제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

전자메일의 중요도에 기반한 이메일문서 필터링 방법

김보미[1], 이원희[2], 이상곤[2]
전주대학교 교육대학원 컴퓨터교육전공[1]
전주대학교 정보기술공학부 언어과학실[2]
{springtwo[1], wony[2], samuel[2]}@jj.ac.kr

Filtering Method based on Importance of E-Mail Document

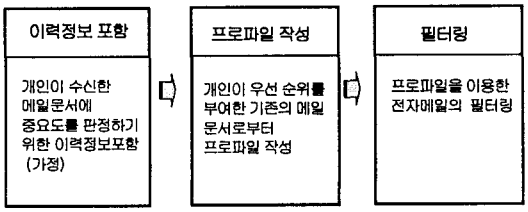
Bo-Mi Kim[1], Won-Hee Lee[2] and Samuel Sangkon Lee[2]
Dept. of Computer Education, Graduate School of Education[1],
Language Science Lab., School of Information Technology & Engineering[2],
Jeonju University[1,2]

요 약

최근 인터넷이 우리생활에 점차 보급됨에 따라 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 등의 다중속성의 조합으로 구성되는 구조적 지식을 획득하여 전자메일을 필터링 하는 방법을 제안한다.

1. 서론

우리생활에 인터넷이 점차 보급됨에 따라 전자메일이 일상의 연락수단으로 사용될 뿐만 아니라 여러 가지 목적의 업무처리에도 하나의 중요한 통신수단으로 이용되고 있다. 편리한 작성과 신속한 전달속도, 비용절감 등의 여러 이점 때문에 대부분의 서류가 전자메일을 통해 전달되고 있다. 또한 개인의 메일서버에는 대량의 메일문서가 수신되고, 중요한 업무메일과 함께 광고메일, 스팸메일이 점차 증가되어 빠른 업무처리에 지장이 있는 실정이다. 이에 따라, 중요도가 높은 메일을 먼저 처리할 수 있는 내용기반(contents-base) 정보 필터링 기술[5, 6]의 필요성이 높아지고 있는 실정이다. 본 논문에서는 (그림 1)과 같은 필터링 방법을 제안한다.



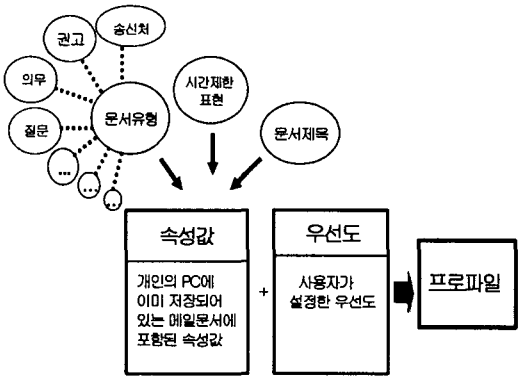
(그림 1) 내용기반 정보 필터링 기술

먼저, 개인이 수신한 메일문서에는 중요도를 판정하기 위한 이력정보가 포함되어 있다고 가정하고, 각 개인이 우선순위를 부여한 기존의 메일 문서로부터 프로파일(profile)을 작성한 후 그 프로파일을 이용하여 필터링을 수행하게 된다. 프로파일[1, 2]의 작성을 위해 (그림 2)에서 예시한 정보를 이용하여 메일의 중요도를 결정하는 요인으로 사용한다. 예를 들어, 송신처, 권고, 의무, 질문 등을 포함한 문서의 유형, 문서의 제목, 시간 제한 표현(이들 모두를 "속성"이라 정의) 등을 이용하여 개인의 PC에 미리 저장되어 있는 메일문서에 포함된 속성값과 사용자가 설정한 우선도를 서로 조합하여 프로파일을 작성한다.

이 프로파일은 사용자가 자신의 생각에 맞게 설정한 다중속성 값의 조합과 새로운 메일문서가 어느 정도의 중요성을 가지고 있는지를 나타낸다. 이 프로파일을 이용하면 각 사용자에게 알맞은 필터링이 가능하게 된다. 적은 규모의 학습데이터나 데이터의 내용이 스파스(sparse)한 경우는 유사한 속성값으로 치환한다. 따라서 본 논문에서는 각 속성의 의미에 따라 치환을 적용하는 순서를 정의하여 시스템을 설계한다.

이하, 2장에서는 메일문서의 중요도에 대해 설명하

고, 어떤 요인으로 중요도를 결정하는가에 대해 논의한다. 3장에서는 프로파일을 이용한 필터링 작성방법에 대해 설명한다. 4장에서는 결론과 향후과제를 서술한다.



(그림 2) 프로파일 작성순서

2. 메일문서의 중요도

메일문서의 내용이 다른 메일문서보다 빨리 읽을 필요성이 있는 내용이나 빠른 회신을 부탁하는 내용 등의 표현이 있는 문서를 중요도가 높은 메일문서라고 정의한다. 중요도는 문서의 유형이나 시간제한표현 어구의 존재 유무, 문서의 제목 등에 의해 판단된다. 또한, 일반문서에는 존재하지 않고 메일문서에만 존재하는 특정정보인 송신처(From), 참조(Cc), 제목(Subject), 과거에 수신한 메일과의 관련성 등을 중

① α : 메일문서의 송신처
② β : 메일문서에 포함된 문장의 유형
③ γ : 메일문서에 포함된 시간적 제한표현
④ θ : 메일문서의 제목

요도의 판단에 이용된다. 형태소 레벨에서 얻어진 정보는 필터링 작업에 유용하다고 판단되고, 다음과 같이 4가지 항목으로 나누어 생각해 볼 수 있다.

각 속성들의 상대적인 중요도를 결정하기 위해 위의 4가지 항목을 "속성"이라 하고, 각 속성들의 값을 "속성값"이라 정의한다. 각 속성 값인 메일문서의 송신처 α는 메일문서내의 From의 항목에서 추출하고, 메일문서에 포함된 문장유형 β는 메일문서의 본문에 포함된 서술어에서 얻는다. 메일문서에 포함된 시간적 제한표현 γ는 문서의 시간표현이나 시간을 나타내는 부사에서 추출한다. 마지막으로, 메일문서의 제목 θ는 사용자 주도의 방법(사용자가 이전에 도착한 메일의 중요도를 판정하여 저장한 방법)에 의해 제목을 결정한다. 이 정보는 제목이 되기 쉬운 명사류를 등록한 사전을 이용하여 얻을 수 있지만, 범용성을 상실할 수 있다. 또한 의미해석과 같은 고도의 해석기술을 이용하여 메일문서의 제목을 결정할 수 있으나 이것은 전자메일의 실시간처리에 어려움이 발생한다. 수신된 메일문서는 미리 사용자에 의해 유사한 내용의 문서에서 수집되어 메일박스의 각 폴더로 분류되어 있다고 전제한다. 각 문서가 저장되어 있는 폴더명을 그 문서의 메일문서의 제목으로 결정한다. 새로 수신된 메일문서에 대해서는 문서분류법과 같은 방법(3.1절에서 설명)으로 각 폴더와의 유사도를 계산하여 가장 유사한 폴더의 이름을 제목으로 선정한다.

2.1 중요도의 개인차

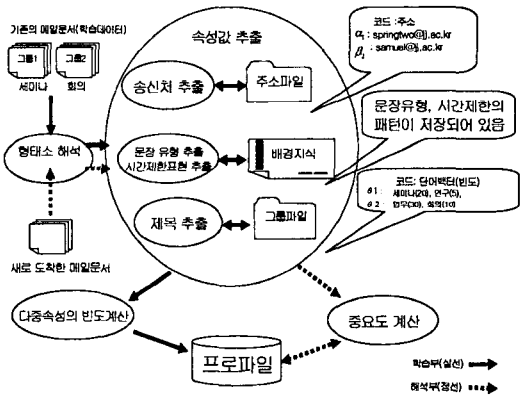
우선 메일문서의 중요성을 판단하는 기준이 개인마다 다르다는 점에 주목해야 한다. 예를 들면, 학생 A는 [이상곤교수]로부터 수신된 메일에 높은 중요도를 두지만, 학생 B는 [이상곤교수]의 메일에 중요성을 두지 않을 수도 있다. 이와 같이 중요한 속성값은 개인별로 다르므로 별도의 지식이 필요하다. 또한, 문서유형을 고려하여 보면, 학생 A는 [이상곤교수]에게서 온 [권고]의 메일문서에는 높은 중요도를 느끼는 반면에 [의뢰]의 내용을 포함하는 것은 중요도가 낮다고 생각할 수도 있다. 이와 반면에 학생 A는 반대의 중요도를 두고 있을 수 있다. 이것은 각 개인별로 속성값의 중요도와 그 조합은 다르게 설정되어 하기 때문이다. 따라서, 중요도에 따른 판정방법을 이용하여 속성 값들끼리 조합할 필요가 있다. 이와 같이 메일문서의 중요도는 각 속성 값에 따라 혹은 개인마다 다른 가중치가 부여되어야 하며, 다중속성 값의 복잡한 조합에 의해 중요도가 새로 결정될 수 있다는 점에 착안하여 다음과 같은 방법을 제안한다.

3. 필터링 방법

사용자의 판단기준에 따라 우선도가 부여된 메일에서 각각의 속성값을 추출하여 추출된 속성값과 우선도의 값을 기초로 구조화된 지식을 이용하여 각 개인에게 적용하기 쉬운 중요도를 산출한다. 이때 "우선도"는 기존의 문서에서 미리 사용자가 부여한 중요성의 레벨값이며 "중요도"는 입력문서에 대하여 시스템이 계산한 중요성의 레벨값을 의미한다. (그림 3)과 같이 메일필터링 시스템의 개요도에 대하여 설명한다.

3.1 학습부

학습부(실선)에서 사용자에 의해 대략적인 우선도가 판단되고, 각 폴더에 분류된 기존의 메일문서(학습데이터)를 형태소 해석한다. 문서의 유형과 시간제한 표현에 대한 속성값은 표현의 패턴수가 유한하며, 개인적인 차이가 적기 때문에 표현패턴을 등록한 배경지식을 별도로 구축하여 속성값을 검출한다. 송신처에 대해서는 검출한 메일주소를 코드화하고 그 코드를 속성값으로 변환한다. 또한, 코드와 주소의 대응은 주소파일에 기재하고 제목에 대해서도 폴더명을 코드화하여 해당코드를 속성값으로 변환한다. 폴더에 저장되어 있는 메일문서에 대하여 제목(Subject)란에서 사용된 단어와 본문에서 출현하는 명사류를 추출하고, 그 폴더에 대응하는 제목을 특정화하는 단어벡터로 작성한다.



(그림 3) 필터링 방법의 개요도

단어벡터는 출현단어와 빈도의 쌍으로 구성되며 제목(속성값의 코드값)과 단어벡터의 대응은 그룹파일에 기재된다. 단, 제목은 본문의 표제이므로 제목 내에 출현한 단어의 빈도는 K배의 가중치를 부여하고, 본문 내에 출현한 단어보다 빈도의 가중치를 크게 해야 한다. 만약 K값을 낮게 하면 제목의 내용이 활성화되지 않고, 반대로 K값을 높게 하면 본문 내에 자주 출현하는 중요어가 무시되어 버린다. 따라서, K의 값은 경험적인 파라미터의 값으로써 제목에서 커버되지 않는 부분이 메일문서의 본문에서 잘 보완되도록 설정하여야 한다. 이상의 순서로 검출한 다중속성 값과 우선도의 조합으로 생기는 다중속성을 각 메일문서마다 생성하고 이들 빈도 모두를 집계한 결과를 "프로파일"이라 한다.

◎ 배경지식

문서의 유형이나 시간제한을 나타내는 속성값을 추출하여 그룹핑하고 배경지식에 저장한다. 배경지식의 일부를 〈표 1〉에 나타내었다. 문서유형은 문장에서 나타나는 서술어를 참고문헌 [3, 4]를 참고하여 분류하였다. 시간제한은 문서전달의 긴박한 정도에 의해 설정하였다. 부사에 대해서는 각 표현이 어느 정도의 긴박도를 갖는가를 학부학생들에게 앙케이트 조사하였고, 그 결과를 각 속성값으로 할당하였다. 시간표현에 관해서는 표현형식의 차이에 의해 다음과 같이 두 종류로 나누었다.

〈표 1〉 배경지식의 예

속성	속성값의 이름	속성값	표현패턴의 예
문서 유형	권고	β_1	~할 것, ~해 줄 것, ~가능한 한 , ~하도록
	의무	β_2	~하고 있습니다, ~해야 하는 것, ~할 필요가 있는 것
	의뢰	β_3	~해 주세요, ~해줘, ~해 주었으면 한다.
	조건부 의뢰	β_4	만약 ~(이)라면 ...주세요,
	권유	β_5	~합시다, ~하자, ~하지 않겠습니까?
	명령	β_6	~해 주세요, ~하고 와, ~해라, ~해, ~할 것
	문의	β_7	~입니까?,~있습니까?, ~일까?, ~없습니까?, ~?
	공지	β_8	~연락합니다, ~알립니다, ~계획하고 있습니다
시간 제한 표현 (단, 시간구간을 표현 하지 않는 시간 표현은 송신시간을 고려하여 결정한다)	1주일 이상	γ_1	1개월 이내, 2~3주일 이내, 1개월 후
	1주일 이내	γ_2	1주일 이내, 7일 이내, 1주일 후에
	3일 이내	γ_3	잊지 않는 동안에, 2~3일 중에
	24시간 이내	γ_4	가능한 한 빨리, 빨리, 24시간 이내,내일
	12시간 이내	γ_5	급방, 급하게, 곧 바로, 1시간 이내에,1시간 후에

(1) 구간을 포함하고 있는 시간표현 어구

(2) 구간을 포함하지는 않지만 시점[9]을 포함하고 있는 시간 표현 어구 등등.

예를 들면 (1)에 속하는 '1시간이내'는 시간구간이 명확히 표시되어 있으면 해당하는 범위의 속성값으로 분류한다. (2)에 속하는 '내일까지'와 같은 표현은 현시점의 시간을 알 수 없으므로 시간구간을 구할 수 없다. 따라서 문서의 송신시간을 기준으로 시간구간을 구하고, 속성값으로 분류하였다. '다음 세미나 시간까지'와 같이 [시점]+[명사]로 이루어진 표현에 대해서는 [명사]에 관한 스케줄정보가 추가적으로 필요하므로 본 논문에서는 논의하지 않는다.

3.2 해석부

해석부((그림 3)의 점선표시 부분)에서 새로 도착한 메일문서에 대해 형태소해석 한 후, 배경지식을 참조하여 문서유형이나 시간제한 표현의 속성값을 추출하여 중요도를 산출한다. 제목의 추출은 입력문의 제목

과 본문에서 추출된 명사들의 그룹파일 내에서 각 단어 벡터를 비교하여 가장 유사한 폴더의 코드를 속성값으로 저장한다. 입력문서에 대하여서는 생성된 다중속성 값의 조합을 프로파일에서 검색하여 어떤 우선도가 주어져 있는가를 기초로 중요도를 확률적으로 계산할 수 있다.

4. 결론

본 논문은 수신된 메일문서에서 다중속성 항목으로 구성된 프로파일을 작성하고, 입력되는 메일문서에서의 중요도를 구하는 방법을 제안하였다. 본 논문의 방법을 이용하면 각 사용자가 중요성을 느끼는 내용을 포함한 메일문서를 가장 먼저 처리 할 수 있고, 중요도가 낮은 메일문서를 우선 처리할 업무에서 배제할 수 있어서 보다 신속하고 효율적인 업무처리 효과를 기대할 수 있다.

참고문헌

[1] Shishibori, M., Ando, K., & Aoe, J., "Filtering Method for E-mail Documents based on Personal Profiles," The 19th International Conference on Computer Processing of Oriental Language(ICCPOL '2001), pp. 069-072, 2001. (in Japanese)

[2] 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)

[3] 임유종 저, 한국어 부사 연구, 한국문화사, 1999.

[4] 신수송 역, 언어와 시간, 역락출판사, 2001.

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

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