DOI: 10.3745/KIPSTB.2008.15-B.5.493

사용자 프로파일에 기반한 전자 메일의 중요도 결정

이 상 곤†

요 약

개인 통신 방법의 수단으로 전자 메일이 널리 사용되고 있으나, 업무에 직접 관련이 없는 쓸모없는 상업용 메일도 대량으로 유포되고 있다. 본 연구에서는 사용자가 작성한 프로파일을 이용하여 메일을 자동으로 그룹핑(grouping) 하는 방법을 제안하고자 한다. 기존의 연구 방법은 단어의 빈도수만을 이용하는 단일 속성을 이용하므로 높은 정확률을 얻을 수 없었다. 그러나 본 논문에서 제안하는 방법은 기존 사용자의 폴더에 수신된 메일의 분류 체계에서 추출된 사용자 프로파일을 이용하여 그룹핑 되는 메일의 정확률을 높이고자 한다. 본 논문에서 적극적으로 이용하는 정보는 다중 속성(송신처, 문서의 주제, 문서의 유형 정보, 시간제한 표현의 어구 등) 값이다. 사용자의 프로파일을 이용함으로써 새로 도착한 메일이 사용자에게 중요한가 혹은 중요하지 않은가의 자동 판단이 가능하도록 시스템을 설계하였다. 학습 데이터를 네 가지 형태로 나누어 실험한 결과 Rocchio와 Widrow-Hoff의 상관계수가 각각 0.40과 0.43인 것 보다 본 논문의 방법이 0.52로 보다 높은 상관계수를 나타내 빈도수만을 이용하는 기존의 연구보다 정확한 방법임을 알 수 있었다.

키워드 : 전자 메일 처리, 다중 속성, 사용자 프로파일, 필터링, 그룹핑, 문서 분류

Decision Method of Importance of E-Mail based on User Profiles

Samuel Sangkon Lee†

ABSTRACT

Although modern day people gather many data from the network, the users want only the information needed. Using this technology, the users can extract on the data that satisfy the query. As the previous studies use the single data in the document, frequency of the data for example, it cannot be considered as the effective data clustering method. What is needed is the effective clustering technology that can process the electronic network documents such as the e-mail or XML that contain the tags of various formats. This paper describes the study of extracting the information from the user query based on the multi-attributes. It proposes a method of extracting the data such as the sender, text type, time limit syntax in the text, and title from the e-mail and using such data for filtering. It also describes the experiment to verify that the multi-attribute based clustering method is more accurate than the existing clustering methods using only the word frequency.

Keywords : E-Mail Processing, Multi-attribute, User Profile, Filtering, Grouping, Document Classification

1. 서 론

현대에는 컴퓨터 네트워크의 기술로 인하여 전자 메일(E-Mail; Electronic Mail)이 중요한 통신 수단이 되었으며, 업무상 중요한 문서도 메일 클라이언트를 통하여 유포되고 있다. 이에 의해 개인별 통신 방법이 간단해지고, 많은 양의 메일이 개인의 컴퓨터에 보내지고 있다. 그러나 수신된 메일이 모두 중요한 메일은 아니다. 예를 들면, 상업용 메일에 관심이 없는 사용자에게는 이러한 상업용 메일이 더 이상 중요한 메일이 아니다. 따라서 사용자별로 중요한 메일과 중요하지 않은 메일을 구별하기 위한 연구 방법이 개발되어야 한다.

본 논문에서는 사용자의 메일 폴더에 저장된 메일이 담고 있는 개인의 이력 정보에 주목하고자 한다. 각 개인의 프로파일[15,16]을 기반으로 중요도를 추출하고, 이 중요도를 순서화하여 새로운 메일의 중요도를 판단하는 속성값을 추출하고자 한다. 이러한 개인 사용자의 프로파일은 사용자가 미리 지정한 우선순위가 부착된 기존의 메일에 존재하는 메일 폴더에서도 추출 가능하다. 본 논문에서 제시하는 방법은 중요도를 결정하는 요소(factor)로서 다중 속성값(Multi-attribute Values)을 이용한다. 이러한 다중 속성값은 메일의 송신처(Sender)나 충고, 요구, 질문 등과 같은 문장의 유형(Sentence Type), 문장의 주제(Theme), 수신자가 메일을 확인하여 (최소한으로) 회신(답변)하여야 하는 시간제한(Time Limit) 등으로 구성되어 있다. 앞에서 언급한 네 가지 요소들 각각을 '속성'이라 정의한다. 그리고 사용자 개인별로 작성된

† 종신회원 : 전주대학교 컴퓨터공학과 부교수
논문접수 : 2008년 2월 21일
수 정 일 : 2008년 8월 14일
심사완료 : 2008년 9월 1일

프로파일은 사용자 개인의 메일 계정 사용 환경에 대한 정보를 의미하며, 사용자가 처음으로 자신의 계정으로 로그인할 때 만들어진다. 이 프로파일은 이미 작성되어 컴퓨터에 저장되어 있는 메일의 속성값과 우선순위를 함께 조합하여 구성할 수 있다. 이 프로파일을 사용하면 여러 명의 사용자가 같은 계정을 사용할 때에도 개인별 메일 사용 환경을 가질 수 있게 된다.

다음은 여러 관련 연구를 문헌하여 본 논문의 방법과 비교 연구에 대하여 기술하였다.

[이지행 외, 2002]은 포털 사이트의 전자메일 자동 분류를 위해 속성 선별 처리 과정과 최대값, 신경망 및 진화 연산 등을 결합하여 다중 신경망을 이용한 문서 분류 방법을 제안하였다[4]. [강영순 외, 2002]는 전자메일의 문서에서 발생하는 잡음 요소를 제거하고 반구조적(semi-structure) 특성을 활용하기 위해 전자메일 분류기의 전처리 모듈을 설계하고 이를 분류기에 통합시키는 방법을 제안하였다[5].

[안희국 외, 2002]는 동적 시소러스와 유전자 알고리즘을 적용하여 사용자의 선호도를 반영할 수 있는 다양하고 광범위한 정보의 추출을 강조하고 있다. 전자메일의 선호도 분류를 위해 1차와 2차 분류로 나누어 제안하고 있다. 1차 분류에서는 사용자의 적합도를 판단하기 위해 동적 시소러스를 구축하고, 구축된 시소러스의 비교를 통해 어떤 메일이 사용자에게 유용한지를 결정한다. 2차 분류에서는 사용자가 지정한 폴더 키워드를 중심으로 사용자 시소러스로부터 유전자 알고리즘을 이용해 추출한 키워드들과의 적합도 비교를 하고, Folder Keyword를 이용하여 특정 폴더로 분류가 이루어진다[6,9].

[류제 외, 2003]은 스팸 메일을 분류하기 위해 몇 개의 단어와 문구를 이용하여 스팸 메일의 여부를 판단한다. 여러 차례의 반복 학습을 통해 특정 속성(feature set을 분리하여 이용하는 방법)과 Co-training을 추출하고 이를 스팸 분류 기술로 이용하고 있다[10].

[현영순 외, 2003]은 대량의 메일을 송수신하는 경우, 메일에 대한 효율적 관리 문제와 불필요한 메일에 대한 관리의 중요성에 대해 메일에 첨부된 파일을 자동으로 분류하는 메일 클라이언트를 설계하여 사용자가 메일을 하나하나 읽고 분류하여야 하는 부담을 덜어 주어야 한다고 주장하였다[11].

[안찬민 외, 2004]은 통계적인 방법인 PCA; Principal Component Analysis와 SVD; Singular Value Decomposition에 기반한 문서 요약 방법과 퍼지 이론을 기반으로 한 동적 분류 체계 방법을 결합하여 전자메일의 다원 분류 방법을 제안하였다[13].

[변영철 외, 2004]의 연구에서는 사용자는 인터넷 사이트 운영자에게 질문하고자 하는 질의 메일을 자동으로 분류하기 위해 신경망에 기반한 분류 기법을 제안하였으나, 전자메일 간 유사도 계산 방법과 질의어에 적응적인 키워드 결정 방법에 관한 연구가 부족하였다. 따라서 본 논문에서는 이에 대한 연구를 하였다[14].

[김보미 외, 2005]는 빠른 업무 처리를 위해 중요도가 높은 메일을 먼저 처리하고자 하는 필터링 기술을 제안하였다. 수신된 메일의 송신처, 제목, 문서 유형, 시간 제한 등의 속성값을 조합하여 구조적인 지식을 획득하고 이를 이용하여 필터링 하는 연구를 수행하였다[15].

[장정효 외, 2006]는 전자메일의 데이터 형식과 구조를 파악하고 메일의 인코딩 방법을 이용하여 전자메일의 필터링 시스템을 구현하였다. 이 연구에서 제안하는 시스템은 사용자가 특별하게 중요성을 느끼고 있는 특정 송신처, 한국어의 부사어구를 토대로 조사한 문서 유형, 특정한 시간을 나타내는 문구 등을 추출하여 사용자의 프로파일 정보를 저장하고 이 정보를 이용하여 신속히 처리하여야 할 전자메일을 업무에 적용시키고자 하였다[16].

이상의 참고 문헌들은 메일에서 주요 단어, 날짜, 장소 등의 언어적인 특징을 추출하고 추출된 표현을 패턴 매칭 하는 규칙기반 방법이다. 메일의 중요도 판정 시 미리 규칙화되어 저장된 중요한 표현 패턴을 매칭 하여 중요도를 계산하는 방법이다. 그러나 중요성의 판단 기준이 각 개인마다 다르고, 메일 문서 내에 포함되어 있는 이력 정보에 따라 좌우되기 때문에 규칙 기반 방법만으로 필터링을 적절히 대응할 수 없다. 정보 검색 방법의 하나인 LSI법[12]을 적용하여, 전자 뉴스 기사와의 유사성을 계산하여 중요한 메일을 추정하는 방법도 있으나 LSI법 자체에 계산량이 너무 많아 메일 문서의 처리와 같이 실시간 처리가 요구되는 분야에는 부적절한 경우가 많다.

메일 문서의 필터링 연구를 주목하면 수신된 메일에 대한 사용자의 행동(참조 시간 등)과 단어의 빈도를 기초로 작성한 프로파일 정보를 중요도의 판정에 이용할 수 있다. 그러나 단어의 단일 속성만을 대상으로 하므로 정밀한 추출이 어렵고, 학습 데이터 수가 적은 경우에는 효과적으로 대처할 수 없다. 메일 문서 내에서 추출한 다중 속성 값의 내용에서 메일 문서를 순위화 하는 방법도 제안되어 있다. 송/수신된 메일의 이력 정보에서 사용자가 중요성을 지정한 속성 값에 높은 가중치를 부여하고, 여러 속성 값의 조합으로 메일 문서의 중요도를 계산한다. 그러나 실제의 메일 문서 내에는 같은 속성이라도 개별적인 속성 값이 서로 다르고, 때때로 중요도도 달라져 속성 값마다 중요 계수를 개별적으로 부여하고 있다. 따라서 다중속성 값의 조합을 고려한 방법이 아니면 메일 문서를 정확하게 필터링 할 수 없다.

이러한 기존의 연구들과 본 논문에서 제안하는 방법과 차별성을 강조하면 다음과 같다.

[강영순 외, 2002]는 전자메일의 특성과 관련된 연구의 필요성을 강조하였다. 본 논문의 방법은 전자메일의 다중 속성 값을 추출하기 때문에 전자메일의 특성을 파악하여 비정형적인 특성을 최소화함으로써 분류 성능을 개선하는데 있다.

본 연구에서는 기존에 수신된 메일 문서에서 학습한 지식을 이용하여 필터링 한다. 또한 다중 속성 값의 조합으로 프로파일을 작성하는 점에서 기존의 연구 방법들과 다르고, 다중 속성 값의 조합을 고려하여 중요도를 산출한다. 앞에서 제시한 참고 문헌들은 제목이 되기 쉬운 중요어구를 수집하

여 등록한 사전을 이용하지만, 본 논문의 방법은 학습 데이터에서 수집한 단어를 자동으로 학습한다. 또한 기존의 연구들은 메일 문서에 대해 우선순위를 사용자의 행동에서 추정하였으나, 본 논문의 방법은 이런 점을 이용하지 않아도 효과적인 필터링이 가능한 장점이 있다.

다음으로 2장에서는 메일의 필터링에 사용되는 정보로써 중요도 측정에 사용되는 여러 가지 요소들에 대해 언급하고, 제 3장에서는 이미 존재하는 메일 문서에서 사용자의 개인별 프로파일을 생성하는 수학적인 방법에 대하여 설명하고, 새로 도착한 메일 문서(우선순위를 알 수 없는 문서)의 중요도를 계산하는 방법에 대해서도 아울러 서술한다. 마지막으로 4장과 5장에서는 실험 결과와 향후의 연구 과제에 대하여 각각 서술한다.

2. 메일의 중요도

2.1 중요도의 결정 요소

이 절에서는 메일 문서에서 흔히 나타나는 중요도의 특징을 다음과 같이 두 가지로 정의한다.

(1) 다른 문서보다 빨리 읽고 처리하여야 하는 메일 문서가 있다.

(2) 빠른 업무 처리를 위해 즉시 답변이 필요한 메일이 있다.

본 연구에서는 위의 두 가지 정의에 의해 문장의 유형, 시간제한 표현의 유무, 문서의 주제 등에 따라 메일의 중요도를 결정하고자 한다. 덧붙여 말하면, 일반 문서에는 없으나 메일 문서에만 나타나는 특정 정보 예를 들면, 메시지의 송신처, 참조(Cc), 따옴표와 본문의 텍스트 사이의 관계, 과거의 메일 문서와의 관련성 등은 중요도를 판정하는데 효과적인 정보이다. 따라서 형태소 분석의 결과 얻을 수 있는 다음의 네 가지 속성들을 중요도 계산을 위한 주요한 요소라 한다.

- α : 송신자(처),
- β : 문장의 유형,
- γ : 시간제한 표현의 유무,
- θ : 메일의 주제(제목) 등

위 네 가지의 개별적인 요소들은 중요도를 결정하는 '속성'이라 정의하고, 각 속성의 값들은 "속성값"이라 정의한다. 이 속성값들은 다음의 추출 방법을 이용하여 얻을 수 있다.

속성값 α는 메일 문서의 헤드 정보 중 "From"에서 얻는다. β는 본문 텍스트에 포함된 문장의 유형을 나타내는 표현어구(영어의 경우, 현재분사나 조동사) 등에서 추출하여 중요한 초점 정보로 사용된다. 속성값 γ는 스케줄을 나타내는 부사나 시간의 정도를 나타내는 명사로부터 얻는다. θ는 문서의 주제를 나타내는 명사 사전을 이용하여 추출한다. 파싱이나 의미 분석과 같은 고급의 자연언어처리 기술을 적용하면 세세한 형태의 주제를 추출할 수 있으나, 실시간 처리가 가능하도록 하기 위해 고급 기술을 적용하지 않는다. 본 논문에서는 메일의 적당한 주제를 추출하기 위해 이미 저장된 메일을 학습 데이터로 이용하여 사용자의 개입을 토대로 문서의 주제를 파악하는 방법을 이용하였다.

본 논문에서는 미리 사용자가 메일 클라이언트에 도착해 보관 중인 각 메일이 사용자의 메일 폴더에 잘 정리되어 있다고 가정한다. 사용자는 이러한 정리 작업을 통해 유사한 메일(비슷한 내용의 주제를 갖는 메일)은 동일한 폴더에 저장한다는 것을 의미한다. 이렇게 메일이 저장되어 있는 메일 폴더의 이름은 θ의 속성값을 이용하여 추출한다. 새로 도착한 메일은 현존하는 메일 폴더에서 가장 유사한 폴더의 이름을 이용하여 θ 값을 결정한다. 유사한 정도를 나타내는 값은 Vector Space Model에 기반하여 단어의 빈도수를 계산하여 결정한다. 즉 동일한 폴더에 존재하는 각 메일들의 단어 빈도수로 구성된 단어 벡터를 구성하고, 새로운 메일에서 신규로 구성한 단어 벡터와 비교한다. 위의 두 가지 벡터는 코사인(Cosine) 방법[12]으로 비교한다. 새로운 메일의 유사성은 기존의 메일 폴더 중 가장 유사한 폴더를 이용하여 결정한다.

2.2 중요도의 개인차

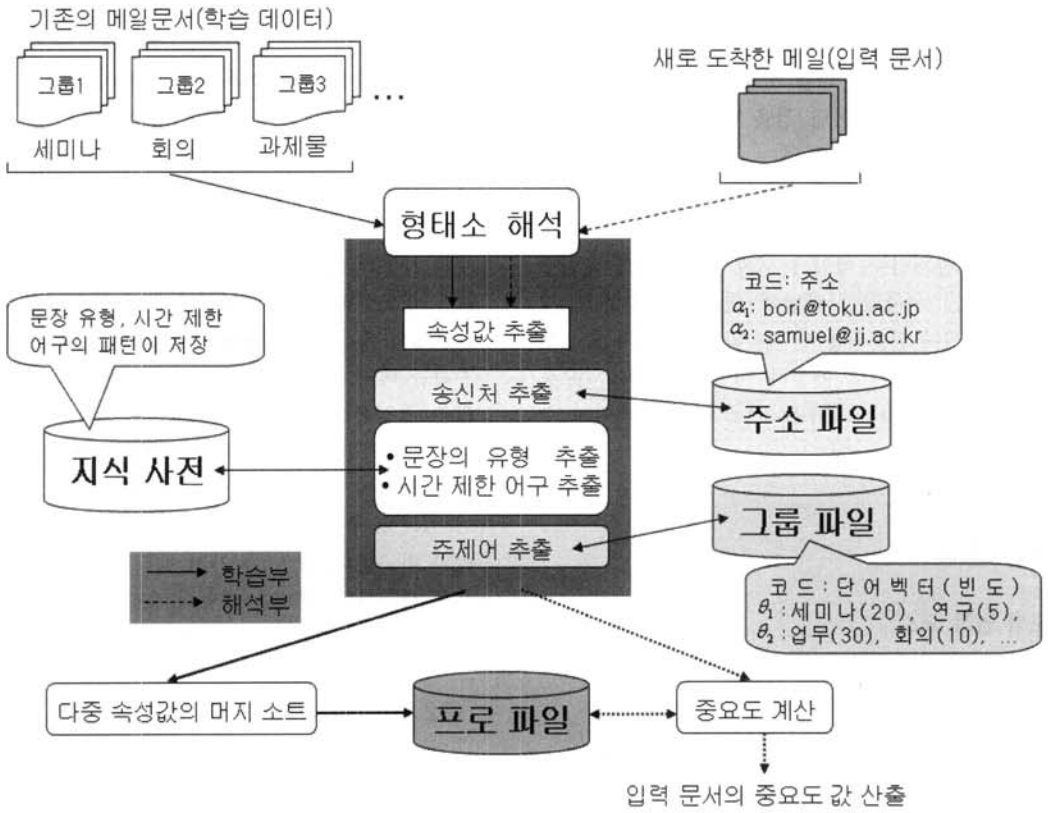
앞의 2.1절에서 서술한 관점으로부터 메일의 중요도를 결정할 때에 개인 사용자 별로 메일의 중요도를 다르게 설정할 필요가 있다. 특정한 송신처의 경우 예를 들면, 학생 A는 교수 X(교수 X의 α의 값)로부터 온 메일에는 매우 높은 중요도를 부여하고 있으나 반면에, 학생 B는 교수 X에게서 온 메일에 별로 중요성을 갖고 있지 않다고 생각할 수 있다. 이와 같이 본 논문에서 제안하는 시스템은 각기 다른 사용자별로 별도의 중요도의 속성값을 갖도록 고려하였다.

위의 경우와 마찬가지로 동일한 송신처를 갖지만, 문서에 포함된 문장의 유형에 따라 다른 중요도가 부여되어야 하는 경우도 있다. 예를 들어, 학생 A는 '교수 X'에게 온 '충고(α의 값은 '교수 X'이며, β의 값은 "충고(advice)"인 경우)'의 문장 유형은 높은 중요도를 갖는다고 생각하지만, 반면에 동일한 송신처라고 하지라도 '요구(α의 값은 '교수 X'이며, β의 값은 "요구(request)"인 경우)'에 대한 문장 유형은 중요하게 생각하지 않을 수 있다. 학생 B의 경우에는 위의 학생 A와는 반대의 경우로 생각할 수 있다. 이와 같이 속성값이 서로 상이하게 조합될 수 있으며, 조합된 속성값도 개인별로 차이가 크게 날 수 있다. 따라서 메일 문서의 중요도는 많은 속성값들의 조합에 의해 결정되어야 하고, 각 속성값 마다의 가중치도 각 사용자 개인별로 차이가 있도록 시스템을 설계하여야 한다.

3. 중요도 결정 방법

3.1 개요

본 논문에서 제시하는 방법을 학습 모듈과 분석 모듈로 각각 나누어 위의 (그림 1)에 설명하였다. 학습 모듈에서는 이미 사용자에게 송신되어 저장된 메일로부터 개별 사용자의



(그림 1) 필터링 방법의 개요도

프로파일이 작성된다. 분석 모듈에서는 새로 도착한 메일에 대한 클러스트링 작업이 수행된다. 이 분석 모듈에 의해, 새로운 메일 문서의 우선순위는 개인 프로파일에 의해 계산된다. 학습 모듈의 흐름은 그림에서 실선으로 나타내었다. 이 모듈에서는 메일이 각 폴더로 나뉘어 분류 및 정리되어 있으며 메일 문서는 형태소 해석되어 사용자가 제시한 우선순위가 부착되어 있다. 형태소 해석 후 각 문서로부터 추출된 속성값이 추출된다. 속성값의 추출은 첫째, 표현 패턴의 수가 유한(finite)하고, 둘째 문장 유형에 해당하는 단어와 시간제한 표현의 어구가 개인별로 큰 차이가 없으며, 셋째 표현되는 많은 수의 패턴이 배경 지식 사전에 기 등록되어 있으므로 메일에 존재하는 문장의 유형에 해당하는 단어와 시간제한 어구 등을 이용한다. 배경 지식 사전에 있는 패턴과 메일에 출현한 단어 사이의 매칭에 의해 속성값이 추출된다. 메일의 헤더(header) 정보로 코드화되어 있는 'From' 항목으로부터 추출된 메일 주소와 송신처가 대응된다. 메일 주소와 송신처의 대응은 숫자로 코드화되어 주소 파일(address file)에 저장된다.

메일을 저장하고 있는 폴더 이름은 "Theme"으로 코드화하여 속성값으로 저장한다. 각 폴더의 서명(signature)으로 구축된 단어 벡터(word vector)를 모든 폴더의 각 메일에 대하여 구축한다. 헤더 정보의 "Subject" 항목에서 추출된 명사와 본문 텍스트에서 추출된 명사는 단어 벡터에 의해 저장된다. 단어 벡터는 원래의 단어와 빈도수의 쌍으로 구성된다. "Theme"과 해당하는 단어 벡터는 그룹 파일(group file)에 저장된다. "Subject"는 메일 내용의 제목이나 요약으로 판단되므로 "Subject" 내에서 출현한 단어는 본문에서 출현한 단어보다 N배하여 단어의 가중치 값을 무겁게 설정한다. 여기서 파라미터로 사용되는 N의 값이 너무 낮게 설정되면 "Subject"에서 출현하는 내용어(명사)를 활성화하기 쉽지 않고, N의 값을 너무 높게 하면 본문에서 출현하는 명사가 무시되는 경향이 있다. 따라서 K는 경험적인 파라미터 값으로 수차례의 예비실험을 토대로 설정한다.

메일 문서의 처리에서 다중 속성값은 위에서 언급한 방법과 우선순위의 정도에 따라 설정되며, 이 두 가지 정보가 포함된 정보를 다중 속성값의 집합(MaVS; Multi-attribute Value Set)이라 정의한다. MaVS는 수신자가 수신하는 모든 메일 문서에 대해 생성되고, 모든 메일에서 추출된 MaVS는 결과 파일에 합산된다. 이러한 결과 파일을 "사용자 프로파일(user profile)"이라 한다.

앞의 (그림 1)에 점선으로 나타낸 흐름은 분석 모듈이다. 중요도 값이 아직 계산되지 않은 채 입력된 메일은 형태소 분석하고, 배경 지식 사전을 참조하여 다음의 두 가지 속성값인 "Type"과 "Time Limit"을 추출한다. "Sender"는 주소 파일을 참고하여 적당한 코드로 숫자화 한다. "Subject"와 본문에서 출현하는 명사를 이용하여 단어 벡터를 생성하고, 적당한 "Theme"을 결정한다. 새로 생성된 단어 벡터는 그룹 파일에 존재하는 단어 벡터와 비교되며 새로운 문서의 "Theme"의 값이 추출되면 가장 유사한 폴더의 코드로 설정된다. 이상과 같은 방법으로 다중 속성값의 집합이 생성된다.

이러한 다중 속성값은 우선순위가 설정되어 있지 않은 값이다. 이 우선순위는 개인별 사용자 프로파일에서 얻을 수 있다. 새로 도착된 메일의 중요도 값은 사용자가 설정한 우선순위(프로파일에서 얻은 다중 속성값의 집합에 부착된 정보)를 기반으로 확률적으로 계산된다. 자세한 계산 방법은 뒤의 3.3절에 제시하였다.

3.2 학습 모듈

3.2.1 지식 사전

문장 유형과 시간제한 어구에서 속성값을 얻기 위해 많은 수의 한국어 표현 패턴을 배경 지식으로 전자사전에 등록하여야 한다. 여러 개의 서로 다른 형태의 표현 패턴이 하나 혹은 둘 이상의 문장 유형으로 분류되지만 최종적으로는 한 개의 속성값으로 결정된다. 한국어 문서의 처리에 적합한 지식 사전의 예를 <표 1>에 제시하였다.

시간제한을 나타내는 표현의 경우 추가적으로 고려하여야 할 점이 있다. 시간을 나타내는 區間(구간)의 정도에 따라 속성값을 구별하는 임계값을 별도로 설정하여야 한다. 특히 한국어에 풍부한 부사의 경우는 구간을 구분하기에 다소 복잡한데, 사람의 인지능력을 이용하여 판단하기 위해 대학(원)생 10명에게 설문 조사를 실시하여 각각의 시간 부사가 얼마나 많은 양 혹은 얼마나 긴 정도의 시간적인 구간(Time Interval)을 가지는가를 조사하여 구분하였다. 이 결과에 의해 한국어의 시간 부사를 크게 다음의 두 가지 유형으로 분류하였다.

(A) 시간의 구간을 포함하는 시간 부사,

(B) 구간을 표시하지 않고 시간상의 점, 어느 때, 특정한 순간 등을 지칭하고 있는 시간 부사 등

예를 들어, "한 시간 이내에"라는 시간 표현은 위의 그룹(A)에 속한다. 왜냐하면 "한 시간"이라는 표현 어구가 시간의 언제부터 언제까지 라는 구간(동안)을 명확하게 지시하고 있기 때문이다. 이러한 표현 어구는 속성값을 시간 구간으로 대응시킨다. 이와 반면에, "내일까지"와 같은 어구는 위의 그룹 (B)에 해당한다. 왜냐하면 이 표현은 "내일"이라는 시간상의 점을 지시하기 때문이다. 이상과 같은 시간 표현 어구들은 현재 시간으로부터 시간상의 점까지의 거리를 구간으로 간주한다. 그러나 이 방법으로 처리하기 곤란한 경우도 있다. 예를 들어, "다음 세미나 시간까지"와 같은 표현의 경우에는 사건의 스케줄을 포함하고 있는 어구들의 처리가 불가능하다. 왜냐하면 시간 구간을 "세미나"라고 하는 사건에 대한 스케줄 정보 없이 특별하게 지칭하기가 불가능하기 때문이다. 따라서 본 논문에서는 "세미나"와 같이 특정 명사의 시간 정보에 대한 별도의 처리는 고려하지 않았다.

3.2.2 사용자 프로파일

사용자 프로파일의 자동 구축을 위해 학습 데이터를 집합 $D = \{d_1, d_2, ..., d_i, ... d_m\}$ 라 정의하면 $d_i = \{x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{in}\}$으로 표현된다. 여기서, D는 문서 집합 전체, d_i는 문서 집합내의 개별 문서, x_{ij}는 다중 속성값의 집합을 의미하며, 다음의 다차원 벡터로 표현할 수 있다.

$$x_{ij} = (\alpha_l, \beta_m, \gamma_n, \theta_q, \kappa_r) \quad \text{..... (1)}$$

이와 같은 다중 속성 집합을 구축하는 알고리즘을 다음에 설명하였다.

〈표 1〉 구축된 지식 사전의 예

종류	속성값	표현 어구의 예
문장 유형 (β)	공 지	계획하고 있습니다, 소개합니다, 안내입니다, 알려 드립니다, 알립니다, 공고, 공지, 모임, 소개, 소식, 안내, 초대, 통지, 행사, 정보, ...
	명 령	가거라, 가라, 가지마라, 거라, 꺼라, 닫아라, 마라, 말라, 말아라, 믿어라, 버려라, 보내라, 보아라, 봐라, 오너라, 와라, 보세요, 의무적으로, 하십시오, 오십시오, 이십시오, 줄래, 하십시오, 하세요, 하여라, 하자, 해야 합니다, ...
	요 청	꼭 부탁드립니다, 말아라, 알려주시기 바랍니다, 하여라, 해라, 해줄 것, 해줘, 도와주시겠어요, 바랍니다, 해주세요, 부탁합니다, 할 수 없겠니, 할 일, ...
	의 문	궁금합니다, 말까, 모르겠어요, 알고 싶다, 알려 주세요, 올까, 울까, 일까, 줄까, 가시겠어요, 가시겠어요, 습니까, 어요, 없습니까, 오셨습니까, 입니까, 있습니까, 했냐, 해줄래, ...
	청 유	갑시다, 합시다, 봅시다, 십시다, 하지 맙시다, 바랍니다, 가자, 물어보자, 잡으세요, 하자, 바란다, 보세요, 않겠니, 않을래, 으면 좋겠다, 좋겠습니다, 주지 않겠니, 하는 게 어떠니, 하는 게 어때요, 하십시오, 하지 않겠습니까, ...
시간 제한 (γ)	24시간 이내	1시간 이내, 1시간 후에, 24시간 이내, 오늘 중에, 지금 당장, 내일까지, 되도록 빨리, 신속하게 처리, 곧바로, 급하게, 빨리, 서둘러서, 시급, 신속하게, ...
	3일 이내	가능한 빨리, 잊지 않는 동안에, 2-3일 중에, 모레까지, ...
	1주일 이내	1주일 이내에, 7일 이내, 이번 주까지, 이번 주말까지, ...
	1주일 이상	2-3 주일 이내, 다음 주까지, 1개월 이내, 1개월 후, 다음달, ...

■ 다중 속성 집합의 구성: 알고리즘

(단계 1) α_l(송신처 속성)을 새로운 다중 속성 값으로 설정한다.

메일 번호 (ID)	(sender, (송신처<α>,	type, 문장 유형<β>,	time limit, 시간제한<γ>,	subject, 주제<θ>,	priority) 우선순위<κ>)	= =	frequency 빈도수	빈도수의 누적 합계 (문장 유형 기준)
1	(홍길동교수,	충고,	12시간,	모임,	3)	=	2	2
2	(홍길동교수,	충고,	24시간,	모임,	4)	=	3	5
3	(홍길동교수,	충고,	12시간,	세미나,	5)	=	* 7	12
4	(홍길동교수,	충고,	*,	세미나,	5)	=	* 3	15
5	(홍길동교수,	충고,	*,	시험,	5)	=	* 2	17
6	(홍길동교수,	충고,	3일,	행사,	5)	=	* 3	* 20
7	(홍길동교수,	요구,	24시간,	세미나,	4)	=	3	3
8	(홍길동교수,	요구,	1주일,	시험,	2)	=	2	5
9	(홍길동교수,	요구,	*,	행사,	1)	=	3	8

(그림 2) 사용자 프로파일의 예

· set (α_l = 'new');

(단계 2) 만약 "Type"에 해당하는 속성값이 추출되면, 추출된 개수만큼 다중 속성값을 중복하여 β_m(문장 유형의 속성)에 저장한다.

· if (β_m = new type)
then for (i = 1 ; i = n ; I++)
store(β_m);

(단계 3) 위의 (단계 2)에서 β_m이 추출되면, 각 문장에서 γ_n("Time Limit")을 추출한다. 대응하는 다중 속성값에 γ_n(시간제한의 속성)을 설정한다.

· extract γ_n for each sentence;

(단계 4) 만약 θ_q("Theme")가 추출되고, κ_r(사용자에 의한 우선 순위값)이 설정되어 있으면 각각을 다중 속성값으로 설정한다.

· set (θ_q='theme') && (κ_r='priority');

모든 문서에 대해 생성된 다중 속성값 집합의 수에 의해 빈도수 $freq(x_{ij})$를 계산한다. 여기서, $freq(x_{ij}) = \frac{1}{n} (1 \le j \le n)$이고, $\sum_j freq(x_{ij}) = 1$이다. 여기서, $\cap_{ij} \neq \varnothing$이므로 동일한 다중 속성 집합이 추가되어, 프로파일 P는 다음의 식 (2)와 같이 구할 수 있다.

$$P = \{p_1, p_2, p_3, \cdots, p_k, \cdots p_z\} \tag{2}$$

p_k는 다중 속성값을 나타내고, $\cap p_k = \varnothing$이다. $freq(p_k)$는 같은 속성 집합의 전체 빈도수이다. 이 프로파일 p_k는 사용자가 여러 가지 속성값의 조합에 의해 어느 정도 중요성을 느끼고 있는가를 결정한다. 어떤 다중 속성 집합이 $(\alpha_l, \beta_m, \gamma_n, \theta_q, \kappa_r)$이면 이후에는 간단히 $p_{<l,m,n,q,r>}$로 표기한다.

3.3 분석 모듈

입력 문서 집합을 $T = \{t_1, t_2, t_3, ..., t_i, ..., t_m\}$라 하면 각 문서는 $t_i = \{y_{i1}, y_{i2}, y_{i3}, \cdots, y_{ij}, ..., y_{in}\}$로 표현된다. 여기서, y_{ij}는 다중 속성 집합이다. 즉, 우선순위 κ를 갖지 않은 다중 속성 집합의 벡터라 할 수 있다. 분석 모듈에서는 모든 y_{ij}에 대해 문서의 중요도가 계산된다. 프로파일 내에 동일한 y_{ij}가 존재하면 정상적인 처리(General Process)가 수행되지만, 존재하지 않으면 근사처리(Approximate Process)가 수행된다.

정상 처리에서는 $y_{ij} = (\alpha_l, \beta_m, \gamma_n, \theta_q)$에 대한 중요도 값은 다음의 식 (3)에 의해 계산된다.

$$y_{ij} = (\alpha_l, \beta_m, \gamma_n, \theta_q) = \frac{freq(p_{<l,m,n,q,r>})}{freq(p_{<l,m,n,q>})} \tag{3}$$

여기서,

$$freq(p_{<l,m,n,q>}) = \sum_r freq(p_{<l,m,n,q,r>}) \tag{4}$$

이다. 완전한 속성값이 추출되지 않는 경우에는 존재하는 속성값 만을 위의 식 (3)에 의해 추출한다. 예를 들면, $y_{ij} = (\alpha_l, \beta_m, *, \theta_q)$[2], 위의 식 (3)은 다음의 식 (5)와 같이 수정하여 자동 계산한다.

$$P(\kappa_r | \alpha_l, \beta_m, \theta_q) = \frac{freq(P_{<l,m,q,r>})}{freq(P_{<l,m,q>})} \tag{5}$$

여기서,

$$freq(P_{<l,m,q,r>}) = \sum_n freq(P_{<l,m,n,q,r>}) \tag{6}$$

식 (3)에 의해 얻어진 값은 우선순위 κ에 의해 계산한 각 순위별의 확률값이다. 따라서 다음의 식 (7)과 같이 수정한다.

$$I_{ij} = \sum_r r \times P(\kappa_r | \alpha_l, \beta_m, \gamma_n, \theta_q) \tag{7}$$

마지막으로, 앞의 3.3절에서 서술한 바와 같이 어떤 입력 문서의 중요도(Importance) I_i는 $I_{i1} \sim I_{if}$의 평균 중요도를 계산하여 얻을 수 있다.

앞의 (그림 2)에서 제시한 사용자 프로파일의 예를 이용하여 분석 작업의 예를 설명한다. 우선, 프로파일 내의

2) 심볼 * 은 문서에 대응되는 속성값이 존재하지 않는 경우를 나타낸다.

y_{ij} = (홍길동교수, 충고, *, 세미나)에서 메일 번호(그림에서 ID로 표시) 3과 4의 다중 속성값의 집합이 추출된다. ID 3과 4 각각의 빈도수의 합이 10 이고, 이 두 ID의 우선순위가 5 이므로 순위 5의 확률이 100%이고, $I_i = 5$이다. 다음으로 y_{ij} = (홍길동교수, 충고, *, *)의 경우, "Theme"의 정보는 얻을 수 없다. 따라서 모든 다중 속성 집합의 모든 ID 1에서 6까지가 참조된다. 이들 여섯 개의 ID의 빈도수의 누적 합이 20 이고, ID 3, 4, 5, 6의 네 개의 우선순위가 5의 우선순위 합이 15(=7+3+2+3)이므로, 순위 5의 확률은 75%(=$\frac{15}{20}\times 100$)이다. 이와 같은 방식으로 순위 4는 20%(=$\frac{4}{20}\times 100\%$), 순위 3은 10%(=$\frac{2}{20}\times 100\%$)이고, 이 때 $I_{ij} = 4.65(=3.75+0.60+0.30)$이다. 왜냐하면 우선순위 5의 중요도 $I_{ij} = 3.75(=5\times 0.75)$, 우선순위 4의 중요도 $I_{ij} = 0.60(=4\times 0.20)$, 우선순위 3의 중요도 $I_{ij} = 0.30(=3\times 0.10)$ 등이다.

다중 속성 집합이 입력 문서로부터 출력되었으나, 프로파일에서는 발견되지 않으면 프로파일 내에 존재하는 다른 다중 속성 집합의 값을 대치하여 근사적인 중요도를 계산하여야 한다. 이러한 대치 작업은 시스템의 성능을 결정하는 매우 중요한 작업이며, 반드시 필요한 작업이다. 각 속성값들의 특징을 고려하여 이 대치 작업을 다음의 순서로 이루어지도록 시스템을 설계하였다.

- $\gamma(Time) \rightarrow \alpha(Sender) \rightarrow \theta(Theme)$

"Type"에 대한 값들 사이에 유사한 정도의 차이가 매우 작기 때문에 위의 대치 작업에 β(Type, 문장 유형)의 값은 사용하지 않았다.

4. 실 험

본 논문에서 제시하는 방법의 효과를 검증하기 위해 전통적인 방법(빈도수만의 속성값을 이용하는 방법)과의 비교 실험을 수행한다. 우선, 이 실험에서 사용된 데이터를 소개하고 실험 결과를 제시한다.

실험 데이터는 네 개의 주제로 구분된 200통의 메일 문서로 구성하였다. 학습용 데이터는 앞의 네 주제로 구분된 실험용 데이터에서 무작위로 30통을 선정하고, 나머지는 학습용 데이터로 이용하였다. 학습용 데이터의 구성을 <표 2>에 설명하였다. "문장 유형"에 대한 282개의 표현 패턴(속성값의 수가 8)과 "시간제한"에 대한 48개의 표현 패턴(속성값의 수가 5)을 지식 사전으로 구성하였다. 우선순위의 수는 5(가장 높은 우선순위)이고, 파라미터 K는 10으로 설정하였다. 프로그램을 구현할 때 사용한 컴퓨터는 Intel Quad Core 6750 (메모리 DDR2 RAM PC 6400 2G)이다.

평가 데이터로부터 얻은 각 속성값의 정확률을 위의 <표 3>에 제시하였다. 이 표에서 "문장 유형"에 대한 높은 정확률을 얻었으나 "시간제한"에 대한 재현율은 다소 낮게 나타났다. 그 이유는 평가용 데이터에 시간상의 점을 나타내는 언어적 표현 어구와 시간을 나타내는 명사가 매우 많이 존재하기 때문이다. "제목"의 경우, 학습 데이터에 유사도의 값이 존재할 때는 적당한 값으로 결정할 수 없다. 아래의 <표 4>에 본 논문에서 제안하는 방법(제목과 같이 단일 속성을 이용한 경우)과 Rocchio 알고리즘[1,3], Widrow-Hoff 방법[2]의 비교 결과를 제시하였다. 이 표에서 사용자에 의한 우선순위와 각 방법의 중요도 사이에서의 상관 계수를 의미한다. 이 상관 계수는 10 Cross Validation에 의한 평균값이다. 이 표에서 보는 바와 같이 전통적인 방법에 비해 본 논문의 방법이 더 높은 상관계수를 보여 더 효과적인 방법임을 알게 되었다.

메일의 중요도를 계산하는데 각 속성들 사이의 어떠한 조합이 효과적인가를 평가하기 위해 속성의 수와 상관 계수를 나타낸 것을 <표 5>에 설명하였다. 각 속성값의 정확률이 100%가 되지 않으면 정확한 결과를 평가할 수 없기 때문에 정확률이 다소 낮은 "시간제한"에 대한 속성값은 이번 실험에서는 취급하지 않았다. 평가용 데이터에서 "제목"의 오른쪽 태그는 인간이 100%가 되도록 부착하였다. 이 표에서 "제목"과 함께 조합한 정확률이 좋은 성능을 보이고, 모든 속성값을 고려한 정확률이 가장 좋은 성능을 보인다.

각 처리 단계의 실행 시간을 측정하기 위해 학습 데이터

〈표 2〉 학습용 데이터의 내역

실험자		A	B	C	D
학습 문서의 수		250	230	150	120
총 바이트 수 (K Byte)		655	501	233	152
송신처 수		59	57	7	28
폴더 수		7	12	6	5
다중 속성의 수		310	291	203	161
우선순위	평균	3.3	3.14	3.63	3.46
	분산	1.27	1.31	2.11	2.07

〈표 3〉 학습용 데이터에서 얻은 각 속성값의 정확률

실험자	제목(%)		문장 유형(%)		시간제한(%)	
	재현율	정확률	재현율	정확률	재현율	정확률
A	66.7	62.5	100	82.5	61.1	100
B	73.3	71.0	95	80.6	47.5	
C	80.0	80.0	100	81.7	44.4	
D	76.7	76.7	94.7	82.2	50.0	
평균	**74.2**	**72.6**	**97.4**	**81.8**	**50.8**	**100**

〈표 4〉 각 방법에 의한 상관계수

실험자	Rocchio	Widrow-Hoff	본 논문의 방법
A	0.58	0.60	0.58
B	0.42	0.46	0.57
C	0.32	0.34	0.53
D	0.29	0.30	0.38
평균	**0.40**	**0.43**	**0.52**

〈표 5〉 각 다중 속성값과 〈표 4〉의 상관계수와의 관계

속성별 / 실험자	단 일 속 성			이 중 속 성			삼 중 속 성
	송신처	유형	제목	송신처+유형	송신처+제목	송신처+제목	송신처+유형+제목
A	0.58	0.48	0.51	0.55	0.57	0.67	0.68
B	0.49	0.29	0.54	0.52	0.55	0.63	0.65
C	0.42	0.27	0.58	0.56	0.60	0.66	0.73
D	0.56	0.14	0.44	0.66	0.21	0.53	0.57
평균	0.51	0.30	0.52	0.57	0.48	0.62	0.66

의 평균 파일 크기가 385 KB이고, 평균 학습 시간이 11.5초 이었다. 평균 분석 시간은 1.2초이었다. 두 가지 데이터 모두 형태소 사전을 주 메모리에 저장하고 있는 경우 형태소 분석에 소요되는 시간이 포함되어 있다. 프로파일의 평균 크기는 4.6 KB이다.

5. 결 론

본 논문에서는 사용자의 프로파일에 기반하여 입력되는 메일을 필터링하는 방법을 제안하였다. 이러한 프로파일은 과거에 사용자가 수신한 메일에서 추출한 다중 속성값으로 구성된 정보이다. 향후에는 메일 문서에서 추출 가능한 다른 형태의 속성값을 본 논문에서 지시하는 시스템에 적용하고자 한다. 예를 들면 사용자의 관심거리를 추적하는 새로운 형태의 정보를 속성값으로 이용하고자 한다.

참 고 문 헌

[1] Buckley C. Salton G. and Allan J., "The Effect of Adding Relevance Information in a Relevance Feedback Environment," Proceeding of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.292-298, 1994.

[2] Lewis D. D., Schapire R. E., Allan J. P. and Papka R., "Training Algorithms for Linear Text Classifiers," Proceeding of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.298-307, 1996.

[3] Rocchio, J. J., "Relevance Feedback in Information Retrieval," The SMART Retrieval System – Experiments in Automatic Document Processing, Salton, G. (Ed.), Prentice Hall, pp.313-323, 1971.

[4] 이지행, 조성배, "전자 우편 문서의 자동 분류를 위한 다중 분류기 결합", 정보과학회 논문지: 소프트웨어 및 응용, 제29권, 제3호, pp.192-201, 2002.

[5] 강영순, 이용배, 김태현, 조숙현, 맹성현, "전자 우편 문서의 효율적인 분류를 위한 전처리", 한국정보과학회 학술발표 논문집(II), 제29권, 제1호, pp.493-495, 2002.

[6] 안희국, 노희영, "동적 시소러스와 GA를 이용한 개별화된 E-Mail 분류 시스템(PECS)", 한국정보과학회 학술발표 논문집(II), 제29권, 제1호, pp.472-474, 2002.

[7] 최승혁, 김용성, 김영천, "지능형 E-mail 문서 관리기 시스템 설계", 한국정보과학회 학술발표 논문집(II), 제29권, 제2호, pp.307-309, 2002.

[8] 박시일, 김두현, 김용성, "지능형 E-mail 지식 관리 시스템 설계", 한국정보과학회 학술발표 논문집(II), 제29권, 제2호, pp.310-312, 2002.

[9] 안희국, 노희영, "유전자 알고리즘을 이용한 전자 메일 분류 시스템의 사용자 선호도 추출 모델링", 한국정보과학회 학술발표 논문집(II), 제29권, 제2호, pp.673-675, 2002.

[10] 류제, 윤성희, 한광록, "특정 속성과 Co-training을 이용한 전자 메일 분류", 한국정보과학회 봄 학술발표 논문집(B), 제30권, 제1호, pp.549-554, 2003.

[11] 현영순, 정옥란, 조동섭, "E-Mail 시스템의 첨부 파일 자동 분류 에이전트 설계", 제19회 한국 정보처리학회 춘계 학술대회 논문집, 제10권, 제1호, pp.1067-1070, 2003.

[12] 권용진, 안준선 역, "정보검색 알고리즘", 도서출판 미래컴, pp.80-86, 2003.

[13] 안찬민, 박선, 김태순, 최범기, 이주홍, "문서 요약 및 동적 분류 체계를 사용한 E-mail 분류의 재구성", 제21회 정보처리학회 춘계 학술발표 대회 논문집, 제11권, 제1호, pp.511-514, 2004.

[14] 변영철, 홍영보, "신경망을 이용한 사용자 질의 전자 메일 분류", 멀티미디어학회논문지, 제7권, 제3호, pp.438-449, 2004.

[15] 김보미, 이상열, 이상곤, "이메일 문서의 속성값에 기반한 필터링 시스템의 설계 및 구현", 한국 컴퓨터 종합 학술대회 2005 논문집, 제32권, 제1(B)호, pp.142-144, 2005.

[16] 장정효, 이상열, 이상곤, 조현준, "한국어 문서의 유형 정보를 이용한 EMFA의 구현", 한국 컴퓨터 종합 학술대회(KCC) 논문집, 제33권, 제1호(B), pp.28-30, 2006.

이 상 곤

e-mail : samuel@jj.ac.kr

1994년 전주대학교 영어영문학과(학사)

1996년 전북대학교 컴퓨터과학과(학사)

1998년 전북대학교 전산통계학과 (이학석사)

2001년 일본 국립 도쿠시마대학교 지능정보공학과(공학박사)

2001년~2002년 원광대학교 음성정보 기술산업 지원센터 연구원

2002년~현 재 전주대학교 공과대학 컴퓨터공학과 부교수

관심분야 : 한국어 정보처리, 한글공학, 자연언어처리, 정보검색, 문서분류 및 자동 요약, 컴퓨터교육, 임베디드 소프트웨어