이메일문서의 속성값에 기반한 필터링 시스템의 설계 및 구현

김보미[1], 이상열[2], 이상곤[2]
전주대학교 교육대학원 컴퓨터교육전공[1]
전주대학교 정보기술공학부 컴퓨터공학 전공 언어과학실[2]
{springtwo[1], pcguy7[2], samuel[2]}@jj.ac.kr

Design and Implementation of E-Mail Filtering System based on Attribute Values

Bo-Mi Kim[1], Sang Yeol Lee[2] and Samuel Sangkon Lee[2]
Dept. of Computer Education, Graduate School of Education[1],
Language Science Lab., Dept. of Computer Science & Engineering,
School of Information Technology & Engineering,[2] Jeonju University[1,2]

요 약

전자메일이 일상의 연락수단일 뿐만 아니라 여러 가지 목적의 업무처리에 있어서도 중요한 통신수단으로 이용되고 있다. 이에 따라 전자메일의 중요도를 자동적으로 판정하는 문서 필터링 방법이 연구되고 있다. 본 논문은 수신된 메일문서에서 송신처, 제목, 문서유형, 시간제한 등의 다중속성을 조합하여 구조적인 지식을 생성하여 전자메일을 자동으로 필터링하는 시스템을 구현한다.

1. 서론

우리생활에 인터넷이 점차 보급됨에 따라 전자메일이 일상의 연락수단으로 사용될 뿐만 아니라 여러 가지 목적의 업무처리에 중요한 통신수단으로 이용되고 있다. 이에 따라 보다 빠른 업무처리를 위해서는 중요도가 높은 메일을 먼저 처리할 수 있는 내용기반(contents-base) 정보 필터링 기술[1, 2]의 필요성이 높아지고 있는 실정이다. 따라서 본 논문에서는 수신된 메일문서에서 송신처, 제목, 문서유형, 시간제한 등의 다중속성을 조합하여 구조적 지식을 획득하고, 이를 필터링에 이용하는 방법을 제안하고자 한다. 이하, 2장에서는 메일문서의 필터링 방법에 대해 설명하고, 어떤 속성을 이용하여 속성값을 결정하는가에 대해 논의한다. 3장에서는 필터링 시스템의 구현 결과를 설명한다. 마지막으로 4장에서는 결론과 향후 연구과제에 대해 논의한다.

2. 필터링 방법

개인이 수신한 메일문서에는 중요도를 판정하기 위한 이력정보가 포함되어 있다고 가정하고, 각 개인이 우선순위를 부여한 기존의 메일 문서로부터 사용자 프로파일(user profile)을 작성한 후, 그 프로파일을 이용하여 필터링을 수행한다. 저장되어 있는 메일문서에 포함된 다중속성값[3]과 사용자가 설정한 우선도를 조합한 속성을 각 메일마다 생성하고, 이미 학습된 문서 시스템이 계산한 중요도를 이용하고 출현단어의 빈도를 집계한 결과를 이용하여 필터링하게 된다.

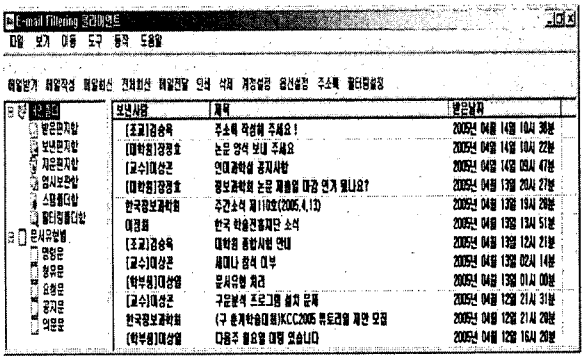
형태소해석에서 얻어진 정보를 기반으로 메일문서의 송신처, 메일 문서에 포함된 문장의 유형, 시간표현 어구의 유무, 메일 문서의 제목에 출현하는 주요명사 등의 4가지 항목을 '속성'이라고 하고, 그 속성들의 값을 '속성값'이라 한다[3]. 문서유형과 시간제한 표현의 속성값은 표현의 패턴수가 유한하며, 대체로 개인차가 적기 때문에 표현패턴을 등록한 배경지식을 미리 구축하여 두고 배경지식을 이용하여 속성값을 검출한다. 배경지식의 일부를 <표 1>에 나타내었다.

<표 1> 속성값의 예

속성	속성값의 이름	표현패턴의 예
문서의 유형	명령	~하라, ~하자, ~할 것, ~하세요, ~해
	의문	~입니까?, ~했냐?, ~않니?, ~까?, ~냐?, ~니?
	청유	~하자, ~할까, ~합시다, ~하지요, ~하는게 어때요, ~으면 좋겠다, ~하지 않겠습니까
	요청	~하여라, ~없겠니, ~하자, ~바란다, ~라, ~니, ~자,
	공지	~알려드립니다, ~알립니다, ~소개합니다, ~안내입니다, ~계획하고 있습니다.
시간제한 표현	24시간 이내	곧바로, 급하게, 빨리, 서둘러서, 가능한 빨리, 신속하게, 오늘 중에, 즉시,
	3일이내	잊지 않는 동안에, 2~3일 중에
	1주일이내	이번주까지, 1주일이내, 7일 이내
	1주일 이상	2~3 주일 이내, 1개월 이내, 1개월 후

3. 메일클라이언트의 설계

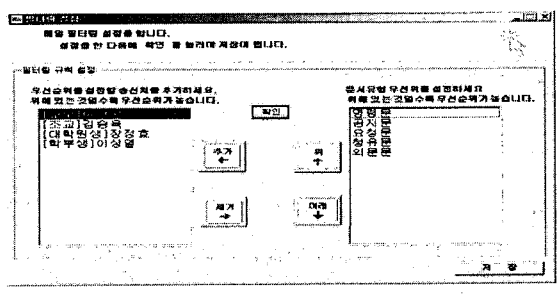
앞장에서 논의된 개념을 기반으로 실제 전자메일 클라이언트에서 필터링되는 동작을 설계하고 구현한다[6, 7]. 문서유형은 243개(속성값의 수는 5개), 시간제한을 나타내는 표현어구은 33개(속성값의 수는 4개)의 표현패턴을 미리 준비하였다. 이들 데이터를 이용하여 사용자 프로파일을 작성했다.



(그림 1) 클라이언트의 받은 편지함의 예

설계된 시스템의 구조를 설명하기 위해 3장의 마지막에 기술한 (그림 4)에 (가)~(라)까지의 순서를 표시하였다. 그림의 (가)로 표시한 부분은 다음과 같다. 전자메일은 형식 혹은 표현방식에 제한이 없으므로, 특정한 기호 삽입이나 띄어쓰기 오류 등이 내재되어 있다. 따라서 시스템의 오류를 줄이기 위해 메일문서의 내용에 포함된 공백을 모두 제거한다. 전자메일 문서에 HTML의 태그 형식이 일반 내용과 함께 색인될 경우 불필요한 색인어들이 점수에 산출되므로, 전자우편문서에 포함된 태그들도 모두 제거한다.

(나)에서 사용자 우선순위 결정의 우선도는 특정 송신처에서 보낸 메일의 주소에 우선순위를 높일 것인지, 제목이나 내용에 포함된 단어나 문장을 검사해서 우선순위를 높일 것인지 등의 여러 단계로 나누어 결정된다. 필터링 환경설정은 각 사용자의 업무 상황에 맞게 송신처를 설정할 수 있다. 또한 각 송신처에 맞게 각자 다른 문장의 유형의 선택이 가능하고, 문서유형에서 선택된 우선순위로 필터링 할 수 있다.

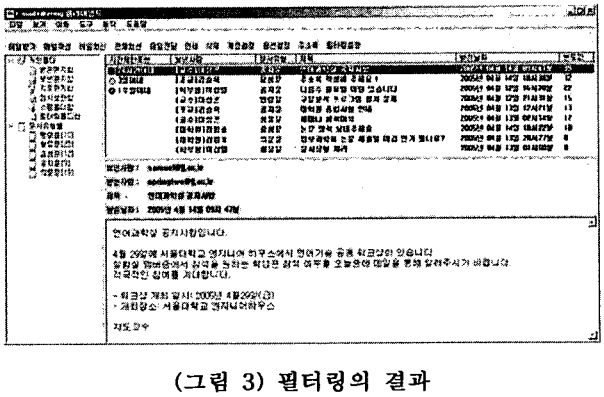


(그림 2) 메일필터링의 우선도 설정 예

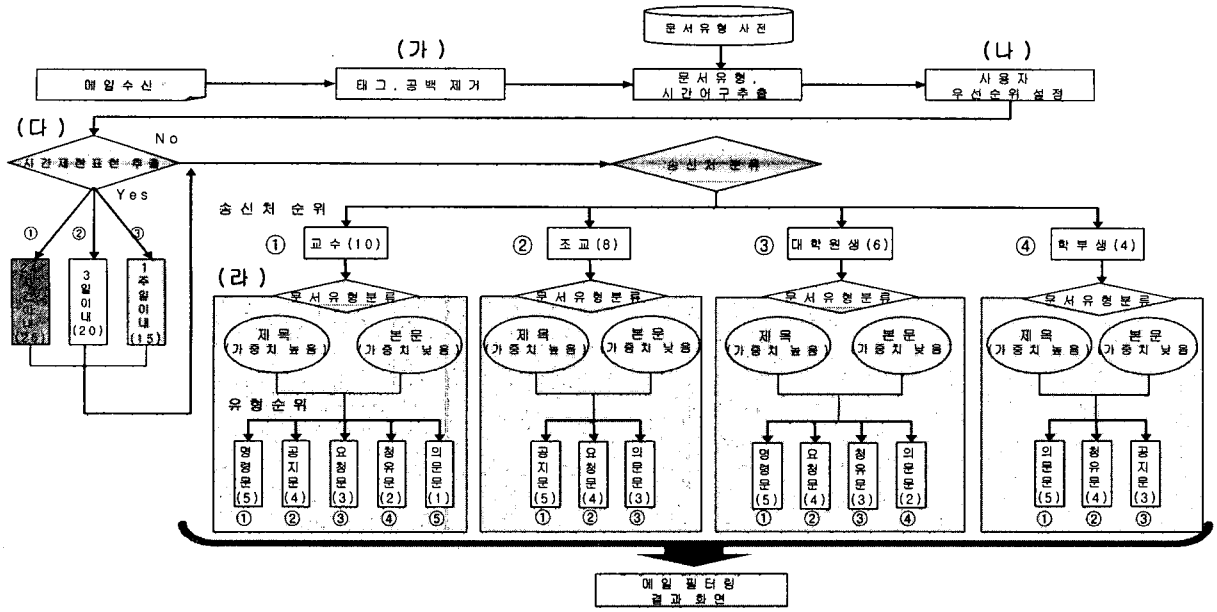
(그림 2)은 사용자에 맞게 사용자가 설정한 송신처와 문서유형의 우선도를 설정하는 화면이다. 그림의 [지도교수]에게서 온 메일의 문서유형 우선도를 ①명령문, ②청유문, ③공지문 등의 순으로 설정을 하였고, 송신처가 [조교]인 메일은 ①공지문, ②요청문, ③의문문의 순서로 서로 다른 우선도를 설정하였다. 다음과 같이 우선도를 설정한 메일을 필터링한 결과는 (그림 3)과 같이 구현하였다.

(다)는 시간제한 표현어구의 처리과정을 설명하였다. 복수개의 표현이 나타나는 경우에는 마지막에 기술된 시간표현 어구를 중심으로 처리한다[3]. 예를 들어, 본문에 "가능한 한 빨리 보내주세요." 혹은 "될 수 있으면 1시간 이내에 해 주시길 부탁드립니다"란 어구가 출현하는 메일은 가장 최후에 발견된 시간표현어구인 "1시간 이내"로 시간제한 표현을 이용하여 결정한다. 시간의 구간을 포함하지 않는 시간표현은 송신시간을 고려하여 결정하였다. 부가기능으로는 시간표현 어구가 포함된 메일에 처리시간의 긴박한 정도를 시간적으로 나타내는 신호등 표시기를 이용하여 사용자에게 보여줌으로써 문서처리의 긴박도를 나타낸다.

(라)는 문서의 유형을 파악하기 위한 처리인데, 이를 위해서는 한국어의 문서유형별로 잘 분류된 사전[4]이 필요하다. 메일의 제목 혹은 본문에 문서의 유형에 해당하는 표현어구를 추출하여 이에 가중치를 부여한 순서대로 매칭 한다. 추출단어가 출현하면 문서의 유형을 지시하는 카운트 값을 증가시킨다. 표현패턴 마다 각기 다른 빈도값을 부여하여 문서유형의 중요도를 판정한다. 예를 들어, 명령문의 경우에는 요청성의 정도를 다음과 같이 네 가지로 나누어 정의하였다. 예시하면 ①강압적 요구, ②온건한 요구, ③정중한 요청, ④희망 등이다. 이러한 분류는 요청성의 강도나 요청하는 정도를 나누어 판단하는 재료가 된다. 제목은 본문의 표제이므로 제목에 나타난 단어들의 변별력을 높이기 위해 가중치를 높여주고, 표현패턴의 출현여부 혹은 출현위치가 문서의 앞부분인지 또는 뒷부분인지에 따라 가중치 값을 다르게 설정하여 문서의 문형을 분류한다.



(그림 3) 필터링의 결과



(그림 4) 시스템의 전체 구조

(그림 3)의 필터링 결과는 (그림 4)에서 제시한 개념을 이용하여 (그림 2)의 수신된 메일을 사용자가 지정한 우선도의 속성을 기반으로 필터링하였을 때의 결과를 출력한 화면이다. 필터링 설정이 되어 있는 않은 송신처에서 수신된 메일은 필터링 결과를 화면에 출력하지 않으며, 시간제한 표현과 사용자가 설정한 우선도 설정을 기반으로 필터링 결과를 나타내게 된다. 입력문서의 다중속성과 동일한 다중속성이 프로파일 내에 존재하지 않는 경우에는 프로파일에 존재하는 다른 속성 값으로 치환하여 근사적으로 계산한다. 각 속성의 특징을 고려하여 치환을 적용하는 우선순위는 ①시간제한 → ②송신처 → ③제목의 순서를 적용하여 구현하였다.

시스템의 구현환경은 메모리 1G를 탑재한 CPU Pentium IV 2.4GHz 속도를 가진 시스템에서 마이크로소프트사의 Visual Studio.Net 2003을 이용하였고, 문서유형을 저장한 사전은 간단히 텍스트 파일로 구축하였다.

4. 결론

본 논문은 수신된 메일문서에서 다중속성 항목으로 구성된 프로파일을 작성하고, 입력되는 메일문서에서의 중요도를 계산하는 방법을 제안하였다. 본 논문의 방법을 이용하면 각 사용자가 중요성을 느끼는 송신처와 문서 유형을 포함한 문서를 우선적으로 처리할 수 있어서 신속하고 효율적인 업무처리를 기대할 수 있다. 향후의 연구과제는 실제 예제들을 중심으로 배경지식을 일반화하고 시간제한 표현의 포함여부를 자동으로 학습하여 긴급하게 처리할 업무는 사용자에게 충고하는 지능적인 메일클라이언트를 구축하고, 긴급한 처리가 필요한 이메일을 모바일 기계에 전송하도록 하는 시스템으로 확장할 계획이다.

참고문헌

[1] 강영순, 이용배, 김태현, 조숙현, 맹성현, "전자우편문서의 효율적인 분류를 위한 전처리", 한국정보과학회 학술발표 논문집(II), 제 29권, 제 1호, pp. 493-495, 2002.

[2] 박시일, 김두현, 김용성, "지능형 E-mail 지식관리시스템 설계", 한국정보과학회 학술발표 논문집(II), 제 29권, 제 2호, pp. 310-312, 2002.

[3] 김보미, 이원휘, 이상곤, "전자메일의 중요도에 기반한 이메일문서 필터링 방법", 한국정보처리학회 학술발표 논문집(상), 제 11권, 제 2호, pp. 811-814, 2004.

[4] 박영순, "한국어 문장의미론", 박이정, 2001.

[5] 강영순, 이용배, 김태현, 조숙현, 맹성현, "전자우편문서의 효율적인 분류를 위한 전처리", 한국정보과학회 학술발표 논문집(I), 제 29권, 제 1호, pp. 493-495, 2002.

[6] 채규혁 역, "인터넷 이메일 프로그래밍", 한빛미디어, 2000.

[7] 김태현, 박한돌 역, "Windows with C#", 정보문화사, 2002.

[8] Shishibori, M., Fujii, M., Ando, K., & Aoe, J., "Filtering Method for E-mail Documents Using Personal Profiles," Transactions of Information Processing Society of Japan, Vol. 41, No. 8, pp. 2299-2308, 2001. (in Japanese)