Personal Profiles 기반의 E-mail 문서 필터링 방법에 관한 연구

최규정*, 이태헌*, 김명기*, 박기홍*
*군산대학교 컴퓨터정보과학과
e-mail:gyujung@kunsan.ac.kr

A Study on Filtering Method for E-mail Documents Based on Personal Profile

Kyu-Jung Choi*, Taehun Lee*,
Myoung-Ki Kim*, Kihong Park*
*Dept of Computer Information Science, Kunsan University

요약

요즘 E-mail은 중요한 통신수단 중 하나로 사용되고 있다. 그러나 상당수의 E-mail 문서들이 상업성 광고 E-mail과 같은 불필요한 정보를 포함한 채 우리들의 컴퓨터에 분포되어 있다. 본 논문에서는 이러한 문제를 해결하기 위하여 각각의 E-mail 문서들의 내용을 판단함으로써 불필요한 문서들을 자동적으로 필터링 하는 방법을 제안하고자 한다. 전통적인 필터링 방법들은 단어의 빈도수와 같은 단일 속성만을 다루기 때문에 높은 정확도를 얻을 수 없다. 따라서 본 논문에서는 각각의 사용자에 의해 이미 수신되어진 E-mail 문서들로부터 Personal Profile을 만들고, 이 Personal Profile를 사용함으로써 새로운 E-mail 문서가 사용자에게 중요한지의 여부를 구별하여 주는 방법에 관하여 제안하고자 한다. 이러한 Profile은 E-mail 문서의 송신자, 테마, 유형과 같은 다중 속성 값으로 구성되어 있다. 실험결과로부터 본 논문에서 제안하는 방법이 전통적인 방법보다 더 나은 정확성을 보이고 있음을 알 수 있다.

1. 서 론

요즘 컴퓨터 네트워크 기술이 기하급수적으로 발전함에 따라 E-mail 또한 현대 사회에서 없어서는 안될 중요한 통신수단중의 하나로 정착되고 있다.

그러나 각각의 사용자들에게 수신된 E-mail 문서들에는 상업성 광고 E-mail과 같은 중요한 내용을 포함하고 있지 않거나, 우리가 원하지 않는 E-mail 또한 수신되어진다. 따라서 내용에 기반을 둔 정보 필터링 기술의 중요성이 높아지고 있으며, 본 논문에서는 그 방법에 관하여 제안하고자 한다.

또한 본 논문에서는 중요도를 결정하는 요인으로서 E-mail을 보낸 송신자, 충고/요구/의문 등의 문장의 유형, 문장의 테마, 시간적 제한을 채택하고, 기존의 E-mail 문서 내에 포함된 각각의 속성 값과 사용자가 설정한 우선도와의 조합으로 이루어지는 Personal Profile을 작성한다.

이 Profile은 사용자가 여러 종류의 속성 값의 조합을 포함한 E-mail 문서에 중요성을 느끼고 있는지를 나타내고 있고, 이 Profile을 이용하여 개인별 필터링이 가능해진다.

2장에서는 E-mail 문서의 중요도란 무엇이며 어떤 요인으로부터 결정되는지를 명확히 설명하고, 3장에서는 Personal Profile의 작성 방법과 Personal Profile을 이용한 중요도의 산출방법을 설명하고, 4장에서는 평가와 요약을 보이고 5장에서는 향후과제에 관하여 설명한다.

2. E-mail 문서의 중요도

2.1 중요도의 요인

본 논문에서는 중요도가 높은 E-mail 문서로 다음과 같이 정의하였다.

1. 다른 E-mail보다 빨리 읽어야 할 필요가 있다.
2. 즉시 답장을 해야할 필요가 있다.

위와 같은 정의를 전제로 하면 E-mail 문서의 중요도는 문서의 테마나, 시간적 제한의 유무, 그리고 문장의 유형 등에 좌우된다고 생각된다. 또한 일반 문서에는 존재하지 않고, 단지 E-mail 문서에만 포함된 특유의 정보로서, 송신자, Cc, 인용문과 본문과의 관계, 과거의 E-mail과의 관련성 등도 중요도의 판정에 유효하지만, 본 논문에서는 형태소 해석에서 얻을 수 있는 정보로 필터링에 유효하다고 생각되는 항목으로 다음과 같은 항목을 중요도를 구성한 요인으로서 채택하였다.

- α : E-mail 문서의 송신자
- β : E-mail 문서 내에 포함된 문장의 유형
- γ : E-mail 문서 내에 포함된 시간적 제한
- θ : E-mail 문서의 테마

위의 네 가지 요소들은 중요도를 결정하기 위해 사용되는 속성이라고 하고, 각각의 속성들이 가지는 값들은 속성 값이라고 한다. 먼저 α의 속성 값은 E-mail 문서의 Header 부분의 "From" 항목으로부터 쉽게 얻을 수 있다. β의 경우에는 E-mail 문서의 본문 내에 포함되어 있는 불변화사나 조동사로부터 얻을 수 있다. γ의 속성 값은 시간을 표현하는 부사로부터 얻을 수 있다. θ의 경우에는 많은 명사들이 등록되어 있는 사전을 이용하는 것이 생각되지만, 적용할 수 있는 영역이 한정적이고, 어구해석과 의미해석과 같은 더욱 발전된 해석 도구를 사용해야 하며 실시간 처리가 곤란하다. 따라서 본 논문에서는 사용자에 의해 수신된 E-mail 문서들을 사용하여 테마를 특정하고자 한다. 먼저, 수신이 끝난 E-mail 문서를 사전에 사용자에 의하여 유사한 내용의 문서마다 분류하여, 메일 박스의 각 폴더로 분류되어 있는 것을 전제로 한다. 수신된 E-mail 문서에 관해서는 각각의 문서들이 저장되어 있는 폴더의 이름을 θ값으로 간주하고, 새로 수신된 입력 E-mail 문서에 관해서는, 각 폴더와의 유사도를 구하고 입력된 E-mail 문서와 가장 유사한 폴더의 이름을 θ값으로 한다.

2.2 중요도의 개인차

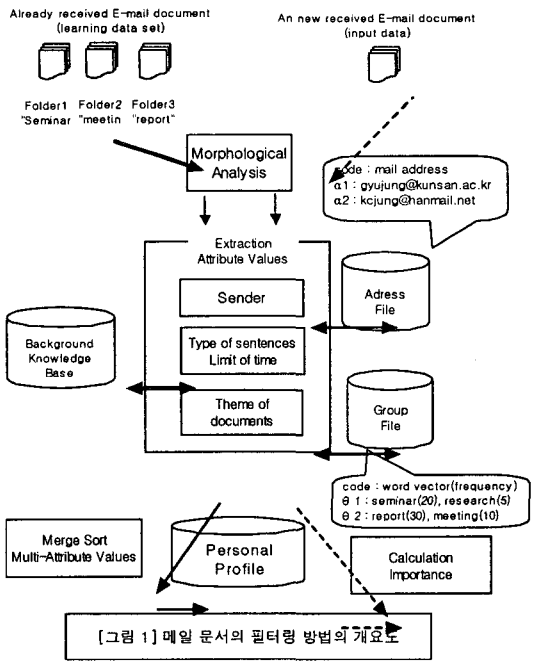
앞서 2.1장에서 말한 관점으로부터 생각하면, E-mail 문서의 중요성의 판단 기준이 개인마다 다른 것은 명백하다. 예를 들면, 송신자에 대하여, 「학생 A」는 「선생 X」로부터의 E-mail 문서에는 높은 중요도를 두고 있지만(이때 α의 값은 「선생 X」가 된다), 「학생 B」는 「선생 X」로부터의 E-mail 문서에는 그다지 중요성을 느끼고 있지 않을지도 모른다.

이처럼, 개개인마다 중요한 속성치는 다르기 때문에, 필터링을 할 때에는 각각의 개인마다의 지식이 필요해진다. 또한 문장의 유형의 경우에도 마찬가지로, 「학생 A」는 「선생 X」로부터 수신된 E-mail 문서에서 '충고'의 E-mail 문서에는 높은 중요도를 느끼고 있지만, '의뢰'의 내용을 포함한 것은 중요도가 낮을지도 모른다. 반면에 「학생 B」는 「학생 A」와 반대의 조합에 중요도를 두고 있을지도 모른다. 이처럼, 각각의 개인마다 중요한 속성 값의 조합은 다르기 때문에, 중요도 판정 지식으로서 속성 값 조합을 위한 정보의 준비가 필요하다. 이렇듯 E-mail 문서의 중요도는 많은 속성 값들의 복잡한 결합에 의해서 결정되어지고, 중요성의 가중치는 각각의 속성 값마다 개별적으로 모두 다르다.

3. E-mail 문서의 필터링

3.1 필터링 방법의 개요

E-mail 문서를 필터링하기 위한 방법의 개요는 [그림 1]에서 보여주고 있고, 이번 장에서는 학습 모듈과 해석 모듈에 관해서 설명하고자 한다. 먼저 학습 모듈([그림 1]에서 실선으로 표시되어 있다)에서는 사용자에 의하여 미리 우선도가 부여되고, 각각 폴더로 분류되었던 기존의 E-mail 문서를 형태소 해석한다. 형태소 해석을 통하여 E-mail 문서로부터의 각각의 속성 값을 얻을 수 있다. 문서 내에 문장의 유형 및 시간적 제한에 대한 속성 값은 표현 패턴 수가 유한 개이고, 또한 개인차가 적기 때문에, 표현 패턴을 등록한 Background Knowledge Base을 이용하고 속성 값을 검출한다. 또 "송신자"에 관해서는, 검출한 E-mail Address를 코드화하여, 그 코드를 속성 값이라고 한다. 코드와 Address의 대응은 Address File에 기재한다. '테마'에 대해서도 폴더 이름을 코드화하여, 그 코드화한 값을 속성 값이라고 한다. 또한 각각의 폴더에 대응하여 그 폴더의 테마를 특징짓는 단어 벡터를 작성한다. 단어 벡터는 단어와 단어의 출현 빈도수로 구성되고, 문서의 테마와 단어 벡터의 대응은 Group file에 기재한다. 단, "Subject"는 본문의 제목이나 또는 요약이므로,



[그림 1] 메일 문서의 필터링 방법의 개요도

"Subject"안에 출현한 단어의 빈도는 'K' 배가되고, 본문 내에 출현한 단어보다 빈도의 비중을 무겁게 한다. 여기에서, 'K'의 값을 낮게 하면, "Subject"내의 내용에 포함된 단어를 충분히 활용할 수 없고, 역으로 높게 너무 하면, 본문 내에 포함된 중요어가 무시되고 버린다. 따라서 'K'의 값은, "Subject"로는 다 커버할 수 없는 부분을 본문에서 알맞게 보충할 수 있도록 설정한 경험적인 패러미터 값이 되어야 한다. 해석모듈([그림 1]에서 점선으로 표시되어 있다)은 중요성이 알려지지 않은 새로 입력된 메일 문서에 대하여 형태소 해석을 적용시킨 후, Background Knowledge Base를 참조하여 "유형" 및 "시간적 제한"의 속성 값을 얻는다. 또한 "송신자"를 나타내는 E-mail Address는 Address File을 참조하여 속성 값을 변환한다. "테마" 관해서는 입력 문서의 "Subject" 및 본문내의 명사와 Group File내의 각 단어 벡터를 비교하여, 가장 유사한 폴더의 코드를 속성 값으로 얻을 수 있다. 이처럼 새로 입력된 문서에 대한 다중 속성(Multi-attribute) Set를 생성한 후, 이 다중 속성 Set과 동일한 조합의 다중 속성 Set을 Profile 안에서 검색하고, 그것들이 어떤 우선도를 부여받고 있는 지를 기초로 하여 중요도를 확률적으로 찾는다.

3.2 학습 모듈

"유형" 및 "시간적 제한"의 속성 값을 검출하기 위하여 Background Knowledge에 저장되어 있는 많은 표현 패턴들 중 같은 속성 나타내는 일부 패턴들은 1개의 속성 값으로 그룹화 한다. "유형"에 관해서는 어미에 나타나는 조술 표현을 문헌을 참고로 분류하였고, "시간적 제한"에 관해서는 시간의 흐름 정도에 따른 속성 값들 사이의 한계 값을 정하였다. 그리고 부사에 관해서는 각각 표현이 어느 정도의 긴박도를 보유 하든지를 10명에게 앙케이트 조사를 하였고, 그 결과로부터 각 속성 값을 할당하였다. 먼저, 학습 문서 데이터를 $D = \{d_1, d_2, \dots, d_i, \dots, d_A\}$라고 하면, 각 문서는 $d_i = \{x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{iB}\}$라고 말한 다중 속성 Set를 갖고, 각 다중 속성 Set는 $x_{ij} = (\alpha_l, \beta_m, \gamma_n, \theta_q, \kappa_r)$이 되고 속성 값은 다 차원 벡터로 된다.

- 다중 속성(Multi-attribute) Set 구성 방법

1) "송신원"의 속성 값 α_l을 다중 속성 Set 대입
2) "유형"의 속성 값 β_m이 복수로 존재한 경우, 그 수만 다중 속성 Set을 복제하고, β_m을 대입
3) 2)번으로 조술 표현이 검출된 동일 문서내 시간적 제한의 속성 γ_n 값을 검출하고, 검출 수만 다중 속성 Set을 복제한 후, γ_n을 대입
4) 복제된 각 다중 속성 Set에 "테마"의 속성 값 θ_q와 우선도 κ_r을 대입

하나의 문서마다 생성되는 다중 속성 Set의 개수에 따라 빈도수 $freq(x_{ij})$를 준다. 단$(1 \le j \le B)$라고 하면, $freq(x_{ij}) = 1/B$이고, $\sum_j freq(x_{ij}) = 1$라고 한다. 여기에서 동일한 다중 속성 Set의 빈도를 집계한 결과를 프로 파일: $P = \{p_1, p_2, \dots, p_k, \dots, pc\}$라고 한다. 단, p_k는 다중 속성 Set를 나타내고 $\cap p_k = \varnothing$, pk와 동일한 내용의 다중 속성 Set x_{ij}의 빈도 합계 치라고 한다. 이 Profile은 사용자가 어떤 속성 값의 조합에 중요성을 느끼고 있는지를 나타내고 있다. 그리고 다중 속성 Set$(\alpha_l, \beta_m, \gamma_n, \theta_q, \kappa_r)$을 간단하게 p<l,m,n,q,r>라고 기록한 것으로 한다.

3. 3 해석 모듈

해석 문서 데이터를 $T = \{t_1, t_2, \dots, t_i, \dots, t_E\}$라고 하면, 각 문서는 $t_i = \{y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{iF}\}$라고 말한 다중 속성 Set의 집합을 갖는다.

정상적인 처리에서의 $y_{ij} = (\alpha_l, \beta_m, \gamma_n, \theta_q)$의 중요도는 다음 식(1)에 의해서 구해진다.

$$P(\kappa_r | \alpha_l, \beta_m, \gamma_n, \theta_q) = \frac{freqp(<l, m, n, q, r>)}{freqp(<l, m, n, q>)} \quad \dots (1)$$

이 식은 다시 아래와 같이 나타낼 수 있다.

freqp(<l, m, n, q>) = Σfreqp(<l, m, n, q, r>)

식(1)에서 구한 값은 확률 값으로서 우선도의 각 Rank마다 요구받기 때문에, 아래 식(2)로 다시 표현할 수 있다.

$R_{ij} = \Sigma\{r \times P(\kappa_r | \alpha_l, \beta_m, \gamma_n, \theta_q)\}$ ..... (2)

최종적으로 입력 문서 t_i의 중요도 R_i는, $R_{(i1)} \sim R_{(iF)}$의 평균 중요도로서 구한다.

4. 평가

본 논문에서 제시하는 방법의 유효성을 확인하기 위해, 학습 데이터 수와 해석 정밀도의 관계, 각 속성의 유효성 및 기존 방법과의 비교에 관한 실험을 했다. 실험에 이용한 데이터 내용을 명시한 후 각종 실험 결과를 개별적으로 나타냈다. 실험 데이터는 기초 데이터로서 기존의 E-mail 문서 150~280통을 각 사용자마다 마다 준비했다. 그리고, 각 폴더 안의 문서수에 비례하여 무작위 하게 추출한 각 30통의 E-mail 문서를 평가용 데이터로 하고, 나머지 문서를 학습용 데이터라고 했다. 또한, Background Knowledge로서, 문서의 유형은 282개(속성 치수 8) 시간적 제한에는 48(속성 치수 5)개의 표현 패턴을 등록했다. 이러한 데이터를 이용하여 각 Profile을 작성한 후, 평가용 데이터의 중요도를 구하였다.

피험자	A	B	C	D
학습 문서의 수	250	230	150	120
합계 크기(Kbyte)	655	501	233	152
송신자 수	59	57	7	28
폴더 수	7	12	6	5
다중 속성 Set	310	291	203	161
우선도의 평균	3.3	3.14	3.63	3.46
우선도의 분산	1.27	1.31	2.11	2.07

[표 1] 학습 데이터의 내용

또한 경험적인 패러미터 값 K를 결정하기 위해 패러미터 K의 값을 1~20까지 1마다 변화시키면서 학습용 데이터에 대한 단어 벡터를 구성하고, 평가용 데이터에 대한 테마의 검출 정밀도(피험자 4 인의 평균)를 구했다. 테마의 검출에 최적인 K의 값은, 문서양이나 내용 등에 의하고 변화한다고 생각되지만, 이번의 실험에서는 평가용 데이터에 최적화한 값으로서 K=10을 이용했다.

피험자	Rocchio	Windrow-Hoff	본 논문
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
평균값	0.40	0.43	0.52

[표 2] 기존의 방법과의 비교 결과

평가 방법으로서는 기존의 방법과 비교 수법을 사용하였고, 산출한 중요도와 사용자가 판정한 우선도와의 상관 계수(1에 가까울 수록 상관이 높다)를 구했다. 또한 신뢰성을 높이기 위해, 10회 교차 검정을 했다. 10 회의 계측 결과의 평균을 [표 2]에 나타내었다. [표 2]로부터, 단어의 빈도수 같은 단일 속성 정보밖에 이용하지 않는 기존의 방법보다도 본 논문에서 제시하는 방법이 사용자의 판단에 도움이 되고 또한 상관을 갖는 것이 확인할 수 있었다.

마지막으로, 각종 처리의 시간 효율은, 평균 학습 시간은 약 11.5초(평균 학습 문서 Size는 385 Kbytes) 해석 시간은 약 1.2 초, 단, 쌍방 모두 형태소 해석시간도 포함하고 있다. 또, Profile의 평균 사이즈는 약 4.6Kbytes이다.

5. 결론

본 논문에서는, 수신이 끝난 E-mail 문서로부터 다중 속성 Set 항목으로 이루어지는 Profile을 작성하고, 중요도가 알려지지 않은 새로운 입력 E-mail 문서의 중요도를 구하는 방법을 제안했다. 본 논문에서 제시하는 방법을 이용하여 각 사용자가 중요성을 느끼고 있는 내용을 포함한 E-mail 문서로부터 제시하거나 또한 중요성이 낮은 E-mail 문서를 배제하는 것이 가능하다. 향후 연구과제로는, Background Knowledge를 충실히 하고 이번 논문에서는 배제하였던 E-mail 문서에만 보이는 특유한 속성을 고려하여, 필터링 정밀도의 향상을 목표로 한다.

참고문헌

[1] Salton G.(1988) "Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer" Addison-Wesley

[2] Buckley C. Salton G & Allan J.(1994) "The Effect of Adding Relevance Information in a Relevance Feedback Environment" Proc. of the 17th Annual International.