전자우편문서의 효율적인 분류를 위한 전처리

강영순[0] 이용배 김태현 조숙현 맹성현
충남대학교 컴퓨터과학과
yskang@hananet.net, {yblee, heemang, shcho, shmyaeng}@cs.cnu.ac.kr

A Preprocessing for Efficient Classification of E-mail Messages

Young-Soon Kang, Yong-Bae Lee, Tae-Hyun Kim, Suk-Hyon Jo, Sung Hyon Myaeng
Dept. of Computer Science, Chungnam National University

요 약

인터넷 사용의 증가는 의사소통 매체의 하나로 전자우편(e-mail)을 일반화되게 했다. 전자우편은 개인적인 목적 뿐만 아니라, 광고, 판매, 서비스 혹은 제품구입 관련문의 등의 특정목적에 이용되고 있는 추세이므로 한꺼번에 많은 메일을 처리 및 관리하기 위해서는 전자우편문서의 자동분류가 필요하다. 전자우편문서는 일반문서와는 달리 반구조적(semi-structure) 구성, 특수문자, 약어 및 속어 등의 특징들이 있으므로 이러한 특성들은 자동분류의 정확도에 영향을 끼치는 요인이 될 수 있다. 본 논문에서는 분류 성능을 향상 시키기 위해 자동분류의 오류가 될 수 있는 특성들을 제거하고, 구조적인 특징을 활용한 분류기의 전처리기를 설계한 방법론을 제시하고자 한다

1. 서론

현재 디지털 형태 문서의 보편화 및 계속적인 증가로 문서를 자동 가공하여 처리하는 문서 자동 분류의 중요성이 널리 인식되고 있다. 문서 자동 분류란 미리 수집되어 있는 문서집합을 바탕으로 부류를 나누어 놓고, 학습을 통해 지식 베이스를 구축한 후, 새로운 문서를 각 부류에 대응 시키는 것을 말한다. 이러한 문서 자동 분류를 이용하여 새로운 뉴스의 분류, 회사로 들어오는 방대한 양의 전자우편을 해당 부서로 라우팅 시키는 등 실제 응용에 관한 연구[1]가 수행되고 있다.

이러한 연구 실험에 적용된 문서집합은 잘 조직된 데이터 세트를 이용하기 때문에 현실에 적용하는 데는 많은 어려움이 있다. 예를 들면, 전자우편문서의 경우가 그렇다. 실험에 적용된 데이터 세트는 신문기사나, 논문 등으로서 전문가에 의해 잘 작성된 문서임에 반하여, 전자우편문서는 일반 사용자들이 작성한 문서로서 그 형식 혹은 표현방식에 제한이 없으므로, 특정 기호 삽입이나 맞춤법 오류 등이 내재되어 있다. 즉, 통신상의 속어나 약어의 표현 등이 대표적 예이다. 이러한 전자우편문서의 속성들은 일반문서와 달리 자동 문서 분류에 잡음(noise)들로서 분류의 정확도를 저하시키는 요인이 될 수 있다. 따라서 본 논문에서는 전자우편문서의 효율적인 자동분류를 위하여, 전자우편문서의 잡음요소를 제거하고, 전자우편문서의 반구조적 (semi-structure) 특성을 활용할 수 있는 전처리기 모듈을 설계하고 분류기와 통합시키는 방법을 기술한다.

2. 관련 연구

기존의 자동 분류 시스템은 정형화된 데이터에 대한 실험만으로 검증되었기 때문에 정형화되어 있지 않고 잡음이 많이 섞인 전자우편문서를 처리할 경우, 그 성능이 좋게 나오지 않기 때문에 그 특성과 관련한 분류 시스템 연구에 대한 관심은 높지만 실질적인 비교 실험 분석은 미비한 실정이다.

분류성능을 향상시키기 위한 전처리 모듈 도입과 관련한 구체적인 실험 분석은 아니지만 전처리기를 설계하기 위한 전자우편문서 자동 분류를 위한 관련 연구는 크게 세 가지로 요약할 수 있다. 규칙기반(rule base)에 근거한 메시지 타입 기반과 규칙세트, 전자우편문서의 편중성(본 실험에 적용된 전자우편문서 1000개 중에서 각 범주비율은 최대 37.1%, 최저 1.3%)을 완화시켜 보려는 일환으로 계층적 문서 클러스터링 방법 및 실제 질의 전자우편에 대한 특징추출 방법의 성능비교에 대한 연구 등이 있다.

2.1 규칙 기반의 메시지 타입 분류기

규칙기반의 메시지 타입 기반의 전자우편문서 자동 분류[2]는 사전에 네 가지 클래스를 기준으로 분류하여 프로그램에 입력된 메일을 세그멘테이션 하여 문서의 로그정보와 텍스트 번호를 붙여주고, 그 다음에 입력된 문자와 사전에 정의된 숫자 범위 매칭으로 표현된다. 이는 출력된 숫자들이 가장 많이 포함된 메시지 그룹을 분류하는 방식으로, 메시지 타입 기반의 전자우편문서 자동 분류는 실질적인 전자우편문서 내용과 사용자가 미리 정한 각각의 메시지 타입의 문서 세트와 문자를 매핑한다. 메시지 타입은 전자우편문서의 어떤 특징적인 요소를 찾아내는데 도움을 줄 수 있는 특성이 될 수 있다. 즉, 특정한 관심 분야 외의 나머지 주제는 무시할 수 있음을 의미한다.

규칙기반의 일종으로서, 전자우편문서의 반구조성을 적용한 RIPPER keyword-spotting rules와 기존의 일반적인 분류기 알고리즘인 TF-IDT 가중치로 전자우편문서를 분류하는 학습 방법을 비교 실험[3]했다. 일반 분류기들은 모든 단어들의 가중치의 결합에 의한 단어의 빈도수로 범주를 결정하는 반면, RIPPER keyword-spotting rules는 의사결정을 하기 위해 몇 개의 키워드로 규칙세트(rule set)를 구축하여 단어의 출현/부재로 그 범주를 결정하여 정확성의 손실정도를 그 평가 목적으로 실험했다. 이러한 결과, 기존의 분류기에 비하여 상대적으로 적은 데이터(100 samples)로 일반화를 얻어므로서, 학습률 속도가 빠르고 에러율 감소를 보였다.

2.2 문서의 편중성을 고려한 분류

전자우편문서의 편중성을 고려한 연구[4]에는 전자우편

문서 집합의 부류 개수의 불균형은 단일 신경망으로 학습할 경우, 공정하지 못한 학습의 유도로 전체적인 성능 저하를 초래 할 수 있다. 이에 대한 한 방안으로서, 다중 신경망은 단일 신경망이 분류해야 할 부류 하나를 모델링하고, 이러한 신경망들 전체의 결과를 종합하여 최종 결론을 도출하는 방식이 있다. 이 방식은 성능 저하를 최소화하며 각 부류별 복잡도에 따른 신경망을 구축하여 효과적인 해결책을 제시하였다.

부류의 연관관계[5]를 고려하여, 계층적 문서 클러스터링 방법을 전자우편문서 분류에 적용하여 단층분류보다 약 11.4%의 인식율 향상을 보이는 실험도 있다.

2.3 전자우편문서의 특징추출 방법에 대한 성능 비교

이 실험은 질의 전자우편문서에 대한 효율적인 특징추출 방법에 따른 성능의 비교[6]를 했다. 즉, 문서빈도(document frequency), 정보이득(information gain), 상호 정보량(mutual information), 카이제곱 통계 , 적합성 점수(relevancy score), 교차비(odd ratio) 및 단순화된 카이제곱 등의 7가지 특징 추출 방법을 사용하여 실제 질의 전자우편문서에 적용한 실험을 했을 때, 카이제곱이 가장 높은 인식비율을 보였다. 이는 전자우편문서의 효율적인 자동분류를 위해서 특징추출에 대한 선택도 고려되어야 함을 의미한다.

3. 전자우편문서의 전처리 설계 및 구현

본 논문에서는 특징 추출 단계에 적합하도록 전자우편문서를 변환하여 정형화된 색인어 집합을 추출하는 과정을 기술하고자 한다. 전처리 단계에서 특수기호 및 태그의 제거, 속어 및 약어에 대한 표준어 변환, 분류 대상 문서의 성격에 따른 대표어 변환과 불용어 제거 등을 처리하여 비정형화된 문서를 정제하고, 제목에 있는 단어들에 대하여 높은 가중치 부여로 분류의 정확율을 높이고자 한다.

문서 자동 분류는 분류기 자체의 성능 뿐만 아니라 분류의 대상이 되는 문서의 특성에 의해 그 결과가 크게 좌우될 수 있다. 따라서 일반적인 문서 자동 분류에서는 정제되고 정형화된 문서를 대상으로 분류를 수행한다. 그러나, 현재 보편적으로 사용되고 있는 통신상의 문서들은 비정형적인 특성을 많이 포함하고 있다. 특히, 전자우편문서와 같은 경우 속어 및 약어의 빈번한 사용과 자유로운 문체로 인하여 그 비정형성이 어느 다른 문서들 보다도 크게 나타나고 있다. 따라서 이러한 문서들을 대상으로 하는 분류의 경우, 분류의 정확도가 떨어지게 된다. 본 논문은 이러한 전자우편문서의 특성을 파악하여 전자우편문서의 비정형적인 특성들을 최소화 함으로써, 전체적인 분류성능의 향상을 꾀하는데 목적이 있다.

본 논문에서 다루고 있는 비정형 문서인 전자우편문서의 전처리 단계에 고려한 사항들로 다음과 같은 것들이 있다.

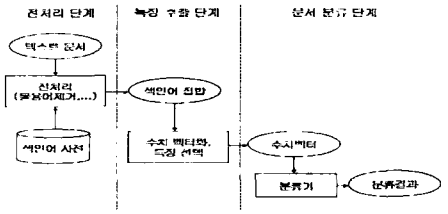
● 색인이 이루어지기 전에 문서 자체에 내재된 에러 요소를 제거해야 한다. 여기에서 말하는 에러 요소는 색인기가 정상적인 색인을 하는데 있어 장애가 되는 요소로, 대표적인 것으로는 2바이트 특수기호(2bytes symbols)가 있다.

● 문서내의 불필요한 정보(잡음요소)를 제거해야 한다. 일반적으로 사용되면서 그 자체는 큰 의미를 갖지 않는 단어들은 제거 되어야 한다. 이는 문서분류에 있어 변별력을 저해하는 큰 요인이 될 수 있다. 이를 위해 본 연구에서는 불용어 목록을 작성하여 색인기에 넘겨 줄 수 있도록 하는 부가적인 모듈을 구성하였다. 이로써, 색인 과정에서 보다 정제된 색인어들을 추출할 수 있도록 한다.

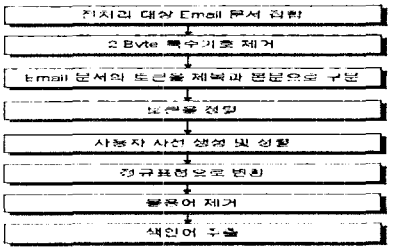
● 전자우편문서에는 HTML의 태그 형식이 허용된다. 따라서, 태그 정보가 전자우편문서의 일반 내용과 함께 색인될 경우 불필요한 색인어들을 산출해 내게 되므로, 전자우편문서에 포함된 태그들도 제거되어야 한다.

● 전자우편문서는 크게 제목과 내용으로 이루어진다. 이러한 구성상의 특성을 이용함으로써 분류의 성능을 높일 수 있다. 본 연구에 사용된 1,000개의 샘플 문서 중 52.3%가 그 제목에 범주와 관련한 정보를 포함하고 있음을 알 수 있었다. 따라서, 제목에 대해 가중치를 부여하는 방법을 도입함으로써 분류 성능을 향상시킬 수 있다.

전자우편문서는 전문가가 작성한 뉴스기사나 논문과 달리 일반 사용자들이 작성한 문서이기 때문에 맞춤법이 틀린 표현이 많고, 통신상의 약어나 속어 등을 많이 사용하므로 색인어 추출이 어렵다. 따라서 일련의 전처리 과정을 거친 전자우편문서를 입력으로 만들어진 색인어 집합을 이용하는 것이 문서를 보다 정형화된 벡터로 표현할 수 있어 분류에서의 잡음 요소들 최소화로 문서 분류의 정확율을 높일 수 있는 좋은 방법이라 할 수 있다. 본 연구에서는 다음의 [그림 1]과 같이 문서 분류 시스템을 구성하고, 전자우편문서 분류를 위해 특수화된 [그림 2]와 같은 흐름도를 가진 전처리기를 제안한다



[그림 1] 문서 분류 시스템의 구성도



[그림 2] 전처리 단계의 흐름도

전처리의 대상이 되는 전자우편문서는 제목과 본문으로 구성되어 있다. 전처리기는 우선 문서의 정보를 표현하는데 있어 불필요한 요소인 특수 기호들을 제거한다. 그리고, 제목과 본문을 구분하여 토큰을 생성하고 이를 정렬한다. 이 단계에서 1바이트 기호들과 태그가 제거된다. 다음 단계로 전처리기는 약어, 속어 및 특정 집합에 속하는 표현들을 표준화하는 표현인 대표어에 대한 사용자 사전을 로딩하여 정렬된 상태로 유지하게 된다.

다음 단계에서 전처리기는, 불필요한 기호 및 태그 정보를 제거하여 만들어진 벡터화 된 전자우편문서와 로딩된 사용자 사전을 매핑하여 전자우편문서에 나타나는 비정형화된 단어들을 정형화된 단어로 변환한다. 더불어 전자우편문서의 제목에 나타난 단어들의 변별력을 높이기 위해 이들을 이중화(duplication)하는 방법을 이용하여 가중치를 높여준다.

전자우편문서 자동 분류의 정확율을 향상하기 위하여, 전처리 과정에서 면밀히 고려해야 할 사항은 사용자 사전 구축이다. 통신상의 약어 및 속어는 어느 정도 일반화 할 수 있겠으나, 전자우편문서 응용의 특성에 따라 다양한 표현에 대한 통일된 표현어 변환은 특정 응용에 국한되어 있으므로 일반화가 어렵다. 본 실험에서는 1000개의 전자우편문서에 국한한 수동 사용자 사전 구축으로 전처리 하였으나, 실제 응용에서는 이보다 훨씬 많은 데이터와 다양한 응용들이 존재하므로, 사용자 사전의 반자동 생성이 필요하다. 이때, 사전은 전자우편문서의 응용환경에 따라 관리자로 하여금 추가, 삭제 및 변경할 수 있는 응용 맞춤형 사전이라 할 수 있겠다.

4. 실험 및 결과

4.1 실험 환경

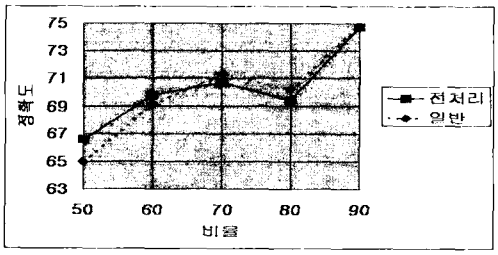
본 논문에서 제안한 전처리기 모듈은 베이지안 분류기[7]에 통합하여 실험했다. 사용된 전자우편문서는 1000개의 샘플 데이터로서 8개의 범주로 나누어 진다. 각 범주에 대한 통계는 [표 1]과 같이 특정 범주에 문서가 편중되어 있음을 알 수 있다.

[표 1] 샘플 데이터의 각 범주(중복범주허용)

범주	문서 수	범주	문서 수
1.애프터서비스	371	5.비밀번호관련	59
2.불편사항	121	6.기종문의	90
3.문의 및 해지분실	13	7.업그레이드	119
4.기타불만	157	8.사용법문의	108

4.2 실험 및 분석

이 실험에서는 각 학습 비율을 50%, 60%, 70%, 80% 및 90%로서, 전처리한 문서와 하지 않은 문서를 구분한 분류([그림 3]과 [표 2])와 전자우편문서 분류에서 제목 부분이 가지는 변별력을 알아보기 위해 제목, 본문(내용) 및 제목+본문([그림 4]) 각각에 대해서도 분류해 보았다.

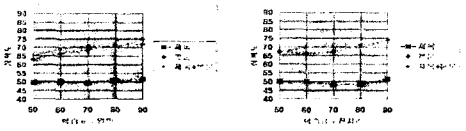


[그림 3] 전처리 및 미전처리 전자우편문서의 분류

[그림 3]에서 문서를 전처리한 분류의 결과가 그렇지 않은 결과에 비하여 약간의 성능 차이만을 보인다. 이는 적은 샘플 데이터의 수, 전처리 대상인 특수문자나 기호들이 전체 문서에서 차지하는 비율이 일부에 불과한 데 그 원인으로 할 수 있겠다.

[표 2]전처리 및 미전처리 전자우편문서의 분류 결과

학습율	일반 메일	전처리 메일	증감률
50%	65.0%	66.6%	+2.4%
60%	69.1%	69.8%	+1.0%
70%	71.3%	70.7%	-0.8%
80%	70.2%	69.4%	-1.1%
90%	74.7%	74.7%	0.0%



[그림 4] 제목, 본문, 제목+본문

[그림 4]는 전자우편문서를 구조적으로 분할하여 실험했다. 이 실험은 전자우편문서의 반구조적인 특성을 반영한다. 제목에 대한 분류 대상은 전체 샘플 데이터 중 1%의 제목란이 없는 것을 제외한 990개의 샘플 데이터로 실험한 결과, 제한된 정보(전자우편문서의 본문이 제공하는 정보량에 비교하여)로 50% 가량의 분류 정확율은 전자우편문서 자동 분류에서 신중한 고려 대상임을 보여준다.

5. 결론 및 향후연구

본 논문은 일반 분류기에 전처리기 모듈을 통합하여 전자우편문서에만 포함된 특성들을 고려하여 보다 효율적으로 분류되도록 했다. 즉, 형식에 제약 받지 않고 맞춤법이나 철자를 고려하지 않은 문자들 및 표준어 형식에 어긋난 속어나 약어의 사용, 신조어, 특성 상품 모델명 등에 대한 용어들을 파악하여 표준화 및 통일된 표현으로 처리할 수 있도록 함으로써 전자우편문서 분류의 성능을 향상 시켰다.

일반문서와 달리, 많은 잡음 요소들이 내재된 전자우편문서에 대한 분류의 정확율을 향상 시키기 위해서는 본 논문에서 제시한 전처리기 모듈 도입 뿐만 아니라, 전자우편문서의 구조적 특성에 적합한 분류기가 필요하다.

향후 연구 과제로, 다양한 응용에 대한 전자우편문서 말뭉치 구축과, 관리자의 선택에 따라 사전의 내용을 추가, 삭제 및 변경할 수 있는 반자동 사용자 사전구축 방법, 전자우편문서에 적절한 분류 알고리즘 개발이 필요하며, 이를 이용한 일반문서에 적용된 기존의 분류기들과의 성능비교가 뒤따라야 겠다.

참고문헌

[1] F. Sebastiani, " Machine Learning in Automated Text Categorization," Technical Report IEI-B4-31-1999, *Istituto di Elaborazione dell' Informazione*, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999.

[2] Andrew D. May, " Automatic Classification of E-mail Messages by Message Type," Journal of the American Society for Information Science, 48(1): 32-39, 1997.

[3] William W. Cohen, " Learning Rules that Classify E-Mail," *AAAI Spring Symposium on Machine Learning in Information Access*, 18--25, 1996.

[4] 이지행, 조성배, " 다중 신경망을 이용한 한메일넷 질의 자동 분류 시스템," 제27회 추계학술발표회, 한국정보과학회, p. 232~234, 2000.

[5] J. Ryu and S.-B. Cho, "Automatic categorization of real world FAQs using hierarchical document clustering," Proc. Korea Fuzzy and Intelligent Systems, Seoul, May 2001. (In Korean)

[6] 홍진혁, 류종원, 조성배," 실세계의 FAQ 메일 자동분류를 위한 문서 특징추출 방법의 성능비교," 제28회 춘계학술발표회, 한국정보과학회, p.232-234, 2001.

[7] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, " Modern Information Retrieval," Addison-Wesley, 1999.