), & if\ P_{max}(d_i) \geq T' \\ & where T' = 1 - \dfrac{P_{max}(d_i)}{\sum_{t=1}^{k} P(d_i|\, C_t)} \\ c_0 \ , & otherwise \end{array} \right\} \quad (12)$$

위 식의 간단한 예제를 들자면,

$$Re(d_i) = \{0.4, 0.5, 0.8, 0.6, 0.2, 0.1\}$$

$$T' = 1 - \frac{0.8}{(0.4 + 0.5 + 0.6 + 0.2 + 0.1)} = 0.692$$

한 카테고리에 해당도에 따른 각각 이메일 문서의 확률값을 보여주고 있는데, $T'$ 값인 0.692 이상만을 해당 카테고리로 제한하는 것이다. 여기에서 $T'$를 이용하여 임계치를 조정함으로써 분류 속도를 향상시키고, 오분류를 막을 수 있게 된다.

## 4. 이메일 추천 시스템의 구현 및 결과분석

### 4.1 이메일 추천시스템 (E-Mail Recommendation Agent)

제안된 시스템의 전체적인 구조는 다음 그림1 과 같이 설계되었다. 사용자 이메일 문서의 정보를 기반으로 베이지안 학습에 의한 필터링 · 분류 · 추천 과정을 보여주고 있다. 구현환경으로는 Windows 2000 Professional, 데이터베이스 컨트롤을 위해 MS Visual C++ 6.0, 주요기능을 컴포넌트화하기 위해 COM+, 기타 기능을 위해 ASP, ASP 컴포넌트를 이용하였다.

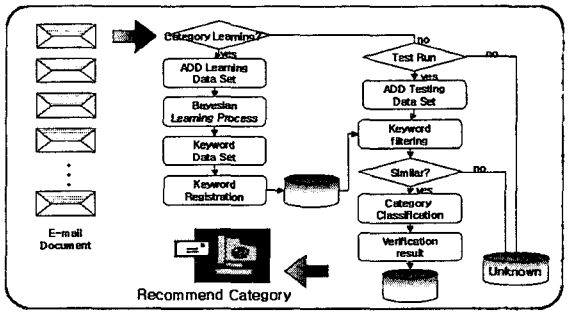


그림 1. 시스템의 전체적인 흐름도
Fig. 1 System Overview

이메일 관리를 위한 추천시스템은 다음과 같은 기능을 가지고 있는 세가지 모듈로 구성되어 있다.

첫째, 새로운 메일이 도착하면 먼저 사용자의 메일 처리 과정을 관찰하여 학습한다. 특징추출 및 규칙형성에 도움을 주는 모듈이며, 또한 사용자가 개인에 맞는 카테고리를 미리 설정할 수 있는 과정이기도 하다.

둘째, 메일 처리 관찰 과정에서 특징을 추출하여 응용된 베이지안 알고리즘을 적용하여 개인에 맞는 규칙을 형성한다.

셋째, 생성된 규칙을 기반으로 새로운 메일이 도착하면 카테고리별 분류를 한 다음 사용자에게 해당 카테고리를 우선순위별로 카테고리를 추천 해 준다.

또다른 방식으로 모듈별로 나타내 보면 시스템 구조는 그림2과 같이 세가지 모듈로 구성되어 있다.

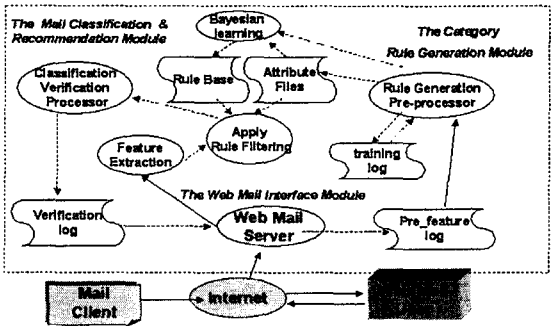


그림 2. 시스템 모듈별 설계
Fig. 2 Modular Design

WMI 모듈(The Web Mail Interface Module)에 속하는 사용자 인터페이스는 사용자 관찰과정에 이용되며, 실제 사용자가 카테고리 생성 및 저장을 할 수 있다. 이 과정을 통해 특징 추출을 한다.

CRG 모듈 (The Category Rele Generation Module)에서는 실제 메일을 카테고리별로 분류할 수 있는 규칙(Rule)을 형성한다.

마지막으로 형성된 규칙을 기반으로 MCR 모듈(The Mail Classification & Recommendation Module)에서 미리 자동분류 후 사용자에게 해당 카테고리를 추천하는 것이다.

마지막 추천 카테고리는 UI는 다음 그림 3과 같다. 즉 받은 메일에 해당되는 특정카테고리를 추천하여 주며, 그 카테고리에 적합한 확률이 99.2% 임을 보여준다.

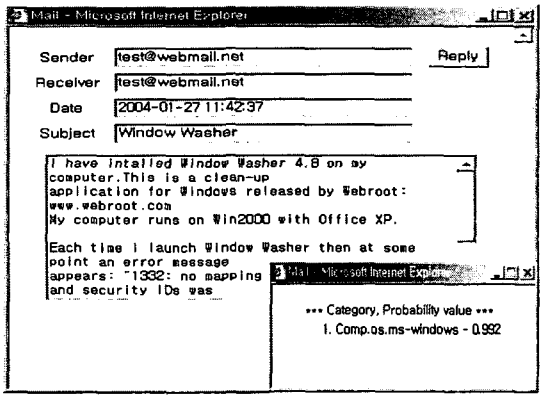


그림 3. 추천 카테고리
Fig. 3 Recommendation Category

본 응용시스템에서 주요 기능인 즉, 특징추출, 학습 및 규칙형성, 실제 분류 및 추천 기능을 COM+로 구현하여 다른 문서 분류 시스템에도 응용 될 수 있도록 확장성 및 재사용성을 고려하였다. 또한 분산 환경에서도 편리하게 이용될 수 있을 것이다. 주요 인터페이스와 메소드는 다음 그림 4와 같다.

```
IRuleFilter
{
HRESULT SetDBOpen(BSTR bstrDBConnect,BSTR
bstrDBID,
BSTR bstrDBPW);
HRESULT MergeMail(BSTR bstrID,BSTR bstrRuleName,
BSTR bstrMailData);
HRESULT CheckFolder(BSTR bstrID, BSTR
bstrMailData,[out,retval] BSTR *pRet);
HRESULT CheckFolders(BSTR bstrID, BSTR
bstrMailData,[out,retval] BSTR *pRet);
};
SetDBOpen(DBOPEN, DBID, DBPW)
MergeMail(ID, RuleName, MailData);
CheckFolder(ID, MailData);
CheckFolders(ID, MailData);
```

그림 4. 룰 필터링 컴포넌트
Fig. 4 Rule Filtering Component (COM+)

를 필터링 컴포넌트의 주요 인터페이스 및 기능은 그림 2의 모듈과 같이 매칭하여 컴포넌트화 하였으며, 각 역할은 다음과 같다.

- SetDBOpen: MS-SQL과 연결을 위해 설정하는 메소드이다 (WMI 모듈).
- MergeMail : 메일을 규칙(Rule) 형성시 필요한 데이터로 바꾸어 규칙을 형성하여 DB에 저장하는 메소드이다 (WMI모듈, CRG모듈).
- CheckFolder : 메일이 어떤 카테고리에 속하는지 체크하는 메소드이며, 가장 큰 확률값을 가진 카테고리를 리턴한다. (MCR모듈)
- CheckFolders: 메일이 어떤 폴더에 속하는지 체크하는 메소드로 모든 폴더의 확률을 체크하여 모든 확률값을 리턴하다(MCR모듈).

### 4.2 실험 설계

#### 4.2.1 문서 빈도수를 이용한 특징추출

일반적으로 문서 분류의 성능을 높이고 분류 계산 시간을 줄이기 위하여 본 실험에서는 속성집합선택을 수행한다. [11]에서는 이것을 위한 대표적인 방법인 문서 빈도(document frequency), 정보이득량 (information gain), 카이제곱통계량($\chi^2$-statistics), 상호정보량(mutual information) 그리고 용어강도( term strength)를 기준값으로 한 방법을 소개하였고, 그것들의 성능를 비교하였다. 본 실험에서는 [11]의 실험 결과를 반영하여 시간적으로 효율성이 좋으면서 분류성능을 높이는 그 결과 앞의 세 방법이 상대적으로 효과적임을 밝혔다. 여기서는 문서빈도를 기준값으로 한 속성선택 기법을 사용 방법을 사용하였다. 앞에서 소개한 5가지 속성집합선택 방법들을 Naive Bayes 문서분류기에 적용한 결과, [11]에서와 같이 앞의 세 방법의 성능이 비슷하게 효과적인 것으로 밝혀졌다. 특정 단어의 문서빈도는 해당 단어가 출현하는 학습 문서의 수를 의미한다. 희귀한 단어는 문서 분류를 하는데 있어 정보를 거의 제공하지 못한다는 것이 이 방법의 기본적인 가정이므로, 이 방법은 문서빈도가 높은 것을 우선하여 속성으로 선택한다.

#### 4.2.2 성능 평가 측정치

문서분류시스템의 적합성은 일반적으로 재현율(recall ratio)와 정확률(precision ratio)이라는 척도에 의하여 측정된다. 재현율은 문서분류시스템에 들어있는 적합한 문헌 중에서 분류 시스템에 의하여 분류된 적합한 문헌의 비율이고, 정확률은 분류된 전체 문헌 중에서 적합한 문헌의 비율이다 [12].

- $a$ : 해당 카테고리에 정확하게 분류된 문서의 수 (true positive)
- $b$ : 해당 카테고리에 틀리게 분류된 문서의 수 (false positive)
- $c$ : 해당 카테고리에 속하지만 이 카테고리로 분류되지 않은 문서의 수 (false negative)
- $d$ : 해당 카테고리에 속하지 않고, 이 카테고리로 분류되지 않은 문서의 수 ( true negative)

즉 $a+c$는 해당카테고리에 속하는 모든 문서의 수이고, $a+b$는 문서분류함수에 의해 해당 카테고리에 실제로 분류된 문서의 수이다. 이를 통해서 precision과 recall를 다음과 같이 정의할 수 있다.

$$precision = \frac{a}{a+b}$$

$$recall = \frac{a}{a+c}$$

이 두 측정치는 모두 문서분류시스템을 평가하는데 있어 매우 중요하기 때문에 이 둘의 중요성을 모두 반영한 측정치가 필요한데, 생각할 수 있는데 이를 위해 $F_1$ 측정치를 사용한다.

다음은 $F_1$ 측정치의 표현식이다.

$$F_1 = \frac{2 * recall * precision}{recall + precision}$$

위에서 설명한 recall과 precision, $F_1$ 측정치는 각 카테고리 성능을 개별적으로 평가하는것이다. 모든 카테고리에 대한 평균적인 성능을 평가하기 위해 여기서는 macro-averaging 방법을 이용한다. 이 방식에서는 각 카테고리 별로 recall, precision, $F_1$ 측정치 등을 계산하고 이들의 평균을 계산하여 전체적인 문서분류시스템의 성능을 평가한다.

#### 4.2.3 실험 결과의 분석

실제 실험은 많은 양의 데이터를 실험해야 하기 때문에 카테고리별 메일을 뉴스그룹에 있는 영문 메일 문서 데이터를 각 카테고리별로 수집한 후 실험하였다.

먼저 학습을 위한 실험단계에서는 데이터메일중에서 불확실성 샘플링 기법을 이용하여 백통정도의 메일을 추출한 후 먼저 규칙 형성을 하였다.

표1은 각 카테고리별 메일 총 수와 각각 재현율, 정확율을 이용하여 계산한 $F_1$ 값을 계산한 후 전체 카테고리를 비교 분석하기 위하여 macro average값을 나타내준다.

표 1에서 볼 수 있듯이 F1(BA)는 기존의 베이지안 알고리즘을 이용한 결과값이고, 기존의 알고리즘을 그대로 적용

표 1 전처리 알고리즘별 분류도 측정치
Table 1 F1 measure and Macro-averaging of each pre-processing algorithm

| 번호 | 카테고리 항목 | 데이터수 | F1 (BA) | F1(SA) | F1(WV) | F1(DT) |
|---|---|---|---|---|---|---|
| 1 | Sale | 3434 | 0.860 | 0.869 | 0.880 | 0.910 |
| 2 | Autos | 2712 | 0.920 | 0.910 | 0.910 | 0.920 |
| 3 | Sports | 1124 | 0.890 | 0.870 | 0.880 | 0.870 |
| 4 | Electronics | 1354 | 0.940 | 0.950 | 0.950 | 0.990 |
| 5 | Politics | 7971 | 0.840 | 0.840 | 0.838 | 0.850 |
| 6 | Computer | 573 | 0.930 | 0.950 | 0.940 | 0.970 |
| 7 | Graphics | 2579 | 0.910 | 0.910 | 0.920 | 0.860 |
| 8 | Hardward | 1578 | 0.830 | 0.830 | 0.840 | 0.850 |
| 9 | Space | 1694 | 0.850 | 0.870 | 0.870 | 0.940 |
| 10 | Talk | 2915 | 0.890 | 0.870 | 0.870 | 0.850 |
| 11 | Language | 1180 | 0.880 | 0.910 | 0.910 | 0.910 |
| 12 | Spam | 3260 | 0.890 | 0.889 | 0.920 | 0.920 |
| | Macro Average | | 0.886 | 0.889 | 0.894 | 0.903 |

하면서 학습 집합만 샘플링 알고리즘으로 적용해서 선택했을때 값은 F1(SA)이며 기존의 F1(BA)의 macro average값보다 0.003 정도의 약간의 향상을 보여주었다.

이 단계에서 F1(WV)는 제목에 가중치를 부여했을때의 결과는 0.894로 기존의 F1(BA)의 macro average값보다 0.008(0.8%)의 개선됨을 보여주었다. 마지막으로 F1(DT)는 동적임계치(Dynamic Threshold)를 적용한 후 결과값이다. 최종적으로 처음 기존의 베이지안 알고리즘을 이용했을때보다 0.017(1.7%)의 향상도를 보여주었다.

본 연구의 전단계에서 우리는 기존의 전처리 알고리즘을 이메일 카테고리별 분류에 적용하여 그 특성을 파악한 후 각각 가중치 부여 기법과 동적 임계치를 적용한 기법을 제안하여 독립적으로 연구진행을 해왔다[1,9]. 그러나, 이방법들을 단계별로 연계하여 적용한다면 좀 더 나은 결과를 볼 수 있다는 점을 착안하여 3단계를 하나의 전처리 알고리즘의 개념으로 설정하고 실험 및 결과 분석하였다.

## 5. 결 론

현재 이메일을 통해 많은 양의 정보들이 오가고 있고, 사용자들은 사용하기 편리한 이메일 인터페이스를 요구하게 될 것이다. 메일 문서를 일반적인 문서분류시스템에 적용하여 사용자가 메일을 카테고리별 분류하는데 정확한 카테고리를 추천하여 준다면 훨씬 편리하게 메일시스템을 관리할 수 있을 것 이다. 본연구에서 자동 분류 방식보다 추천방식으로 설계한 이유는 메일문서는 다른 일반문서보다 개인적 성향이 매우 강하므로 마지막 단계에서는 사용자의 개입이 가능하도록 한 것이다.

이메일 문서의 정확한 카테고리별 분류를 위해 본 연구에서는 기존의 방법 샘플링 알고리즘과 본연구에서 제안한 가중치와 임계치를 개선한 방법을 단계별로 연계 적용하는 방식인 3단계 전처리 알고리즘을 제안하였다. 갈수록 늘어나는 메일을 관리할 때 3단계 전처리 알고리즘을 이용한 이메일 추천 시스템은 매우 유용하게 활용될 것이다.

향후연구로는 제안한 전처리 알고리즘을 문서 분류 추정 알고리즘 Naive Bayes 에 적용하였지만 이를 최근 들어 문서분류시스템 구축하는데 자주 사용되고 있는 SVM 알고리즘 등에 적용하여 볼 것이다. 또한 주요 모듈 부분들이 컴포넌트화 되어 있으므로 일반 텍스트 문서에 적용했을 때의 결과와 비교 분석하기 위해 텍스트 분류의 표준 데이터인 Reuter-21578 문서집합을 이용해 보고자 한다.

**감사의 글**

이 논문은 2004년도 두뇌한국21사업에 의하여 지원되었음.

## 참 고 문 헌

[1] Ok-Ran Jeong, Dong-Sub Cho, "A Personalized Recommendation Agent System for E-Mail Document Classification" , Computational Science and Its Applications-ICCSA 2004, LNCS3045, Springer Verlag, Vol 3, pp.558-565, 2004

[2] Ian H. written and Eibe Frank, "Data Mining," Morgan Kaufmann Publishers, Inc., 2000.

[3] Pedro Domingos and Michael Pazzani. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier," In Proceedings of the 13thInternational Conference on Machine Learning, pp105-112, 1996

[4] F.Sebastiani, "Machine Learning in Automated Text Categorization," Technical Report IEI-B4-31-19

[5] David D. Lewis and William A.Gale. A Sequential Algorithm for Training Text Classifiers. In Proceedings of the 17thAnnual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 3-12, 1994.

[6] David D. Lewis and Jason Catlett. Heterogeneous

Uncertainty Sampling for Supervised Learning. In Proceedings of the 11th International Conference on Machine Learning, pages 148-156, 1994

[7] M. Trensh, N. Palmer, and A. Luniewski. Type Classification of Semi-structured Documents. In Proceedings of the 21st ACM SIGMOD International Conference on Management of Data, 1995.

[8] 강영순, 이용배,김태현,조숙현, 맹성현, "전자우편문서의 효율적인 분류을 위한 전처리", 제 29회 춘계학술발표회, 한국정보과학회, 제29권 제1호 pp.493-495, 2002.

[9] 정옥란, 조동섭, "개인화된 분류를 위한 웹 메일 필터링 에이전트", 정보처리학회논문지B, 제 10-B권 제7호, pp.853-862, 2003.

[10] Tom Mitchell, MaGraw Hill, "Machine Learning", McGRAW-HILL International Edition, 1997.

[11] M. Trensh, N.Palmer, and A. Luniewski, "Type Classication of Semi-structured Documents," In Proceedings of the 21st ACM SIGMOD International Conference on Management of Data, 1995.

[12] Yiming Yang, Jan O. Perdersen, "A Comparative Study on Feature Selection in Text Cateforization", Proc. of ICML97, pp.412-420, 1997.

## 저 자 소 개

**정 옥 란 (鄭 玉 蘭)**

1993년 전북대학교 전자계산학과 졸업.
1998년 동대학교 대학원 전산정보학과 졸업(이학석사)
1999년 9월~현재 이화여자대학교 컴퓨터학과 박사과정
Tel : 02-3277-2309
Fax : 02-3277-2306
E-mail : orchung@ewhain.net
orchidjeong@acm.org

**조 동 섭 (趙 東 燮)**

1979년 서울대학교 전기공학과 졸업.
1981년 동대학교 대학원 전기공학과 졸업(공학석사).
1986년 서울대학교 대학원 컴퓨터공학과 (공학박사)
1985년~현재 이화여자대학교 컴퓨터학과 정교수
Tel : 02-3277-2309
Fax : 02-3277-2306
E-mail : dscho@ewha.ac.kr