유사어 사전을 이용한 자동범주화 모델 개발

Automatic Text Categorization Model by Synonym Dictionary

김규환, 중앙대학교 문헌정보학과, emm1996@hanmail.net
이두영, 중앙대학교 문헌정보학과, leety0521@hanmail.net

Qu-Hwan KIM Graduate School of LIS, Chung-Ang University
Too-Young Lee, Dept. of LIS, Chung-Ang University

기존의 문서분류는 학습문서에 출현하는 자질에 대해 가중치를 계산하여 그 순위에 따라 상위 자질로 구성된 지식베이스를 사용하였다. 그리고 새로운 문서가 들어왔을 때 자질 지식베이스를 근거로 새 문서를 색인하였다. 결국 자질 지식베이스와 정확히 일치하지 않는 키워드는 색인대상에서 제외되는 문제가 있었다.

본 고에서는 이 문제를 해결하기 위하여 분류될 문서의 특징을 나타내는 범주별 자질과 유사한 의미를 가지나 형태가 변형되어 기술된 단어에 대하여 유사어 사전을 구축하였으며 이를 통해 새로운 문서가 범주에 할당될 가능성을 높여 자동 문서 범주화 시스템의 성능을 향상시키고자 한다.

1 서 론

1.1 연구의 필요성 및 목적

자동 문서 범주화는 미리 정의된 범주에 문서를 자동으로 할당하는 기법과 관련된 연구분야로서, 대량의 문서의 효율적인 관리 및 검색을 가능하게 하는 동시에 방대한 양의 수작업을 감소시키는 데 그 목적이 있다.(Yang, 1997)

자동 문서 범주화는 일반적으로 문서를 어떤 자질을 통해 표현할 것인가를 다루는 자질 추출과정과 자질로 표현된 문서를 어느 범주로 할당할 것인가를 결정하는 문서 분류 과정으로 구성된다. 자질추출과정에는 문서를 표현하는데 사용할 자질을 선택하는 자질 선택과정과 선택된 자질로 어떻게 문서를 표현할 것인가에 대한 색인과정이 포함되며, 가장 일반적인 색인 방법은 이른바 벡터 공간 모델이다(Slaton, 1983).

자질 선택과정은 전처리 단계를 통해 문서에 나타나는 후보 자질들 중에서 범주화 구분에 유용하게 사용될 만한 자질을 찾아내는 과정이다. 학습 문서에 나타나는 후보 자질의 수는 다양하기 때문에 모든 후보 자질이 자질로 선택된다면 학습 및 분류 시간이 오래 걸리게 된다. 그러므로, 문서 범주화의 성능의 저하 없이 자질의 수를 줄이기 위하여 학습문서에 나타나는 후보 자질의 정보량을 계산하고 정보량이 큰 자질만을 선택하려는 연구가 활발히 진행되어 왔다.

자질 선택 기법 중에 대표적인 방법으로는 정보 획득량(Information Gain)과 카이 제곱 통계량(x2 Statistic)이 가장 좋은 성능을 보이는 것으로 평가된다. 이렇게 선택된 자질은 문서를

표현하는 색인과정에 사용된다.

기존의 문서분류는 자질에 대한 정보량을 계산하여 그 순위에 따라 상위 자질들을 지식베이스로 사용하여 새로운 문서가 들어왔을 때 이를 기반으로 색인을 하였기 때문에 자질 지식베이스와 일치하지 않는 자질은 추출되지 않았다.

본 고에서는 분류될 문서의 특징을 나타내는 범주별 자질과 유사한 의미를 가졌지만 그 형태가 변형되어 표현된 자질에 대하여 유사어 사전을 구축하여 새로운 문서의 분류 가능성을 높여 문서 자동범주화시스템의 전체적인 분류 성능을 향상시키고자 한다.

1.2 연구 방법

기존 자동범주화시스템에서 기 분류된 범주별로 문서를 할당하는 방법은 주로 범주를 대표하는 대표 자질과 신규로 입력되는 문서의 자질과의 단어 및 유사도 비교에 의해 할당하게 된다. 이런 대표 자질의 선택은 학습문서에 나타나는 여러 후보자질들 중 범주화 구분에 유용하게 사용될 만한 자질들만 선택하는 작업을 통해서 이루어진다.

본 고에서는 범주별 학습문서에 대한 전처리 과정을 걸쳐 명사만을 후보자질로 추출한다. 추출된 후보자질들을 카이 제곱 통계량을 이용하여 범주별 가중치를 부여하고 순위화 한다. 그리고 "자질 수에 따른 문서 분류율" 성능평가를 수행하여 문서 분류기의 최적의 자질 수를 구한다. 추출된 범주별 자질 수 내에서 유사어 사전을 구축한다. 자질 유사어는 학습문서에서 추출하며 유사어 사전에 첨가된 키워드들은 이미 할당된 동일 의미의 범주 자질이 가지는 가중치와 동일한 수치벡터를 가진다.

문서표현(색인)은 벡터 공간 모델(Vector Space Model)에 의하여 표현한다. 자질 선택 과정에서 추출된 자질들은 단어의 출현빈도(TF)와 역범주빈도(ICF)에 의해 가중치를 부여한다. 이는 문서 분류기를 통해 문서를 분류할 때 범주별 자질의 벡터값으로 사용하기 위한 것이다.

본 고에서는 제안한 방법의 성능을 평가하기 위하여 현재 자동 문서 범주화 분야에서 사용되고 있는 선형회귀모형(Linear Regression Model)을 사용하였다.

2 범주별 자질 유사어 사전을 이용한 자동문서 범주화 실험 설계

실험에 사용할 데이터은 모회사의 VOC(Voice of Comsumer)데이터 25만 건의 데이터베이스를 대상으로 하였다. 먼저 분류하고자 하는 범주는 '물류'를 비롯하여 10개 영역으로 분류하였다. 그리고 주제전문가에 의하여 수작업으로 부여된 키워드 사전을 사용하여 25만 건에 대한 사전 자동분류 작업을 수행하였다. 그 결과 추출된 범주별 문서 중에서 범주별 상위 20개씩 분석대상 자료를 200건을 추출하였다. 주제전문가들에 의해 범주별로 적합문서 판정 작업을 수행하여 분류 정확도를 측정하였다. 그리고 기 분류된 실험문서집합 200건에 대하여 본 연구자가 구축한 유사어들을 부여한 후 분류 정확도를 측정하여 자동분류알고리즘에 대한 성능평가를 수행하였다. 그리고 본 실험은 실험데이터의 중복 범주할당을 인정한다.

2.1 전처리과정

수집된 문서를 본 시스템에서 사용하기 위해서 우선 기계적 처리가 가능하도록 변환한다. 문서의 내용이나 특징을 잘 반영하는 단어를 내용어라고 한다. 이런 내용어를 추출하기 위해서 먼저 형태소 분석기를 사용하여 문장을 각 형태소 별로 나누어 품사를 결정한다. 본 시스템은 품사 중에서 명사만을 추출하여 내용어로 사용한다.

명사는 개념을 도입하고 설명하는데 쓰이므로

내용어로 가장 많이 등장하는 중요한 품사이다. 기존 연구결과를 보면, 자동 분류를 주제로 한 논문 대부분이 명사만을 규칙에 넣어서 자동분류를 하고 있다. 아울러 텍스트 마이닝에서도 명사이외의 품사를 규칙에 포함할 경우 성능이 떨어진다는 연구 결과가 있다(김재훈 2000).

2.2 범주별 자질 추출

범주별 자질 추출은 문서 범주를 대표할 수 있는 자질을 추출하는 과정이다. 본 시스템에서는 우선 주제전문가 집단에 의해 이미 추출된 키워드 사전에 카이 제곱 통계량(X^2 Statistic)을 이용하여 범주별 적합문서를 대상으로 후보 자질들을 추출하여 상위 5%로 자질 축소를 실시하여 범주별 자질 지식베이스를 구축하였다. [표 3]은 범주별 자질의 예를 나타낸 것이다.

[표 3] 범주별 자질의 예

범 주	자질수	자 질
물류	12	배달, 택배, 착하불량, 설치지연, 설치기사, 배달기사, 물류, 타모델 배달, 운송, 운반, 해외운송, 운반중
영업	31	판매, 대리점, 판매원, 판매점, 매장, 점원, 가격, 광고, 선전, 홍보, 이벤트, 행사, 할인, 프로모션, 판촉, 세일, 쇼핑몰, 마케팅, 방문, 보상판매, 영업, 영업시간, 출하, 구입가, 구입처, 할부, 생산, 단종, 대량구매, 구매, 구입
서비스	17	애프터서비스, A/S, 수리, 수리내역, 수리비용, AS, 수리불가, 단종, 서비스센터, 출장비, 보증기간, A/S 지연, 환불, 수급지연, 수리방문, 제품분실, 에프터 서비스
제품	14	제품기능, 제품특성, 제품사양, 제품스펙, 제품용량, 제품크기, 상품성능, 상품기능, 상품특성, 상품사양, 상품스펙, 상품용량, 상품크기, 상품성능
불만	49	해명, 이해, 조치, 바가지, 황당, 불만, 불편, 불가능, 부족, 사과, 고장, 노력, 화, 손해, 손상, 상처, 지연, 불친절, 우롱, 허위, 결함, 불량, 반품, 실해, 환불, 답답, 단종, 오류, 가격, 포기, 실망, 손해, 가열, 에러, 고장, 횡포, 악평, 오점, 결여, 상담, 대기, 최악, 시정, 무시, 이상, 상심, 짜증, 엉망, 리콜
요청	17	드라이버, 사용설명서, 설명서, 매뉴얼, 메뉴, 유상, 무상, 교체, 구성품, 가이드, 설치파일, 설치프로그램, 자료, 요청, 부탁, 기사, 다운
칭찬	14	칭찬, 감동, 감탄, 훌륭, 찬사
정보	12	경쟁사, LG, 엘지, 엘쥐, SK, KTF, 타회사, 동향, 타사, 경쟁, 시장, 주가경제
제안	7	실험, 개선, 특허, 제휴, 제안, 고안, 아이디어
문의	27	사용방법, 전화번호, 대리점, 위치, 에이에스센터, 변상, 배상, 지점, 출시, 서비스센터, 설치, 구성, 가격, 요금, 기종, 구입, 디자인, 신제품, 제품, 용량, 크기, 배달, 재송, 홈쇼핑, 장단점, 특소세, 문의

2.3 범주별 자질에 대한 자질값 부여

추출된 대표 자질에 가중치를 부여하기 위해서 다음과 같은 용어빈도(TF : Term Frequency)와 역범주빈도(ICF : Inverse Category Frequency)를 사용한다.

각 범주별 학습문서에서 용어 t_i의 용어 빈도(TF)는 아래 식으로 계산한다.

TF_{ij} = j번째 범주에서의 용어 t_i의 출현빈도

범주간의 분리도가 높은 단어에 높은 가중치를 주기 위하여 제안된 기법에서는 역범주 빈도을 사용한다(조광제, 김준태 1997). t_i를 포함하는 범주의 개수는 CF_i이고 총 범주의 개수를 M이라고 할 때 역범주 빈도는 아래 식과 같다.

$$ICF_i = \log(M) - \log(CF_i)$$

위에서 계산된 용어 빈도(TF_{ij})와 역범주 빈도(ICF_j)를 이용해서 용어 t_i의 j번째 범주에서의 가중치 w_{ij}는 다음과 같이 계산된다.

$$w_{ij} = TF_{ij} \times ICF_i = TF_{ij} \times (\log(M) - \log(CF_i))$$

2.4 범주별 자질에 대한 유사어 사전 구축

2.4.1 유사어 사전의 필요성

자동문서범주에 대한 실험에서 문서들의 오분류의 가장 큰 이유는 자질이 추출되지 않음으로 인하여 잘못 구성된 특징 벡터에 기인한다. 자질이 추출되지 않는 이유는 자질이 등록되어 있지 않거나 등록되어 있지만 자질의 형태의 변형이 그 원인이다. 자질의 변형이란 범주 자질과 동일하거나 유사한 의미를 가진 단어가 문서 작성자들의 자질 표현능력의 차이로 인하여 자질 추출의 가능성을 반감시키는 것을 말한다. 문서 작성자들에게 있어 자질 표현능력은 작성자들의 개별적 전문성에 따라 상당히 차이가 난다. 이러한 문서 작성자들의 차별적 자질 표현능력을 일반화 및 제고시키기 위한 방법으로 자질 변형에 대한 유사어 사전을 구축하여야 한다. 이를 위해서 범주별 자질 지식베이스에 유사어 사전을 추가적으로 첨가하여 새로운 문서가 삽입되었을 때 새문서에 포함된 단어가 통합지식베이스(자질DB + 유사어사전)를 통해 자질로 전환될 가능성을 높일 수 있다. 즉 자질의 변형으로 인해 자질 가중치가 잘못 배정되거나 미배정됨으로 인해 발생할 수 있는 새 문서의 오분류 원인을 감소시킴으로써 분류 성능을 향상시킬 수 있다.

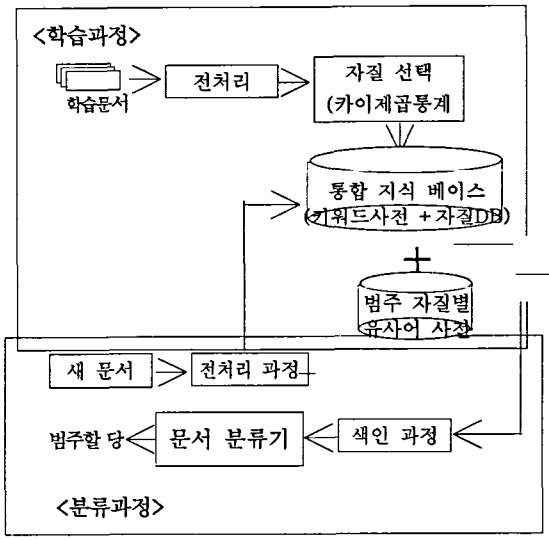
2.4.2 유사어 사전 구축

유사어 사전을 구축하기 위하여 학습과정을 통해 오분류된 문서들을 사용한다. 오분류된 학습문서들에 포함된 자질들은 주제전문가들에 의해 배정된 원범주에 속하는 문서들의 자질과 동일한 의미나 유사한 의미를 가짐에도 불구하고 문서 작성자의 자질 표현의 차이로 인하여 추출되지 못하였다고 가정하고 오분류된 문서 집합의 자질을 원래 범주의 자질과 비교하여 유사어를 수집하였다. [표 4]는 일부를 보여준다.

[표 4] 범주 자질에 대한 유사어 일부 예

범주	유사어의 예
물류	[배송, 전달, 지연설치, A/S기사 운송중.....]
영업	[판매센터, 판매점, 매장직원, 출고, 출원, 출시.....]
서비스	[a/s, 애프터써비스, 수리비, 설치비, 교환.....]
제품	[제품규격, 상품규격.....]
불만	[즉답, 결함, 장애, 문제, 증상, 에러, 실패, 오작동, 오작동, 경고.....]
요청	[다운로드, download.....]
칭찬	[감명, 감격, 짱, 굳, 감사,,.....]
정보	[SKT, LGT, 019.....]
제안	[개선, 개선사항, 반짝.....]
문의	[질문, 의문, 사용법, 이용방법, 이용법, 출고, 출하,서비스 센타, 고객센터, 써비스 센터, A/S센터, 설치방법, 설정, 제품가격, 제품가격, 상담, 문의내용.....]

본 실험에서 제안한 시스템의 전체적인 구성은 그림으로 나타내면 다음과 같다.



[그림 1] 제안한 시스템의 전체 구성도

2.4.3 유사어 사전의 기대효과

(1) 범주별 자질 발췌능력 제고
(2) 문서 작성자들의 자질 표현능력의 개별적 차이 해소
(3) 자동 문서 범주화의 오분류률 감소
(4) 자동 문서 범주화의 정확률과 재현률 개선

3 실험 결과와 분석

본 실험은 범주별 자질의 유사어 사전이 문서 분류 정확도를 향상되는 것을 발견하고자 하였다. [표 5]는 유사어 사전을 부여하기 전과 유사어 사전을 부여한 후의 성능평가를 비교한 것이다. 유사어 사전을 부여한 후 평균 문서 분류 정확도는 2.5%의 상승하였다. 그러나 범주별 성능을 비교할 경우 '영업'범주와 '제품'범주는 유사어 사전을 부여하기 전과 동일한 성능이 나왔는데 이는 유사어 사전에 포함된 단어들이 범주별 자질과 유사한 의미를 가지기는 하나 사용빈도가 그리 높지 않은 것들이 대부분이기 때문에 성능향상에 영향을 주지 못하는 것으로 분석된다. '서비스'범주에서는 유사어 사전을 첨가함으로써 10%의 성능향상이 있었다. '불만'범주, '제안'범주, '칭찬'범주, '정보'범주들에서는 유사어 사전을 부여했을 경우 오히려 분류 성능을 떨어지게 하는 결과가 나왔다. 이는 보다 정교하게 정련된 유사어 사전이 검색효율에 유의한 영향을 준다는 것을 확인시켜 주는 결과이다.

[표 5] 유사어 사전 부여 전 / 후 성능비교

범 주	유사어 미부여시 분류 정확도(%)	유사어 부여시 분류 정확도(%)
물류	15%	33%
영업	40%	40%
서비스	50%	60%
제품	20%	20%
불만	75%	66%
요구	30%	50%
칭찬	70%	60%
정보	65%	62%
제안	70%	60%
문의	65%	74%
평 균	**50%**	**52.5%**

4 결 론

기존의 문서분류는 자질에 대한 정보량을 계산하여 그 순위에 따라 상위 자질들을 지식베이스로 사용하여 새로운 문서가 들어왔을 때 이를 기반으로 색인을 하였기 때문에 자질 지식베이스와 일치하지 않는 자질은 추출되지 못하는 한계를 해결하기 위해서 본 실험에서는 분류될 문서의 특징을 나타내는 범주별 자질과 유사한 의미를 가지나 형태가 변형되어 기술된 단어에 대하여 유사어 사전을 구축하여 성능평가 실험을 하였다.

분석 결과 범주별 대표 자질에 대한 유사어 사전을 부여하였을 때 큰 분류성능 차이가 나타난 것은 아니지만 전체적으로 유사어 사전을 부여하기 전보다 문서 분류 정확도가 2.5%로 향상되었다. 그러나 각각의 범주별로 보면 분

류 정확도에서 유사어 사전을 부여하기 전과 분류성능이 거의 동일한 정확도를 보였으며 절반정도의 범주에서는 유사어를 부여하였을 때 오히려 분류 정확도가 떨어지는 결과를 보여졌다.

본 실험의 결과를 통해 볼 때 보다 정교하게 정련된 유사어 사전이 앞으로의 실험을 통해 구축되어야 함을 알 수 있었다. 다시 말해 정련화는 검색효율에 유의한 영향을 미칠 수 있다 것을 확인할 수 있었다.

참고문헌

Y. Yang and J. O. Pederson, 1997. "A Comparative study on feature selection in text categorization." In Proceedings of the 14th International conference on Machine Learning.

G. Salton, E. A. Fox and H. Wu. 1983. " Extended boolean information retrieval. " Communications of the ACM, Vol. 26, No. 12, pp. 1022-1036

김재훈. 2000. "도합유사도를 이용한 한국어 추출문서 요약" 한국해양대,

안영훈, 서정연. 2002. "자동 문서 범주화 기법을 이용한 질의 유형 분류 시스템" 석사학위논문, 서강대학교 대학원, 컴퓨터학과.

박진우, 서정연. 2001. "문장 중요도를 이용한 자동 문서 범주화". 석사학위논문, 서강대학교 대학원, 컴퓨터학과.

고영중, 서정연. 2002. "문서관리를 위한 자동문서범주화에 대한 이론 및 기법" 정보관리연구, Vol. 33. No. 2. pp. 19-32

조광제, 김준태. 1997. "역카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류", 한국 정보과학회 봄 학술발표논문집 (B), pp. 73-79

황재영, 이응봉. 2003. "자동문헌분류를 위한 대표색인어 추출에 관한 연구" 한국정보관리학회 하계 학술발표논문집, pp. 55-64