韓國컴퓨터情報學會 論文誌
第16卷 第2號, 2011. 2.

2011-16-2-1-5

효과적인 이메일 분류를 위한 빈발 항목집합 기반 최적 이메일 폴더 추천 기법

문 종 필*, 이 원 석**, 장 중 혁***

A proper folder recommendation technique using frequent itemsets for efficient e-mail classification

Jong Pil Moon*, Won Suk Lee**, Joong Hyuk Chang***

요 약

이메일이 중요한 정보 전달과 의사소통의 수단으로 널리 활용된 이래 사람들은 이메일을 내용에 따라 적절하게 분류하는 작업에 많은 노력을 기울려 왔다. 이메일은 문서의 길이나 문체가 다양하며 사용되는 단어들이 비정규적이다. 또한 이메일 분류 기준은 일반적으로 해당 이메일 사용자의 주관에 따라 정의된다. 따라서 기존의 일반적인 문서 분류 기법으로는 이메일을 효율적으로 분류하는데 어려움이 있다. 상업용 이메일 프로그램에서 제공되는 분류 기능은 메일 클라이언트에서 지원하는 텍스트 필터링을 이용한다. 한편 이메일의 자동 분류에 관한 연구는 확률 기반의 나이브 베이지안 기법을 응용하여 정확도를 높일 수 있는 연구가 주로 진행되어 왔으며, 대부분 영문 이메일에 대한 연구이다. 본 논문에서는 빈발 패턴 마이닝 기법을 적용하여 한글 이메일에 대한 개인 맞춤형 폴더 추천 기법을 제시한다. 이메일의 맞춤형 폴더 추천 기법은 이메일에 대한 전처리 과정과 빈발 항목집합을 이용한 메일 폴더의 프로파일 생성 과정으로 구성된다. 생성된 프로파일은 분류 대상이 되는 각 메일이 개인별 맞춤형 기준에 따라 가장 적합한 이메일 폴더로 효과적으로 분류되는데 활용된다. 또한 제안된 기법을 적용한 이메일 분류 시스템을 구현한다.

▸ Keyword : 이메일 분류, 개인 맞춤형 추천, 개인 맞춤형 분류, 빈발 항목집합, 문서 분류

Abstract

Since an e-mail has been an important mean of communication and information sharing, there have been much effort to classify e-mails efficiently by their contents. An e-mail has various forms in length and style, and words used in an e-mail are usually irregular. In addition, the criteria of an e-mail classification are subjective. As a result, it is quite difficult for the conventional text classification technique to be adapted to an e-mail classification efficiently. An e-mail classification

• 제1저자 : 문종필 교신저자 : 장중혁
• 투고일 : 2010. 09. 30, 심사일 : 2010. 11. 02, 게재확정일 : 2010. 12. 13.
* KT 이노츠 연구개발본부(KT Innotz)
** 연세대학교 컴퓨터과학과 정교수(Dept. of Computer Science, Yonsei University)
*** 대구대학교 컴퓨터IT공학부 교수(Dept. of Computer & Information Technology, Daegu University)
※이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국과학재단의 국가지정연구실사업(No.2010-0008007) 및 특정기초연구(No.2008-0052335)지원으로 수행된 연구임

technique in a commercial e-mail program uses a simple text filtering technique in an e-mail client. In the previous studies on automatic classification of an e-mail, the Naive Bayesian technique based on the probability has been used to improve the classification accuracy, and most of them are on an e-mail in English. This paper proposes the personalized recommendation technique of an email in Korean using a data mining technique of frequent patterns. The proposed technique consists of two phases such as the pre-processing of e-mails in an e-mail folder and the generating a profile for the e-mail folder. The generated profile is used for an e-mail to be classified into the most appropriate e-mail folder by the subjective criteria. The e-mail classification system is also implemented, which adapts the proposed technique.

▸ Keyword : E-mail classification, Customized recommendation, Customized classification, Frequent itemsets, Document classification

Ⅰ. 서 론

정보화 시대에 있어서 이메일(e-mail)은 중요한 정보 전달의 수단이다. 이메일은 광고, 청구서, 회사 업무 등은 물론 개인의 사적인 생활 등과 관련하여 생활의 모든 부분을 표현하고 있으며, 다양한 분류 방법이 존재한다. 인터넷의 대중화와 정보화로 인하여 개인이 수신하는 이메일의 수가 지속적으로 증가하고 있고, 개인은 다양한 이메일을 목적에 맞게 분류하여 메일 폴더에 저장하고 있다. 또한 이메일 사용자들은 하나의 메일 계정만을 사용하지 않고 목적 별로 여러 개의 메일 계정을 사용하고 있다. 이러한 환경에서 메일 분류 작업을 수행하지 않는다면 방대한 수의 이메일 중에서 목적에 맞는 이메일을 찾는 일은 많은 시간이 소요된다. 한편 이메일을 용도와 목적에 따라 적합한 폴더로 분류하는 작업은 객관적인 분류 기준보다는 개인의 주관적인 기준에 따라 진행되는 특성을 갖는다.

메일 분류를 위한 제품화된 솔루션은 메일 클라이언트에서 제공하는 단순한 메일 규칙들이 주를 이룬다. 이러한 규칙을 제공하는 대표적인 메일 클라이언트로서는 아웃룩과 아웃룩 익스프레스, 썬더 메일 등이 있다. 이들은 제목이나 본문에 특정 단어나 문장을 포함하면 특정 메일 폴더로 자동으로 이동시키는 규칙과 보낸 사람을 기준으로 메일을 분류시키는 기능을 주요 기능으로 제공하고 있다.

메일함 관리 기준은 사용자의 필요에 따라 다양한 방법이 적용될 수 있다. 특히 이메일의 수신자/발신자 정보 등과 같이 정형화된 정보를 활용하여 분류 작업을 수행하기도 한다. 그러나 동일한 사람이 업무에 관련된 메일과 취미에 관련된 메일을 보내는 예시와 "카드"라는 단어를 포함한 메일이 청구서 폴더와 축하 메일 폴더 중 어떤 폴더에 분류될 것인가와 같은 예시를 고려해 본다면, 정형화된 정보를 기반하는 메일 분류 규칙만으로는 개인이 의도하는 바를 만족시키기에 한계가 있음을 알 수 있다. 따라서, 메일의 내용 등을 고려한 사용자의 주관적인 분류 기준을 만족시킬 수 있는 정확도가 높은 메일 분류 기법을 필요로 하고, 이 기법이 제품화된 솔루션에 포함된다면 사용자들의 분류에 대한 시간과 노력을 줄일 수 있다.

이메일은 약어, 속어 등의 사용이 빈발하고, 문체가 자유로우며 문서의 길이가 비정규적이라는 특징을 가지고 있다[1]. 따라서, 메일을 분류하는 작업은 정규화된 문서를 분류하는 작업보다 정확도가 낮다. 이러한 단점 때문에 불용어 제거기의 성능이 중요하게 부상되고 있다. 또한, 영어권에서의 메일 분류에 관한 연구는 활발하게 진행되고 있지만, 한글 이메일에 관한 연구는 미비한 상태이다[2]. 따라서, 불용어 제거기를 활용한 한글 메일 분류 기법의 필요성이 증가하고 있다.

본 논문은 단어의 조합을 이용하는 데이터 마이닝 기법을 사용하여 메일 분류의 정확도를 향상시킨다. 폴더에 속해있는 메일들을 하나의 트랜잭션으로 보고, 메일에 있는 단어들을 하나의 항목으로 본다면, 항목들의 집합을 정의할 수 있다. 이러한 항목집합(itemsets) 중에서 빈발 항목집합을 구하고 이를 활용하여 해당 메일이 속하게 될 폴더를 추천한다. 또한, 사용자의 실제 메일 데이터를 사용함으로써 비정형적인 한글 메일에 대한 분류 기법의 정확도를 측정한다. 메일 데이터를 훈련 집합과 시험 집합으로 나누고, 훈련 집합을 기반으로 빈발패턴을 적용한 메일 폴더의 프로파일을 생성하고, 시험 집합에 각 폴더의 프로파일을 적용함으로써 가장 좋은 결과를 보여주는 폴더를 사용자에게 추천한다.

본 논문의 구성은 2장에서 관련 연구에 대하여 살펴보고, 3장에서 폴더의 프로파일 생성을 위한 알고리즘을 설명한다. 4장에서는 맞춤형 폴더 추천 기법을 구현하는 방법에 대해서 기술하고, 5장에서 실험 결과에 대한 설명과 나이브 베이지안 기법과의 비교 결과를 기술한다. 끝으로 6장에서 결론을 기술한다.

Ⅱ. 관련 연구

새로운 이메일에 대해서 이메일이 속할 폴더를 추천해주는 기법은 나이브 베이지안 학습에 기반한 기법[1,3-6]과 중심점 기반 분류기를 이용한 기법[2], 그리고 추천 시스템을 이용한 기법[7]이 있다.

나이브 베이지안 분류기는 문서 내 모든 속성들이 주어진 클래스 내에서 서로 독립이라는 가정 하에 통계적인 방법으로 문서를 분류하는 기법이다[2]. 나이브 베이지안은 두 가지 모델이 존재한다. 문서에서 단어의 존재 여부를 기반으로 이진 속성 벡터에 의해서 문서를 표현하는 다변량 베르누이 이벤트 모델과 문서가 발생한 빈도에 의해서 가중치를 부여하고 문서를 표현하는 다항식 이벤트 모델이 존재한다[2]. 문서의 단어가 크거나 단어들이 적절하게 선택되었을 경우에, 다항식 이벤트 모델이 다변량 베르누이 모델보다 우수한 성능을 보인다는 연구 결과가 있다[8].

다변량 베르누이 모델을 기초로 이메일을 분류하는 연구는 메일의 각각의 폴더에 대한 확률 값을 구하고, 구해진 확률 값 중 가장 높은 확률을 가진 폴더에 문서를 분류하는 기법이다. 이를 응용해서 고정된 임계치를 동적으로 개선하여 필터링의 적합도를 향상시킨 연구[3]는 메일 폴더의 중요 단어에 가중치를 부여하는 전처리를 한 후, 나이브 베이지안 알고리즘을 응용해서 실험 했을 때 88.6%의 정확률을 나타냈으며, 동적인 임계치를 적용했을 때 89.1%의 정확률을 나타냈다.

나이브 베이지안의 다항식 이벤트 모델을 기초로 한 연구는 문서에서의 단어의 발생 빈도에 가중치를 주는 방법으로 기존의 나이브 베이지안 기법을 수정하였으며, 이를 이용해서 카드 회사에 들어오는 고객의 전자 메일을 분류한 연구[2]가 존재한다. 이 연구에서는 총 메일 폴더의 수가 3개이고 메일수가 658개일 때 정확도 0.88을 나타냈고, 메일 폴더의 수가 6개이고 메일의 수가 1210개일 때 0.81의 정확도를 나타냈다.

메일이 아닌 다른 문서에서의 분류 연구를 살펴보면, 빈발 단어 집합을 나이브 베이지안에 적용하여 정확도를 높인 연구[4,6]가 존재한다. 이 연구에서는 단어를 1-크기의 빈발단어 집합으로 보고, k-크기의 빈발단어 집합을 나이브 베이지안에 적용함으로써 정확도를 높이고자 했다[4]. 그리고 한국어 정보검색 시스템의 성능평가용 데이터 집합인 KTset95 문서 중 2400개의 문서를 훈련 집합으로 사용한 연구[9]에서는 연관단어를 기반으로 한 지식베이스 기반의 베이지안 분류법을 연구하였으며, 그 결과 역문헌 빈도에 가중치를 준 베이지안 분류 방법보다는 2.18%, 단순 베이지안 방법보다는 4.11% 성능이 향상되었다.

추천 시스템은 일반 상품 관련 추천에서부터 영화, 음악, 뉴스, 웹 페이지 관련 추천까지 여러 분야에 걸쳐 진행되고 있는 연구이며, 주로 내용 기반 필터링과 협업 필터링으로 나누어진다[7]. 내용 기반 필터링에서는 폴더 별로 프로파일을 생성하고, 이를 벡터 공간 모델(Vector Space Model) 기반으로 코사인 유사계수를 사용하여 유사성을 판단한다. 이 연구에서는 사용자의 사전을 구축하고, 사용자의 지식을 이용한 단어 간의 유사성을 반영하여서, 총 318개의 메일에 대해서 실험을 하였다. 그 결과, 시험 집합의 메일에 대하여 한 개의 폴더를 추천하였을 경우 정확도 0.6을 나타냈고, 세 개의 폴더를 추천하였을 경우 정확도 0.8을 나타냈다.

중심점 기반 분류기는 이메일을 벡터 공간 모델을 이용하여 표현한다. 이 기법[2]에서는 이메일을 단어-빈도(TF)와 역문서빈도(IDF)를 이용하여 벡터로 표현하였다. 그리고 메일 폴더 별로 중심점 벡터를 계산하고, 테스트 메일에 대하여 TF-IDF 벡터와의 유사성을 코사인 함수를 이용하여 판단한다. 이 연구 방법은 총 메일 폴더의 수가 3개이고 메일수가 658개일 때 정확도 0.88을 나타냈고, 메일 폴더의 수가 6개이고 메일의 수가 1210개일 때 0.79의 정확도를 나타냈다.

메일 분류의 정확도를 높이는 알고리즘과 병행하여 전처리의 방법에 대한 연구[3]도 진행되고 있다. 이 연구에서는 3가지의 전처리 방법에 대해서 제안하고 있다. 첫 번째 방법은 평균 절대편차 값을 이용하는 불확실성 샘플링 알고리즘을 적용하고, 두 번째 방법으로 제목 부분에 본문보다 가중치를 더 줄 것을 제안했다. 세 번째 방법은 나이브 베이지안을 적용함에 있어서 고정 임계치를 동적 임계치로 수정하는 방법이다. 이러한 전처리 방법을 뉴스그룹에 있는 영문메일을 대상으로 적용하고 실험한 결과는 불확실성 샘플링 알고리즘을 통해 정확도가 0.3% 향상되었으며, 제목에 가중치를 부여했을 때 0.8% 향상되었고 동적 임계치를 적용했을 때 1.7% 향상되었다.

데이터 분류 과정에서 분류 정확도를 높이기 위한 방법으로 연관규칙 또는 빈발 항목집합에 기반한 분류에 대한 연구[10,11]들도 수행되었다. 해당 연구들에서는 데이터 분류 과정에서 연관규칙 탐색에 의해 얻어진 연관규칙들 중에서 분류 작업한 유용한 규칙들을 찾고 이를 활용하여 분류 작업을 수행함으로써 훈련용 데이터 집합에 대한 단순 분석을 통해 규칙을 찾고 이를 분류 작업에 적용하는 기존의 방법들에 비해 분류 정확도를 향상 시켰다. 본 논문에서 제안하는 이메일 분류 기법은 이러한 기법을 이메일 분류 과정에 적용함으로써 개인의 주관적인 기준을 효율적으로 반영하여 각 메일에 적합한 폴더를 추천하고 이를 통해 개인 맞춤형 이메일 관리 시스템을 구현할 수 있도록 하는데 목적을 둔다.

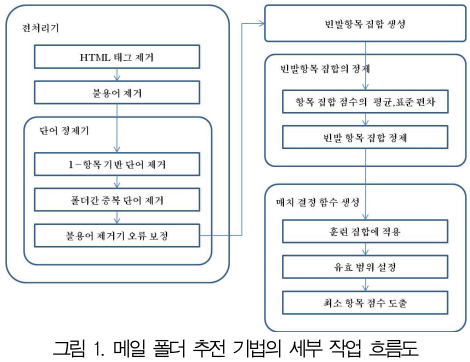


그림 1. 메일 폴더 추천 기법의 세부 작업 흐름도
Fig. 1. Detailed processes of an e-mail folder recommendation system

III. 메일 폴더 추천 기법

메일 폴더 추천 기법은 각 폴더에 존재하는 이메일들에 대해서 데이터 마이닝 기법을 적용하여 폴더를 대표하는 프로파일을 생성하고, 생성된 프로파일들을 신규 이메일에 적용하여서 가장 적합한 폴더를 추천하는 기법이다. 본 논문에서는 빈발 패턴을 사용하여 메일의 단어들을 대상으로 빈발 항목집합을 생성하고 생성된 빈발 항목집합을 기반으로 하는 프로파일을 생성 방법을 제안한다.

그림 1은 메일 폴더 추천을 위해서 본 논문에서 사용할 알고리즘이다. 메일에 포함된 단어들을 유효한 단어들로 정제하기 위해서 전처리기를 통한 전처리를 하고, 그 단어들로부터 빈발 항목집합을 생성한다. 그리고 생성된 빈발 항목집합들을 유효한 빈발 항목집합으로 정제를 하고, 정제된 빈발 항목집합을 바탕으로 매치 결정 함수를 생성한다. 생성된 매치 결정 함수는 메일 폴더를 대표하는 프로파일이며, 새로운 메일에 대하여 이 함수를 적용함으로써 추천에 가장 적합한 폴더를 제안하게 된다.

1. 전처리 단계

일반 문서에 대한 자동 분류 연구는 정제가 잘 되어 있고 정형화 되어 있는 문서를 사용한다. 그러나, 이메일은 비정형적인 특성이 많고, HTML 문서이며, 속어 및 약어의 사용이 많다. 그리고 동사에서는 어미를 제외한 어간만이 유효한 단어이며, 관사와 조사는 분류 작업에서 유효하지 않다. 빈발 항목집합을 찾을 때, d개의 항목들을 포함하는 데이터 집합에서 추출 가능한 연관규칙(association rule)들의 개수 R은 다음과 같다[12].

$$R = 3^d - 2^{d+1} + 1$$

많은 수의 단어에 대하여 마이닝을 한다면 시간적 제약과 메모리 사용의 제약이 따른다. 따라서, 빈발 항목집합을 추출하기 전에 효율성과 정확도의 향상을 위하여 유효하지 않은 단어를 제거 할 필요가 있다.

전처리 과정에서는 정규식을 이용하여 메일 본문에서 HTML 태그를 제거하고, 형태소 분석기를 이용하여 조사와 어미 등의 불용어를 제거한다. 이어서 단어의 수를 다음 3가지 단계로 정제한다.

첫 번째 단계로, 각 폴더에서 지지도가 특정 값 미만인 단어들을 제거함으로써 빈발 단어들을 얻으며, 이때 데이터 마이닝 기법 중의 하나인 빈발 항목집합[12] 탐색 기법을 이용한다. 데이터베이스 D에서 I={i1,i2,…,id}를 모든 항목들의 집합이라고 하고, T={T1,T2,…,Tn}을 모든 트랜잭션의 집합이라고 하자. 각 트랜잭션 ti는 I로부터 선택된 항목들의 부분집합을 포함하고 있다. 이 때 0 또는 다수 항목의 집합을 빈발 항목집합이라고 부르고, 한 빈발 항목집합에 k개의 항목들을 포함하면, k-항목집합이라고 부른다. |D|를 D에 포함된 트랜잭션의 개수라고 하고, |X|를 D에서 X를 포함하는 트랜잭션의 개수라고 할 때, 지지도 SUP(X)란 |X|를 |D|로 나눈 값을 의미한다.

$$SUP(x) = \frac{|X|}{|D|}$$

최소 지지도(s)란 사용자가 지정하는 지지도의 임계값을 말하고, 빈발 항목집합은 I에 속하는 항목집합 X 중 최소지지도 s 이상의 지지도를 갖는 모든 항목집합을 지칭한다.

전처리의 첫 번째 단계로서 지지도 기반으로 단어를 제거하는 작업은 Apriori 원리를 응용한 단계이다. Apriori 원리는 "만약 한 항목집합이 빈발하면, 그것의 모든 부분집합들 역시 빈발해야한다"[12]이다. 이를 바꾸어 말하면, 항목집합 J에 대해서 J가 빈발하지 않다면, J를 포함하는 모든 집합(superset)은 빈발하지 않다고 추론할 수 있다. 이러한 추론에 따라서 전처리 시에 사용할 전처리 최소 지지도(PSmin)를 결정할 수 있는데, 이는 빈발 항목집합을 구할 때 사용될 최소 지지도보다 작은 값으로 설정해야 한다. PSmin 이하의 지지도를 가지는 1-항목집합은 빈발 항목집합 추출 과정에서 의미가 없는 단어들이 되고, 전처리 시 제거해야 할 단어들이 된다. 즉, 폴더 F가 가지고 있는 메일 수를 Mf라고 하고, 단어 w가 포함되어 있는 메일의 수가 Nw 일 때, 제거해야 할 단어들의 집합 I는 다음과 같다.

$$I = \{w \in F \mid \frac{N_w}{M_f} < PS_{\min}\}$$

두 번째 단계로, 다수의 메일 폴더에서 출현하는 단어들을

제거한다. 첫 번째 단계를 통해 정제된 단어들 중에서 다수의 메일 폴더에서 나타나는 단어는 프로파일의 오류율을 높이는 원인이 될 수 있다. 여기서 오류율이란 잘못된 폴더로 분류될 확률을 말한다.

세 번째 단계로, 단어의 길이가 1인 단어들을 제거한다. 이 단계는 본 논문에서 사용한 불용어 제거기의 에러를 보정하는 단계이다. 불용어 제거기를 사용한 후의 정제된 단어들의 집합에는 제거가 되지 않은 불용어들이 존재하였다. "할 수 없다"라는 문장에서는 "수"가 제거가 안 되었고, 이름에 공백을 넣은 "홍 길 동" 과 같은 문장에서는 각각의 "홍","길","동"을 제거하지 못하였다. 단어의 길이가 1인 단어는 실제로 의미를 가지고 있는 단어보다 그렇지 않은 단어가 많았고, 그러한 단어는 전처리 시에 제거 대상이 되어야 한다.

2. 빈발 항목집합 생성 및 정제

다음 단계로서, 전처리 과정에서 정제된 단어 집합들에 대하여 Apriori 알고리즘을 적용해서 빈발 항목집합을 추출해야 한다. 빈발 항목집합을 추출하기 위해서는 메일 폴더 별로 최소 지지도를 설정해야 한다. 모든 폴더에 동일한 최소 지지도를 설정한다면, 특정 폴더는 추출된 빈발 항목집합의 개수가 많을 것이고, 특정 폴더는 빈발 항목집합이 없을 수도 있다. 따라서, 폴더마다 최소 지지도를 다르게 설정해야 하고, 결과로 추출된 빈발 항목집합을 정제하는 정규화를 통하여 이를 보정해야 한다.

빈발 항목집합을 정제하기 위해서는 빈발 항목집합 점수와 정제된 빈발 항목집합에 대한 정의가 필요하며 각각 [정의 1] 및 [정의 2]에서와 같이 정의된다. 빈발항목 집합의 관심도나 중요도를 결정하기 위한 방법은 다양하게 고려될 수 있다. 예를 들어 항목집합의 길이에 따른 관심도를 정의하는데 있어서 길이와 관심도의 관계를 단순 관계로 간주하거나 또는 지수 관계로 간주하는 등의 방법이 고려될 수 있다. 본 논문에서는 항목집합의 지지도 및 단순 길이를 고려한 관심도를 정의한다. 즉, 일반적으로 길이가 긴 항목집합일수록 관심도가 큰 것으로 간주될 수 있으며 이를 고려하여 하나의 항목집합에 대한 항목집합 점수를 정의한다.

[정의 1. 항목집합 점수 (Itemset score: IS)] 빈발 항목집합 i의 지지도가 SUP(i)이고, 항목의 길이가 Li일 때 i의 빈발 항목집합 점수 IS(i)는 다음과 같이 정의된다.

$$IS(i) = SUP(i) \times L_i \quad \blacksquare$$

[정의 2. 정제된 빈발 항목집합 (Refined frequent itemset: RFI)] 메일 폴더 f에 속하는 모든 빈발 항목집합들을 I={i1,i2,…,in}라고 하고, I에 있는 빈발 항목집합의 항목집합 점수의 평균을 Ef, 표준 편차를 σf 라고 할 때 해당 메일 폴더에 대한 정제된 빈발 항목집합 RFIf는 다음과 같이 정의된다.

$$RFI_f = \{i \in I \mid (IS(i) \geq (E_f - \sigma_f)) \wedge (IS(i) \leq (E_f + \sigma_f))\}$$

길이가 길면서 지지도가 높은 항목집합과 같이 IS가 지나치게 높은 항목집합은 해당 폴더에 속하는 메일의 보편적인 특성을 나타내는 것이 아니라 작은 수의 예외적인 특징이 지나치게 높게 고려될 수 있는 단점이 있다. 즉, 데이터 마이닝을 활용하여 데이터 집합의 특성을 나타내는 프로파일을 생성하는데 있어서는 예외적으로 두드러지는 특징보다는 보편적으로 적용될 수 있는 특징을 고려하는 것이 보다 적합하다. 본 논문에서는 이러한 특이 상황을 배제하기 위해서 추출된 빈발 항목집합들 중에서 항목집합들의 IS의 평균값을 중심으로 표준편차의 범위 내에 있는 IS를 갖는 빈발 항목집합들을 RFI로 정의한다. 따라서 길이가 짧고 낮은 지지도를 가지고 있는 빈발 항목집합과 길이가 길면서 높은 지지도를 가지고 있는 예외적인 빈발 항목집합들을 제외시킨다. 이때, IS의 허용범위를 보다 크게 설정하는 경우 길이가 길거나 상대적으로 큰 지지도를 갖는 항목집합 등과 같이 보다 다양한 빈발 항목집합을 RFI로 포함할 수 있다. 즉, RFI를 얻기 위한 IS의 허용범위를 다양화하여 각 폴더의 프로파일 특성을 조절할 수 있다. RFI는 메일 폴더의 프로파일에서 일치 여부를 판단할 수 있는 기반이 되는 빈발 항목집합이 된다. 즉, 메일 폴더를 대표하는 빈발 항목집합이다.

3. 매치 결정 함수의 생성

RFI를 이용하면 폴더의 프로파일인 매치 결정 함수를 정의할 수 있다. 이 함수를 정의하기 위해서는, 메일 항목 점수와 최소 메일 항목 점수의 정의가 필요하며 [정의 3]에서와 같이 정의된다.

[정의 3. 메일 항목 점수 (Mail Score for RFI)] 폴더 f의 RFI를 RFIf 라고 하고, 메일 m의 단어들이 항목집합의 집합 R ∈ RFIf, R={r1,r2,…,rn}에 있는 빈발 항목집합에 대해서 모두 일치한다면, 메일 m의 폴더 f에 대한 메일 항목 점수 MS(f, m)은 다음과 같이 정의된다.

$$MS(f, m) = \sum_{i=1}^{n} IS(r_i)$$

즉, 메일 항목 점수는 메일 m을 RFI에 적용했을 때, RFI에 속해있는 빈발 항목집합 중 일치되는 빈발 항목집합들의

IS의 합으로 정의된다. ■

그리고 메일 폴더 별로 폴더 추천의 판단 기준이 되는 MS의 최소값을 정의할 수 있는데, 그 값을 최소 메일 항목 점수라고 하고 [정의 4]에서와 같이 정의된다.

[정의 4. 최소 메일 항목 점수 (Least Mail Score for RFI: LMS)] MSi를 메일 폴더 f의 RFI에 포함된 빈발 항목집합 i의 MS 값이라고 할 때 f에 속하는 모든 m에 대해서 'MS(f,m)-MSi≥0'를 만족하는 MSi를 구할 수 있으며 이 값을 해당 메일 폴더의 최소 메일 항목 점수 LMSf 라 정의한다. 최소 메일 항목 점수는 각 폴더마다 서로 다른 값으로 구해진다. ■

다음으로 매치 결정 함수를 [정의 5]에서와 같이 정의할 수 있으며, k 개의 메일 폴더가 존재하는 상황에서 하나의 메일 m을 적합한 폴더로 분류하고자 하는 경우 매칭 결정 함수 값이 가장 큰 폴더가 해당 메일 m에 대한 추천 폴더가 된다.

[정의 5. 매치 결정 함수 (Function for Matching: FM)] 메일 m과 폴더 f가 존재할 때, 매치 결정 함수는 다음과 같이 정의된다.

$$FM(f,m) = MS(f,m) - LMS_f$$ ■

한편, LMS의 값은 빈발 항목집합을 구하는 데 사용되었던 훈련 집합의 RFI에 대한 MS를 바탕으로 설정할 수 있다. 훈련 집합에 존재하는 메일들에 대한 MS 값의 범위에서 유효 범위를 정하고, 유효 범위 내에서의 최소 MS의 값을 해당 폴더의 LMS로 설정한다.

Ⅳ. 메일 분류 시스템 프로토타입 설계 및 구현

메일 폴더 추천 기법을 활용한 메일 분류 시스템은 그림 2에서와 같은 순서로 세부 작업이 진행된다. 즉, 메일 폴더에 있는 메일들에 대해서 훈련 집합과 시험 집합으로 나눈 뒤 훈련 집합으로 빈발 항목집합을 통한 프로파일을 생성하고 시험 집합의 각 메일들로 프로파일을 테스트함으로써 해당 메일이 속할 적합한 폴더를 찾아 분류한다.

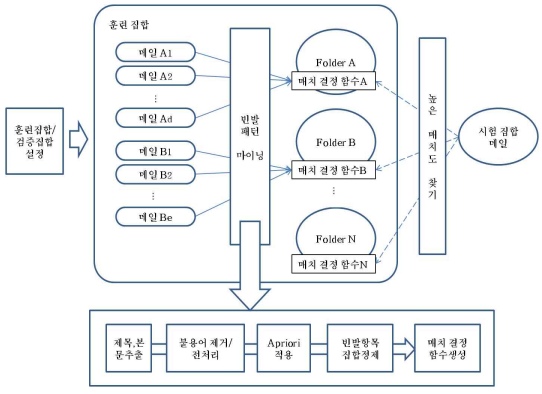


그림 2. 최적 폴더 추천을 위한 작업 흐름도
Fig. 2. Detailed processes for finding proper folder

1. 제목 및 본문 추출

이메일 본문은 메일 헤더와 본문으로 구성되어 있다. 메일 헤더에는 보내는 사람, 받는 사람, 받은 서버 정보, 제목 등 메일에 대한 메타 정보가 포함되어 있고, 메일 본문은 텍스트 본문, HTML 본문, 첨부파일로 구성되어 있다[13]. 메일 본문 추출은 본문에서 텍스트 본문 부분을 추출해야 한다. 그러나 대다수의 웹메일과 회사의 대량 메일 발송 시스템은 텍스트 본문을 넣지 않고 HTML 본문만을 메일에 삽입하여 발송하고 있다. 따라서 본 논문에서는 메일의 제목과 HTML 본문만을 추출하고, HTML 본문에서 HTML 태그를 제거하여서 데이터를 구성한다. 그림 3은 제목 및 본문 추출 절차를 보여준다.

메일 클라이언트는 신규 메일이 수신되었을 때 메일을 파싱하고 제목과 HTML 본문을 별도의 테이블에 저장하게 된다. 본 실험에서는 메일 파싱이 이미 처리가 되어있다는 가정하에 제목과 본문 추출을 SQLITE에서 제공하는 C언어 API를 사용하여 테이블에서 직접 추출한다. 이 과정에서 메일 본문이 HTML 포맷으로 되어 있기 때문에 HTML 태그를 제거해야 할 필요가 있다. HTML 태그 제거 과정에서는 기본적인 HTML 문서를 표현하는데 이용되는 모든 정규식 패턴을 메일 본문에서 찾아 공백으로 대체한다. 제목 및 본문 추출 단계에서는 메일 폴더 별로 파일 시스템에 디렉토리를 생성하고, 각각의 메일에 대하여 제목과 본문이 같이 포함된 포맷으로 텍스트 파일을 생성하여 해당 디렉토리에 저장한다. 메일 폴더별로 개별 디렉토리를 생성하여 저장함으로써 다음 단계에서 실행될 불용어 제거 프로그램이 효율적으로 실행하는데 도움을 줄 수 있다.

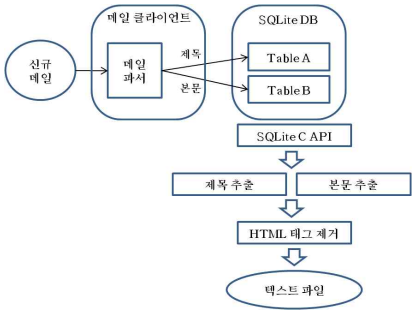


그림 3. 제목 및 본문 추출 절차
Fig. 3. Processes for extracting the title and body of an e-mail

그림 4는 SQLITE의 C API를 사용해서 메일 클라이언트의 데이터베이스에서 제목 및 본문을 추출하고 HTML 코드를 제거하는데 사용할 프로그램이다. SQLITE 데이터 파일을 선택하고, 결과 텍스트 파일이 저장될 디렉토리를 선택한 후 내용 추출을 진행한다.

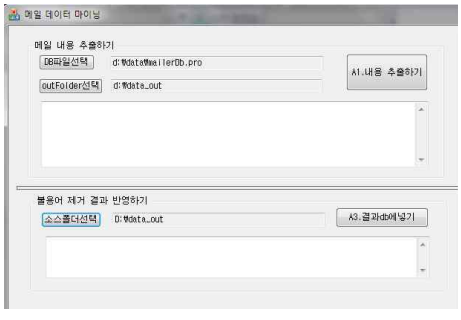


그림 4. 제목 및 본문 추출 프로그램
Fig. 4. A program for extracting the title and body of an e-mail

2. 불용어 제거 및 단어 수 정제를 통한 전처리

HTML 태그를 제거한 후에는, 각각의 텍스트 파일에서 형태소 분석을 통한 불용어를 제거해야 하는데, 본 논문에서는 국민대학교 자연어 처리 연구소에서 제공하는 한국어 불용어 제거기[14]를 사용하였다. 불용어 제거기는 실행파일을 직접 실행시키거나 라이브러리로 개인이 작성한 프로그램에 삽입하는 2가지 방법을 제공하는데, 본 논문에서는 라이브러리로 프로그램에 삽입을 하였다. 불용어 제거 프로그램은 각 메일 폴더에 대응되는 메일들을 추출하여 저장하고 있는 디렉토리 경로를 입력으로 하고, 불용어를 제거한 결과를 각 메일 폴더 디렉토리의 하위 디렉토리인 Index에 텍스트 파일로 출력한다.

불용어가 제거된 파일은 다시 데이터베이스로 입력이 되어야 한다. 본 실험에서는 속도 향상을 위하여, MS-SQL server에서 제공하는 유틸리티인 BCP를 사용하였다. 이 유틸리티는 콘솔에서 실행할 수 있는 프로그램으로서, 텍스트 파일의 내용을 직접 데이터베이스에 대량 삽입(Bulk Insert)하는 용도로 사용된다. 예시로 BCP는 5개의 컬럼을 가진 테이블에 이십만 건의 레코드를 삽입하는데 3.84초(인텔 코어2듀오, 2.13GHz)가 소요된다. 그림 5는 불용어 제거 과정의 흐름을 보여준다. 그리고 그림 6은 BCP를 실행시키기 위한 입력 파일의 포맷으로서 불용어 제거기에서 결과로 출력되는 파일 내의 텍스트 포맷이다.

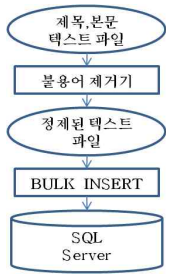


그림 5. 불용어 제거 과정
Fig. 5. Processes for removing disused words

그림 6. 불용어 제거 후 얻어진 결과 파일
Fig. 6. A resulting file after removing disused words

다음 과정으로, 단어 수의 정제를 구현한다. 본 논문에서는 빈발 항목집합 생성을 위하여 Apriori 알고리즘을 사용하는데, Apriori를 구현한 프로그램은 입력 단어 수가 많을수록 긴 실행 시간과 높은 메모리를 사용하는 비효율성이 있다. 따라서 효율성과 정확도를 높이기 위해서 중요성이 낮은 단어를 마이닝 분석 대상 집합에서 제외하는 단어 수 정제 과정을 수행한다. 첫 번째 단계로, 모든 폴더에서 전처리 최소 지지도를 0.04로 설정하고 0.04 미만의 지지도를 가지는 단어들을 제거한다. 두 번째 단계로, 3개 이상의 메일 폴더에서 중복 출현하는 단어들을 제거한다. 제거 방법은 데이터베이스에 존재하는 단어들의 집합에 대해서 SQL 쿼리의 Group by 절을 이용한다. 세 번째 단계로, 단어의 길이가 1인 단어들을 제거한다.

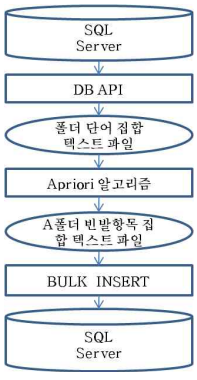


그림 7. 빈발 항목집합 추출 과정
Fig. 7. Processes for mining frequent itemsets

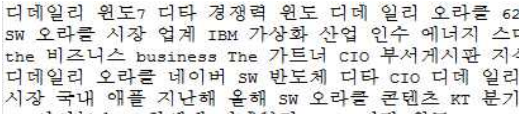


그림 8. Apriori 적용을 위한 뉴스 폴더의 입력 파일 예제
Fig. 8. An example e-mail in a news folder for the Apriori algorithm

3. Apriori 알고리즘을 적용한 빈발 항목집합 추출

전처리 단계가 끝난 후에는 단어 집합을 대상으로 각 폴더별 빈발 항목집합을 추출해야 한다. 빈발 항목집합 추출은 그림 7에 있는 순서로 진행된다. 빈발 항목집합 추출을 위해서 Apriori 알고리즘을 사용할 것이며 본 논문에서 사용한 Apriori 프로그램은 Christian Borgelt[15]가 개발한 프로그램으로서, 텍스트 파일로 입력을 받고, 인수로 지정한 최소 지지도를 넘는 빈발 항목집합을 결과 텍스트 파일에 출력 한다.

우선 데이터베이스로부터 메일 폴더 내의 단어들의 집합을 텍스트 파일로 추출해야 한다. 이 텍스트 파일은 Apriori 프로그램의 입력이 될 것이고, 메일 폴더 별로 생성이 된다. 메일 폴더의 고유키를 텍스트 파일의 이름으로 하고, 파일의 내부 포맷은 한 라인이 메일 하나를 표현하고, 라인에는 메일에 포함되어 있는 단어들이 공백을 구분자로 입력이 되어 있어야 한다. 그림 8은 텍스트 파일의 포맷에 대한 예시이다. 이러한 방법으로 데이터베이스 API를 사용하여 프로그램에서 메일 폴더 별로 텍스트 파일을 해당 디렉토리에 추출한다.

추출된 텍스트 파일들에 Apriori 알고리즘을 적용하는데 있어서, 모든 폴더에 동일한 최소 지지도를 설정한다면, 빈발 항목집합이 없는 폴더가 발생할 수 있다. 따라서 휴리스틱 기법으로 폴더 별로 최소 지지도를 다르게 설정해 준다. 이렇게 추출된 빈발 항목집합을 다시 데이터베이스로 입력하는 방법은 데이터베이스 서버의 대량 삽입 기능을 사용한다. 대량 삽입의 속도를 높이기 위해서 결과 파일들을 우선적으로 병합하고 텍스트 포맷을 빈발 항목집합을 저장할 테이블의 구조에 맞게 변경한 후에 데이터베이스로 대량 삽입을 한다. 그림 9는 빈발항목 추출 프로그램으로서, Apriori 프로그램을 각 폴더 별로 적용하고, 이를 데이터베이스에 입력하는 기능을 한다.

그림 9. 빈발 항목집합 추출 프로그램
Fig. 9. A program for mining frequent itemsets

4. 빈발 항목집합 정제 및 매치 결정 함수 생성

빈발 항목집합을 데이터베이스로 저장을 한 후에, 각각 다른 최소 지지도로 추출이 된 빈발 항목집합들에 대해서, 빈발 항목집합의 IS값의 평균과 표준편차를 구한다. 그리고 IS 값의 평균으로부터 표준편차 이내의 범위에 있는 집합을 구하고, 그 이외의 빈발 항목집합을 삭제함으로써 빈발 항목집합을 정제한다. 이 과정은 프로그램의 개발 과정 없이 SQL 서버의 저장 기능을 이용하여 수행한다. 정제된 빈발 항목집합(RFI)을 훈련 집합에 적용하고, 훈련 집합의 메일들이 RFI에 매치되는 정도의 비율을 고려하여 RFI에 대한 LMS의 값을 결정하며, 해당 과정은 그림 10과 같다.

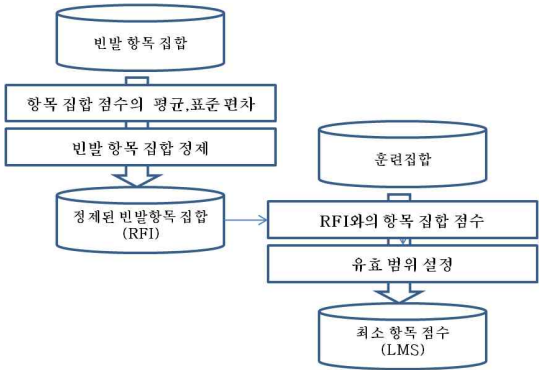


그림 10. 빈발 항목집합 정제 및 최소 항목 점수 도출
Fig. 10. Refining of frequent itemsets and getting a LMS

V. 실험 결과 분석

본 실험을 위한 메일 클라이언트는 직접 개발한 메일 클라이언트를 사용하였고, 클라이언트 내의 메일 데이터는 서버 없이 독립적 실행이 가능한 파일 DB인 SQLITE에 존재한다. 그리고 데이터의 처리 속도를 위해서 추출된 폴더와 메일 별 단어 목록은MS-SQL Server Express에 저장을 하였다. 데이터베이스에서 데이터를 추출하는 방법은 Visual C++을 사용한 프로그램으로 개발하였고, 데이터의 처리는 SQL 질의를 사용하는 저장 프로시져를 이용하여 SQL Server내에서 처리를 하였다.

표 1. 실험에 사용된 데이터의 폴더 구성
Table 1. Folders used in the experiments

폴더	폴더의 주요 컨텐츠
청구서	여러 회사에서 발송한 청구서 (휴대폰, 통신비, 신용카드)
뉴스	여러 회사에서 발송한 IT관련 뉴스레터
취업	여러 회사로부터 받은 구인 제의 메일, 관심 있는 회사의 구인 정보를 알려주는 취업사이트 소식지
동문 모임	대학원 동기들의 대학 생활 및 친목 도모 메일
쇼핑_개인	쇼핑몰에서 구매 후에 발송되는 확인 메일, 배송 관련 메일, 쇼핑몰 관련 Q&A
회사 업무	회사 개발 업무에 관련된 메일

표 2. 폴더별 기초 정보
Table 2. Detailed information of each folder

폴더	메일 개수	단어 수 평균	단어 수 표준편차	단어 길이 평균	단어 길이 표준편차
청구서	231	80.06	93.22	2.88	0.60
뉴스	156	239.42	173.98	2.75	0.68
취업	102	173.84	121.27	2.80	0.54
동문 모임	211	135.81	196.74	2.80	0.84
쇼핑_ 개인	247	134.38	75.05	2.64	0.65
회사 업무	166	110.21	93.23	3.07	0.93

표 3. 훈련 집합과 시험 집합의 메일 개수
Table 3. The numbers of e-mails in a training set and a test set

	1차 – 4차		5차	
	훈련	시험	훈련	시험
청구서	185	46	184	47
뉴스	125	31	124	32
취업	82	20	80	22
동문 모임	169	42	168	43
쇼핑_개인	198	49	196	51
회사 업무	134	33	132	34

1. 실험 데이터 집합

본 논문에서 제안된 기법은 이메일 분류에 필요한 최적 폴더를 추천하는데 있어서 맞춤형 분류를 지원하는데 목적을 두고 있다. 따라서 제안된 기법의 성능 검증을 효과적으로 검증하기 위해서는 정형화된 메일 데이터 집합이 아니라 개인 사용자의 실제 이메일 데이터 집합을 활용하는 것이 실험 결과에 대한 신뢰도를 높일 수 있으며, 이를 고려하여 본 논문의 실험에서는 개인 사용자가 실제 사용중인 이메일 데이터 집합을 활용하였다. 실험에 사용된 데이터 집합은 실험에 사용된 데이터는 6개의 메일 폴더로 구성된다. 메일 폴더의 주요 내용은 표 1과 같고 각 폴더 별 메일 개수와 기초 데이터는 표 2와 같다.

본 논문에서는 실험 결과의 신뢰도를 높이기 위한 방법으로 교차 검증 방법 중 다중 교차 검증 방법을 이용하였다. 교차 검증은 통계적 분석의 결과가 독립적인 데이터 집합으로 일반화 될 수 있는가를 측정하는 방법[12]이다. 다중 교차 검증은 샘플 데이터 집합을 K개의 서브 집합으로 나눈 후, 하나의 집합을 시험용으로 하고, 나머지 K-1개의 집합들을 훈련 집합으로 한다. 그 과정을 모든 서브 집합이 정확히 한번만 시험 집합이 되도록 K번 반복하는 과정을 거쳐서 검증을 하게 되고, 결과는 시험 집합에 대한 결과의 평균값으로 결정된다. 이 방법은 모든 샘플 데이터가 훈련용과 시험용에 다 사용된다는 장점이 있다. 본 논문에서는 전체를 임의의 5개의 서브 집합으로 분할하고, 4개의 서브 집합들을 훈련 집합으로 하고, 1개의 서브 집합을 시험 집합으로 하였다. 즉, 각 폴더의 모든 메일들에 대하여 총 5개의 서브집합으로 나누고, 총 5차례에 걸쳐 다중 교차 검증방법을 사용하였으며, 표 3은 폴더 별 훈련 집합과 시험 집합의 메일 개수를 나타낸다.

2. 전처리 및 빈발 항목집합 추출 과정의 성능 평가

먼저 각 폴더에 대한 전처리 수행에 따른 결과를 확인하기 위한 것으로서 그림 11은 전처리 전과 전처리 후의 단어의 개수를 나타낸다. Apriori 알고리즘은 단어 수가 많을수록 많은 시간과 높은 메모리를 요구하므로, 불용어 제거 및 단어 수정제 등의 전처리 과정을 통해 중요도가 낮은 단어들이 제거되고 빈발 항목집합 추출을 위한 단어 수가 크게 감소된 것을 알 수 있다.

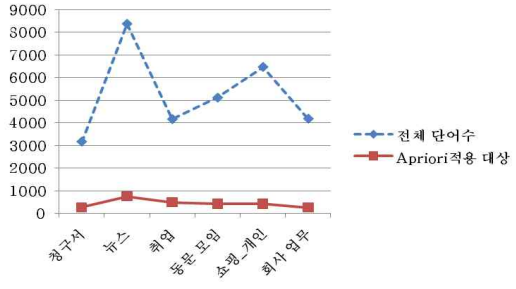


그림 11. 전처리 수행 후 단어 수
Fig. 11. The number of words after preprocessing

그림 12는 다섯 번의 실험에 걸쳐서 추출한 빈발 항목집합의 개수를 나타낸다. 빈발 항목집합을 포함하는 텍스트 파일을 대량 삽입 포맷에 맞게 변경하고, SQL 서버에 저장을 하는 단계는 추출된 텍스트 파일들의 크기에 따라 달랐지만, 프로그램에서 직접 레코드 단위로 데이터베이스에 입력을 하는 방법과 비교해서 약 2배의 속도적인 향상을 볼 수 있었다. 이때 빈발 항목집합을 탐색하기 위한 최소 지지도는 각 폴더에 포함되는 메일들의 특성에 따라 다르게 설정되었다. 각 폴더의 특성을 고려한 최적의 최소 지지도를 설정하는 문제는 보다 복합적인 분석을 필요로 하는 것으로서 향후 좋은 연구 주제가 될 수 있을 것으로 판단되며, 본 논문에서는 휴리스틱 기법으로 빈발 항목집합이 원활히 수행될 수 있는 값으로 최소 지지도 값을 설정하였다. 이러한 과정을 통해 얻어진 빈발 항목집합은 폴더 별로 서로 다르게 설정된 최소 지지도로 인하여 정규화의 필요가 있다. 이를 위해서 IS값으로 평균과 표준편차를 구한다. 그림 13은 빈발 항목집합들의 IS평균과 표준편차를 보여주며, 평균에서 표준편차의 범위 내에 있는 빈발 항목집합들이 최종적으로 정제된 빈발 항목집합이 되며 그림 14에서와 같이 얻어진다.

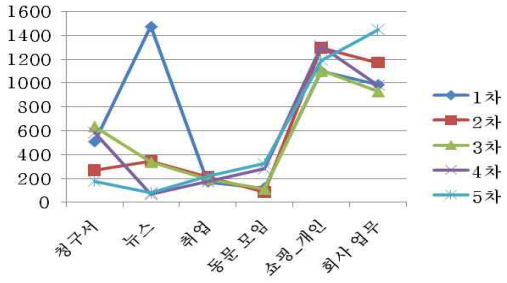


그림 12. 폴더별 빈발 항목집합의 수
Fig. 12. The number of frequent itemsts in each folder

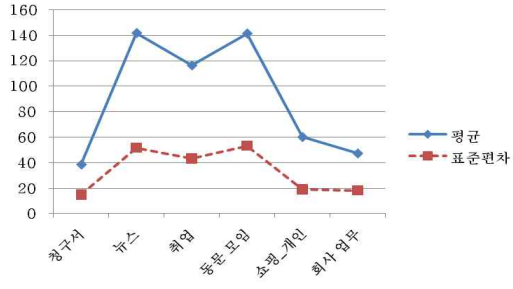


그림 13. 빈발 항목집합의 항목집합 점수의 평균 및 표준편차
Fig. 13. The average and the standard deviation of IS for frequent itemsets in each folder

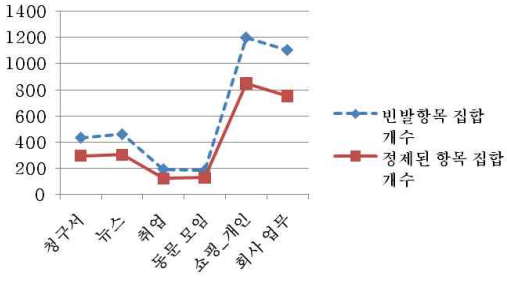


그림 14. 정제된 빈발 항목집합의 개수
Fig. 14. The number of frequent itemsets in a revised set

3. 메일 분류의 정확도 및 오류도

매치 결정 함수의 추출을 위해서는 RFI를 훈련 집합에 적용을 해야 하는데, 그림 15는 각 폴더의 RFI를 훈련 집합에 적용했을 때, 폴더의 훈련 집합의 메일 중에서 RFI와 일치된 빈발 항목집합들을 가지고 있는 메일들의 비율을 나타낸다.

즉, 폴더 f에 대해서 MS(f,m) > 0 을 만족하는 m들의 개수를 의미한다. 꺽은선 그래프는 1차에서 5차 실험까지의 데이터를 표시하고, 이들의 평균은 막대 그래프로 표시하였다. "취업"폴더는 RFI에 일치된 메일의 개수가 타 폴더와는 다르게 낮게 나타나 있는데, 이는"취업"폴더에 있는 메일들의 응집도가 낮음을 보여준다. 응집도가 낮다는 것은 훈련 집합에서 추출한 규칙들을 다시 그 훈련 집합에 적용했을 때 결과가 안 좋게 나오는 것을 의미한다. 응집도가 낮은 폴더는 RFI가 폴더를 대표하는 빈발 항목집합이 될 수가 없음을 보여주고, 시험 집합에 적용했을 때에도 좋지 않은 결과를 나타낼 가능성을 포함하고 있다.

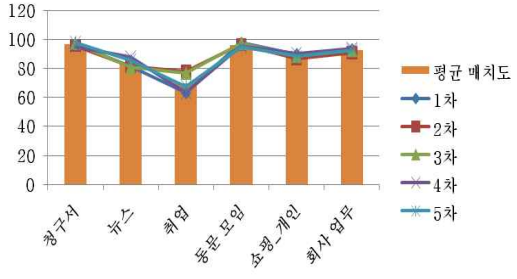


그림 15. RFI를 훈련 집합에 적용했을 때의 매치도
Fig. 15. The matching rate in a training set using RFI

표 4. 폴더별 유효 범위에 따른 정확도 평균값
Table 4. The average accuracy of each folder

	85% 유효범위		90% 유효범위		95% 유효범위	
	정확도	LMS	정확도	LMS	정확도	LMS
청구서	82.28	112.38	86.63	77.32	91.80	69.00
뉴스	81.37	197.50	80.73	123.30	80.73	123.30
취업	69.00	81.36	69.00	81.36	69.00	81.36
동문 모임	92.89	575.00	94.29	552.50	94.29	552.50
쇼핑_개인	76.83	49.92	79.70	44.00	79.70	44.02
회사 업무	87.90	37.22	89.71	32.14	90.93	33.22

폴더 f에 속한 훈련 집합의 MS 값의 분포에서 폴더에의 자동 분류를 결정할 수 있는 MS의 유효 범위를 결정할 수가 있다. 본 실험에서는 MS의 유효 범위를 85%, 90%, 95% 3가지 값으로 하고, 각각의 유효 범위를 만족시킬 수 있는 LMS의 값을 설정하였다. 그리고 시험 집합에 있는 메일들에 대해서 폴더 추천을 테스트 하였다. 시험 집합의 메일 m에 대해서 모든 폴더의 매치 결정 함수 FM(f,m) 값을 적용한 결과가 모든 폴더에 대해서 0이라면 분류를 찾지 못함(Not Found, NF)으로 하고, 0보다 크다면 제일 큰 결과값을 출력하는 폴더를 추천 하였다. 표 4는 각 폴더에 대해서 유효범위별 정확도를 보여주고 있으며, 다섯 차례의 훈련 집합에 대한 실험에서 얻어진 정확도의 평균값을 보여준다.

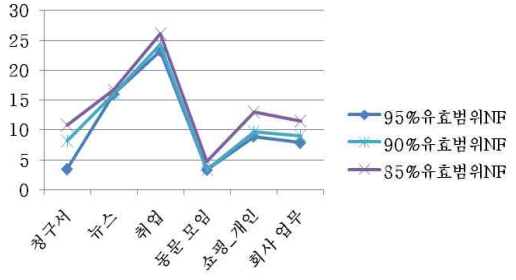


그림 16. 추천 폴더를 찾지 못한 메일의 비율
Fig. 16. The NF rate for each folder

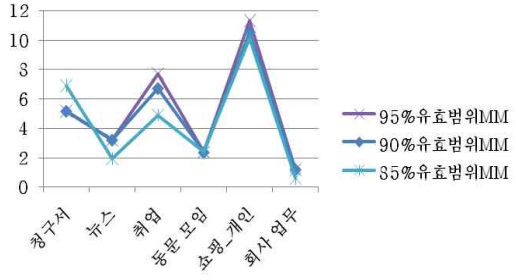


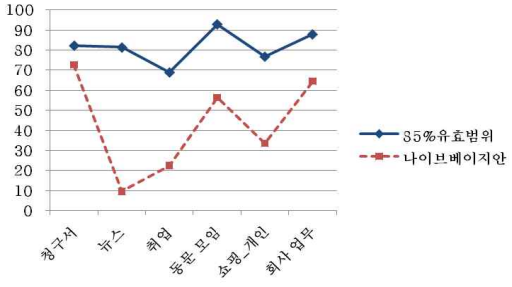
그림 17. 적합한 폴더를 잘못 추천한 메일의 비율
Fig. 17. The MM rate for each folder

다음으로 메일에 대한 폴더 추천 과정에서 오류율을 분석하였다. 오류율은 정확하게 추천하지 못한 메일들의 비율로서, 추천할 폴더를 찾지 못한 경우와 추천할 폴더를 잘못 찾은 경우를 조사하였다. 그림 16은 추천할 폴더를 찾지 못한 비율이고, 그림 17는 추천할 폴더를 잘못 찾은 비율이다. 이 비율은 각 경우의 메일 개수를 폴더 내의 메일 개수로 나눈 값이다. 유효 범위가 클수록, 찾지 못하는 비율은 낮아지고 잘못 분류할 비율은 커진다는 것을 알 수 있었다. 그리고 응집도가 낮은 "취업"폴더는 분류를 찾지 못하는 메일이 잘못 분류된 메일보

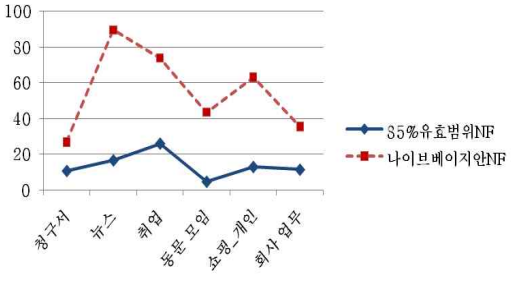
다 많음을 보여준다. 잘못 분류된 메일의 비율이 가장 높은 "쇼핑 개인"폴더의 데이터를 살펴보니, 대부분의 메일이 "청구서"로 분류가 되어 있었다. 이는 "쇼핑 개인" 폴더의 메일 내용이 주로 쇼핑몰에서 온 구매 결과이거나 정보 변경이었고 메일 내용에 카드 결제 내역이 포함되어서였다.

4. 기존 방법과의 비교

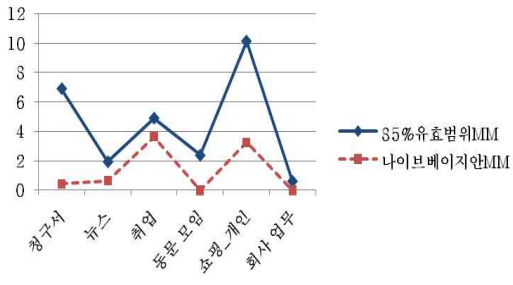
본 논문에서 제안된 기법의 성능을 비교하여 검증하기 위해서 문서 분류 기법 중 가장 널리 쓰이는 나이브 베이지안 기법[2]과 비교 실험을 수행하였다. 나이브 베이지안 기법은 불용어를 제거하는 전처리 과정을 거친 후에, 다항식 이벤트 모델 기반인 단어 빈도에 가중치를 주는 방법을 사용한다. 나이브 베이지안 기법에 대하여 본 논문이 사용하는 교차검증 방법을 동일하게 적용하였으며, 5차례 실험을 거친 결과의 평균값을 본 논문에서 제안된 기법의 결과와 비교하였다. 그림 18은 이러한 비교 실험 결과를 보여주며, 실험에서는 본 논문에서 제안된 기법의 유효범위는 85%로 설정되었다.



(a) 정확도(classification accuracy)



(b) 미추천 오류율 (NF rate)



(c) 잘못된 추천 오류율 (MM rate)

그림 18. 나이브 베이지안 방법과의 비교
Fig. 18. Comparing with the Naive Bayesian method

먼저 그림 18-(a)는 단어 빈발 가중치 기반의 나이브 베이지안 분류기와 본 논문에서 제안된 기법의 정확도의 비교 실험 결과이다. 정확도는 다섯 번의 실험에서 얻어진 결과의 평균값으로 구하였다. 그림에서 알 수 있듯이 본 논문에서 제시한 방법에서 적합한 폴더로 이메일을 분류하는 비율이 높음을 알 수 있다. 이메일은 정규화가 되지 않은 불규칙적인 문서로서 나이브 베이지안이 비정규적인 문서에 대한 분류 정확도가 낮고 또한 단어 간의 유사어를 설정하는 사용자 사전을 반영하지 않았기 때문이라고 볼 수 있다[2]. 하지만 본 논문에서 제시한 기법도 사용자 사전을 반영하지 않았음을 고려한다면 논문에서 제안된 기법이 우수한 성능을 보임을 알 수 있다. 그림 18-(b)와 그림 18-(c)는 해당 실험에서 적합한 폴더를 찾지 못하는 오류율과 적합한 폴더를 잘못 찾은 오류율을 나타낸다. 그림 18-(b)에서 보듯이 본 논문에서 제안된 기법에서는 나이브 베이지안 방법에 비해 미추천 오류률이 매우 낮다. 즉, 상대적으로 많은 수의 메일들에 대해서 적합 폴더를 추천한다. 이로 인해 그림 18-(c)에서와 같이 잘못된 추천 오류률은 상대적으로 다소 높게 나타났다. 한편, 표 4에서 보듯이 본 논문에서 제안된 기법에 있어서 유효범위가 85%인 경우는 정확도가 가장 낮은 경우로서 보다 높은 유효범위로 설정하는 경우 정확도가 증가되어 비교실험에서 보다 우월한 성능을 보인다.

Ⅵ. 결 론

본 논문에서는 메일의 개인 맞춤형 폴더 추천을 위하여 전처리와 빈발 패턴을 적용하고, 그 결과로 폴더의 프로파일을 생성하는 기법을 제안하였다. 객관적인 분류 기준이 명확한 웹 문서나 뉴스와는 다르게 주관적인 분류 기준을 가지고 있는 이메일을 대상으로 마이닝 기법을 적용함으로써 개인마다 다른 분류 기준을 수용할 수 있었다.

제안된 방법은 한글로 작성된 정형화되지 않은 실제의 이메일을 대상으로 실험하였다. 결과를 보면, 동호회 활동이나 회사업무와 같이 의미의 집중도가 높아서 응집력이 좋은 폴더는 정확도가 90%수준으로 높았고, 의미 집중도가 낮은 메일 폴더는 69%로 낮게 나왔다. 정확도가 낮은 폴더는 추천 폴더를 잘못 찾은 메일 수보다 추천할 폴더를 못 찾은 메일이 더 많았다. 현업에서 폴더 추천 기법을 적용한다면, 적합한 폴더를 못 찾은 메일들에 대해서 사용자가 적정 폴더를 지정하는 사용자 인터페이스를 제공함으로써 오류율을 개선할 수 있다.

메일함 관리 기준은 사용자의 필요에 따라 다양한 방법이 적용될 수 있다. 특히 이메일의 수신자/발신자 정보 등과 같이 정형화된 정보를 활용하여 분류 작업을 수행하기도 한다. 본 논문에서 제안한 기법은 이와 같은 기존의 방법들을 대체하기 위한 것이 아니라 메일함 관리에 적용될 수 있는 여러 방법들 중의 하나로서 제안된 것이며, 특히 정형화된 정보를 활용한 분류가 아니라 이메일의 내용에 기반한 분류를 수행하는데 그 특징을 두고 있다. 근래 들어 웹 문서를 포함한 일반적인 문서 분류 및 검색에서도 정형화된 정보에 기반한 분류뿐만 아니라 내용에 기반한 분류 및 검색에 대한 관심이 증가되고 있으며, 이를 고려할 때 내용 기반 이메일 분류 기법도 메일함 관리를 위한 주요 방법이 될 수 있으리라 판단된다. 한편, 문서 분류 과정에서 정확도를 높이기 위한 방법으로 특정 위치에 나타나는 단어나 특정한 의미를 갖는 단어에 대해서 상대적으로 높은 가중치를 부여하거나 출현빈도는 높지 않으나 중요한 의미를 갖는 특이 단어를 고려하기도 한다. 이러한 기법들을 본 논문에서 제안된 이메일 분류 시스템에 적용하는 경우 이메일 분류 정확도를 보다 향상시킬 수 있다.

한편, 하나의 폴더에 존재하는 메일 집합에 대하여 빈발 항목집합으로 정의되는 특성 프로파일을 구하는 과정으로 제안된 기법에서는 해당 메일 집합들에 대한 마이닝을 통해 빈발 항목집합을 얻은 후 항목집합점수 등에 근거하여 정제된 빈발 항목집합을 얻고 이를 해당 폴더의 특성 프로파일로 활용한다. 이 과정에서 항목집합 점수에 대한 정의 및 빈발 항목집합을 정제하기 위한 허용범위 등에 따라 하나의 메일 폴더에 대한 특성 프로파일이 다양화될 수 있으며, 사용자의 요구 등을 고려하여 최적의 설정 값을 구하는 것도 흥미로운 연구 주제가 될 수 있다. 또한 논문에서 제안된 기법에서는 각 폴더의 빈발 항목집합 탐색을 위한 최소 지지도를 휴리스틱에 의해 설정하며, 빈발 항목집합이 크게 증가하는 임계치를 기준으로 설정을 하였다. 따라서 실험 과정에 대한 전체적인 자동화가 어려웠다. 빈발 항목집합 탐색을 위한 최소 지지도 등의 매개변수에 대해서 처리 대상이 되는 데이터 집합의 특성 등을 고려하여 해당 매개변수의 최적 설정값을 구하거나 이를 자동으로 설정함으로써 제안된 방법의 성능을 향상시키는 문제 또한 관련 분야에서 흥미로운 향후 연구 주제가 될 수 있다. 또한 하나의 폴더에 속하는 메일들 간의 의미 집중도가 낮은 경우 빈발 항목집합 탐색 과정에서 유효한 결과를 얻는데 어려움을 겪게 되고, 따라서 추천의 정확도가 낮아질 수 있다. 이러한 단점을 보완하기 위한 연구들도 흥미로운 향후 연구 주제가 될 수 있으며, 이러한 향후 연구들을 통해 제안된 기법의 성능을 향상 시킬 수 있을 것으로 판단된다.

참고문헌

[1] O.-R. Jeong, D.-S. Cho, "A Three-Step Preprocessing Algorithm for Enhanced Classification of E-Mail Recommendation System," The Transactions of The Korean Institute of Electrical Engineers, Vol. 54D, No. 4, pp. 251-258, 2005.

[2] K. P. Kim, Y. S. Kwon, "Performance Comparison of Naive Bayesian Learning and Centroid-Based Classification for e-Mail Classification," IE Interface, Vol. 18, No. 1. pp. 10-21, 2005.

[3] O.-R. Jeong, D.-S. Cho, "A Recommendation Agent System for E-mail Classification," The Proc. of the KISS Spring Conference, pp. 94-96, 2003.

[4] J.M. Lee, "An Improvement of Accuracy for NaiveBayes by Using Large Word Sets," Journal of Korean Society for Internet Information, Vol. 7, No. 3, pp. 169-178, 2006.

[5] S.J. Ko and J.H. Lee, "Weighted Bayesian Automatic Document Categorization Based on Association Word Knowledge Base by Apriori Algorithm," Journal of Korea Multimedia Society, Vol. 4, No. 2, pp. 171-181, 2001.

[6] S.J. Ko and J.H. Lee, "Bayesian Automatic Document Categorization Using Apriori - Genetic Algorithm," Journal of the KIPS, Vol. 8, No. 3, pp. 251-260, 2001.

[7] M. Ryu, J.S. Park, and J.K. Kim, "A Knowledge-based Folder Recommendation Procedure for e-mail Classification," The Proc. of the KIISS Fall Conference, pp. 349-357, 2004.

[8] Diao, Y., H. Lu, and D. Wu, "A Comparative Study of Classification Based Personal E-mail Filtering," in Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data

Mining, Current Issues and New Applications, pp. 408-419, 2000.

[9] H.-J. Kim, J. J. Jeong, and G.-S. Jo, "Spam-Mail Filtering System Using Weighted Bayesian Classifier," Journal of the KISS: Software and Applications, Vol. 31, No. 8, pp. 1092-1100, 2004.

[10] Yin, X., J. Han, "CPAR: Classification based on Predictive Association Rules," in Proceedings of the third SIAM International Conference on Data Mining, pp. 331-334, 2003.

[11] Liu, B., W. Hsu, Y. Ma, "Integrating classification and association rule mining," in Proceedings of the fourth International Conference on Knowledge Discovery and Data Mining, pp. 80-86, 1998.

[12] Tan, P.-N., Introduction to Data Mining. INFINITY BOOKS, 2007.

[13] Wood, D., Internet Email Programming. Hanbit Media, 2000.

[14] HAM, *http://nlp.kookmin.ac.kr/HAM/kor/index.html.*

[15] Apriori program, *http://www.borgelt.net/apriori.html.*

저 자 소 개

문 종 필

2000년 : 연세대학교 컴퓨터과학과 졸업 (이학사)

2010년 : 연세대학교 공학대학원 컴퓨터공학과 졸업 (공학석사)

2010년1월~현재 : KT Innotz 연구개발본부

E-mail : allblack@naver.com

이 원 석

1985년 : Boston University 컴퓨터과학 (공학사)

1987년 : Purdue University 컴퓨터과학 (공학석사)

1990년 : Purdue University 컴퓨터과학 (공학박사)

2004년3월~현재 : 연세대학교 컴퓨터과학과 정교수

E-mail : leewo@database.yonsei.ac.kr

장 중 혁

1996년 : 연세대학교 컴퓨터과학과 졸업 (이학사)

1998년 : 연세대학교 대학원 컴퓨터과학과 (공학석사)

2005년 : 연세대학교 대학원 컴퓨터과학과 (공학박사)

2006년1월~2008년7월 : UIUC, Wright State Univ. 박사후 연구원

2008년9월~현재 : 대구대학교 컴퓨터IT공학부 교수

E-mail : jhchang@daegu.ac.kr