용어의 문맥활용을 통한 문헌 자동 분류의 성능 향상에 관한 연구

A Study on Improving the Performance of Document Classification Using the Context of Terms

송성전(Sung-Jeon Song)*
정영미(Young-Mee Chung)**

초 록

자동 분류에서 문헌을 표현하는 일반적인 방식인 BOW는 용어를 독립적으로 처리하기 때문에 주변 문맥을 반영하지 못한다는 한계가 있다. 이에 본 연구는 각 용어마다 주제범주별 문맥적 특징을 파악해 프로파일로 정의하고, 이 프로파일과 실제 문헌에서의 문맥을 비교하는 과정을 통해 동일한 형태의 용어라도 그 의미나 주제적 배경에 따라 구분하고자 하였다. 이를 통해 주제가 서로 다름에도 불구하고 특정 용어의 출현만으로 잘못된 분류 판정을 하는 문제를 극복하고자 하였다. 본 연구에서는 이러한 문맥적 요소를 용어 가중치, 분류기 결합, 자질선정의 3가지 항목에 적용해 보고 그 분류 성능을 측정했다. 그 결과, 세 경우 모두 베이스라인보다 분류 성능이 향상되었고 가장 큰 성능 향상을 보인 것은 분류기 결합이었다. 또한 제안한 방법은 학습문헌 수가 많고 적음에 따라 발생하는 성능의 편향을 완화하는데도 효과적인 것으로 나타났다.

ABSTRACT

One of the limitations of BOW method is that each term is recognized only by its form, failing to represent the term's meaning or thematic background. To overcome the limitation, different profiles for each term were defined by thematic categories depending on contextual characteristics. In this study, a specific term was used as a classification feature based on its meaning or thematic background through the process of comparing the context in those profiles with the occurrences in an actual document. The experiment was conducted in three phases; term weighting, ensemble classifier implementation, and feature selection. The classification performance was enhanced in all the phases with the ensemble classifier showing the highest performance score. Also, the outcome showed that the proposed method was effective in reducing the performance bias caused by the total number of learning documents.

키워드: 자동분류, 문맥프로파일, 용어가중치, 분류기 결합, 자질선정
document classification, context profile, term weighting, ensemble classifier, feature selection

* 연세대학교 문헌정보학과 대학원(hellossong@naver.com) (제1저자)
** 연세대학교 문헌정보학과 명예교수(ymchung@yonsei.ac.kr) (공동저자)

■ 논문접수일자 : 2012년 5월 30일 ■ 최초심사일자 : 2012년 5월 30일 ■ 게재확정일자 : 2012년 6월 26일
■ 정보관리학회지, 29(2), 205-224, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.2.205〕

1. 서 론

최근 컴퓨터의 보급과 인터넷의 발달로 생산되는 전자문헌의 수가 기하급수적으로 늘어나게 되었다. 이 문헌들을 사람이 수작업으로 가공하여 처리하기에는 인력과 시간이 많이 소요된다. 이에 따라 시스템에서 자동적으로 문헌을 분류하는 자동 분류에 대한 중요성이 부각되기 시작했다. 이러한 현상은 현재까지도 지속되어 더 정확한 분류 결과와 실생활에 활용 가능한 분류 시스템을 만들기 위한 연구가 이루어지고 있다.

텍스트 형태의 문헌을 자동 분류하기 위해서는 먼저 분류 알고리즘을 적용할 수 있는 형태로 문헌을 표현해야 하는데 이는 자동 분류 분야뿐만 아니라 문헌에 수치 계산을 적용하는 모든 경우에 필요한 과정이다. 일반적으로 자동 분류에서는 문헌을 하나의 벡터 형태로 표현한다. 이 벡터는 말뭉치(corpus)에 출현한 용어의 종류만큼 요소(element)를 갖게 되는데 벡터의 요소 하나는 용어 하나와 대응되고, 요소의 값은 해당 용어의 가중치가 적용된다. 이러한 방식의 문헌표현(document representation)을 BOW(bag-of-words)라고 한다. 이 BOW 방식은 비교적 간단하게 문헌을 표현할 수 있는 효과적인 방법으로 평가받고 있지만 문헌에 표현된 정보를 모두 반영하지는 못한다는 한계점이 지적되고 있다(Gabrilovich & Markovitch, 2009; Wang & Domeniconi, 2008).

그 중 하나가 용어의 문맥을 통해 얻을 수 있는 정보가 무시된다는 점으로 특정 용어와 그 앞이나 뒤에 오는 용어들을 함께 고려했을 때 얻을 수 있는 정보를 반영하지 못한다. 가령, '운동'이라는 용어가 한 문헌에서 '근육', '자세'와 함께 사용되었다면, 이때의 '운동'은 체육학과 관련된 것임을 알 수 있고, '질량', '법칙'과 함께 사용되었을 때는 물리학과 관련해 사용된 것을 알 수 있다. 비록 '운동'이라는 용어가 동일하게 출현하였지만 문헌에 출현한 다른 주변 용어들을 통해 주제적으로 다른 상황에서 사용되었다는 것을 파악할 수 있게 되는 것이다. 하지만 BOW 방식은 기본적으로 각 용어를 독립적으로 취급하는 형태이기 때문에 두 문헌에서 사용된 '운동'이라는 용어가 주제적으로 비슷한 상황에서 출현하였는지 그렇지 않은지 파악할 수 없다. 단지, 두 문헌 간 공통적으로 출현하는 용어의 가중치에 따라 두 문헌의 유사성을 산출할 뿐이다. 이러한 점은 문헌의 자동 분류에도 적용되는 문제이다. 많은 자동 분류 기법이 있지만 기본적인 형태는 생성된 분류기와 입력 문헌 간 유사도를 산출해 해당 주제범주로 배정할 것인지를 판단하는 것이다. 이 때, 대부분의 기법이 벡터기반의 유사도 산출 기법을 적용하기 때문에 동일한 한계점이 발생한다.

이에 본 연구에서는 이러한 한계점을 개선할 수 있는 방법을 고안하고, 이를 바탕으로 자동 분류 실험을 하고자 한다. 본 연구의 목적은 첫째로, 주목한 한계점을 개선했을 때 문헌의 자동 분류 성능이 향상되는지 파악하는 것이다. 이를 통해 한계점이 자동 분류 성능 저하에 많은 영향을 미치고 있는지 여부를 파악할 수 있을 것이고, 더불어 자동 분류에서 문맥을 활용하는 것 자체가 효용성이 있는지도 확인할 수 있을 것이다. 둘째로, 문맥을 활용하는 것이 자동 분류에 효용성이 있다고 할 때, 자동 분류 과정에 적용할 수 있는 다양한 방법을 생각해 보고 가장 큰

성능 향상을 이끌어 주는 방법을 찾는 것이다. 이를 통해 단지 문맥 활용이 효과가 있다는 점을 넘어 자동 분류 과정 중 어떤 부분에 적용하는 것이 가장 효과적인지 판단할 수 있을 것이다.

문맥을 통해 특정 용어가 사용된 주제적 배경을 파악하기 위해서는 사전에 각 주제별로 특정 용어에 대한 문맥적 특징이 정의되어야 한다. 특정 문헌 내에서의 문맥적 특징과 주제별로 정의된 문맥적 특징을 비교함으로써 주제적 배경을 파악할 수 있기 때문이다. 이를 위해 본 연구에서는 학습문헌 집합을 통해 '주제별 문맥 프로파일'이라는 명칭으로 특정 용어에 대한 각 주제범주의 문맥적 특징을 정의하였다. '주제별 문맥 프로파일'을 외부자원을 사용해 구성하지 않고 학습문헌 집합을 사용한 이유는 실험문헌과 동일한 정보원으로부터 생산된 문헌들이기 때문에 사용된 용어의 종류나 분포가 외부자원보다는 더 유사할 것이기 때문이다. 또한 외부자원을 사용해 분류자질을 추가하는 형태의 연구는 이미 많이 이루어진데다 본 연구는 분류자질의 수를 추가하는 등의 물리적 변형을 하지 않은 동일한 텍스트에서도 더 좋은 분류 성능을 얻고자 하기 때문이다. 생성된 문맥 프로파일의 문맥 비교를 통해 얻은 결과는 용어 가중치, 분류기 결합, 자질선정의 3가지 요소에 반영하여 자동 분류 실험을 실시해 보았고, 각 경우의 자동 분류 성능은 아무런 처리를 하지 않은 베이스라인과 비교하였다.

2. 선행연구

자동 분류와 관련한 다양한 연구들 중 문헌을 표현하는 방식에 문제를 제기하는 연구가 많이 이루어지고 있다. 이들 대부분의 연구는 BOW 방식이 가지는 한계를 극복하려는 시도이다.

먼저, Sable과 McKeown, Church(2002)는 품사태깅을 통해 문장 내 주어와 술어를 추출하여 이를 각 주제범주의 주어, 술어와 비교하는 과정을 통해 단순히 독립된 용어가 아닌 언어학적 관계를 자동 분류에 적용하고자 하였다. 이 접근방법은 언어학적으로 주어, 술어 정보가 중요하다고 판단한 것이지만 관계의 설정을 품사태깅에만 의존했다는 점에서 다소 제한적이었다. 또한, Caropreso와 Matwin, Sebastiani(2001)은 n-gram 색인기법이라는 구조적 접근방법을 통해 하나의 단어가 아닌 2개 이상의 단어로 구성된 요소를 자동 분류에 적용하기도 하였다. 하지만 이 경우 모든 문헌을 n-gram으로 나누기 때문에 의미적 요소와 일치하지 않을 수 있고, 각 요소가 무엇을 의미하는지 파악하기 어렵다는 단점이 있었다.

반면, 사람이 직접 구성한 외부 지식자원을 통해 한계점을 극복하려는 연구들도 수행되었는데 주로 *WordNet*과 *Wikipedia*를 외부 지식자원으로 이용하였다. Rodriguez와 Hidalgo, Agudo(1997)는 주제범주에 따라 학습문헌 수가 달라 발생할 수 있는 용어부족을 *WordNet*을 통해 보충해 주는 연구를 진행하였다. 이는 *WordNet*을 자동 분류 과정에 성공적으로 적용한 초기 연구로 *WordNet*을 적용했을 때, Rocchio 알고리즘과 Widrow-Hoff 알고리즘에서 더 좋은 성능을 보이는 것으로 나타났다. 정은경(2009)도 *WordNet*을 이용하여 문헌의 제목, 키워드, 분류어 등의 동의어를 분류자질로 추가하는 형태

의 연구를 진행하였다. 특히, *WordNet*의 synset과 문헌과의 유사성 비교를 통해 선택적으로 추가하는 방법을 적용하기도 하였다. 그 결과, 분류 성능이 월등하게 향상된 것으로 나타났다. Wang과 Domeniconi(2008)는 *Wikipedia*를 이용해 기존의 문헌표현의 한계점을 극복하려 하였다. 2개 이상의 단어로 구성된 단어구 형태를 자질로 사용하기 위해 문헌에 출현한 단어들을 어구 형태로 조합하고, 조합된 단어구가 *Wikipedia*에 존재하는 것은 기존의 단어를 단위가 아닌 단어구 단위로 변환하였다. 또한 다의어와 동의어 문제를 개선하기 위해 *Wikipedia*의 문헌들로부터 용어의 인접행렬을 구성해 해당 용어와 관련 있는 용어들을 추가하였다. Gabrilovich와 Markovitch(2009)도 *Wikipedia*를 이용해 문헌표현을 개선하고자 하는 연구를 진행하였는데 ESA(Explicit Sematic Analysis)라는 분석방법을 제안하였다. ESA는 *Wikipedia*로부터 자동적으로 시소러스를 구축하여 이를 활용하는 것으로 기존의 벡터에 관련 개념을 추가하는 방식으로 자동 분류에 적용되었다. 특히, 한 문헌을 단어, 문장, 문단, 문헌 전체의 단위의 지역적 문맥으로 나누어 *Wikipedia*와의 비교하는 과정을 반복해 공통적인 개념을 추출하였다. 추출된 개념들은 문헌에 추가되었고, 자질선정 과정을 거친 후 문헌을 자동으로 분류하였다.

이외에 각 색인어를 분류자질로 사용하지 않고 문헌의 벡터 전체를 자질로 사용한 연구(이재윤, 2005), 의견어를 설정하고 이를 잠재적의미색인(Latent Sematic Indexing) 기법을 통해 의견 문서를 분류한 연구(이지혜, 정영미, 2009), 이용한 문헌의 구조적 분석을 통해 용어 간의 관계를 추출하고 이를 용어 가중치에 적용한 연구(Huynh et al., 2011) 등이 있다.

3. 실험 설계

3.1 실험 개요

본 연구의 목적은 문맥적 요소를 자동 분류에 반영하여 분류 성능을 향상시키는 것으로 주제범주에 따른 용어별 문맥을 정의한 후 실험문헌의 문맥과 비교하여 문맥적으로 유사한 주제범주를 파악하게 된다. 구체적으로 분류과정에 반영하는 방법으로는 동적 가중치를 통해 가중치를 조정하는 방법과 문맥 적합성 점수를 기존 분류기 결과에 결합하는 방법, 그리고 주제범주에 따른 문맥의 상이성을 고려한 자질선정의 3가지 방법을 적용하였다. 각각의 방법에 문맥을 적용했을 때 기대하는 효과는 다음과 같다.

첫째, 동적 가중치는 동일한 형태의 용어라도 문맥에 따라 다른 가중치 값을 적용하기 위한 것으로 기존의 가중치는 문맥에 관계없이 용어에 가중치를 부여했기 때문에 문맥을 통한 용어 가중치의 정교화는 가중치로 인한 자동분류의 잡음 문제를 완화할 수 있을 것으로 기대한다. 둘째, 문맥 적합성 점수를 분류기 점수와 결합하는 것은 문맥적 관점이 무시되는 기존 분류기의 분류 결과를 일정 부분 보완할 수 있는 방법으로 기존의 분류기 분류점수와 문맥의 관점으로만 판단한 분류점수를 결합하기 때문에 문맥을 반영한 분류결과를 얻을 수 있을 것이다. 셋째, 자질선정에 있어서 문맥을 고려

하면 주제범주에 따라 문맥적 특징이 서로 상이한 용어를 식별할 수 있으며, 이를 통해 식별력이 큰 자질을 선정할 수 있을 것이다.

본 연구의 자동 분류 실험은 〈그림 1〉과 같이 다섯 단계로 구성된다.

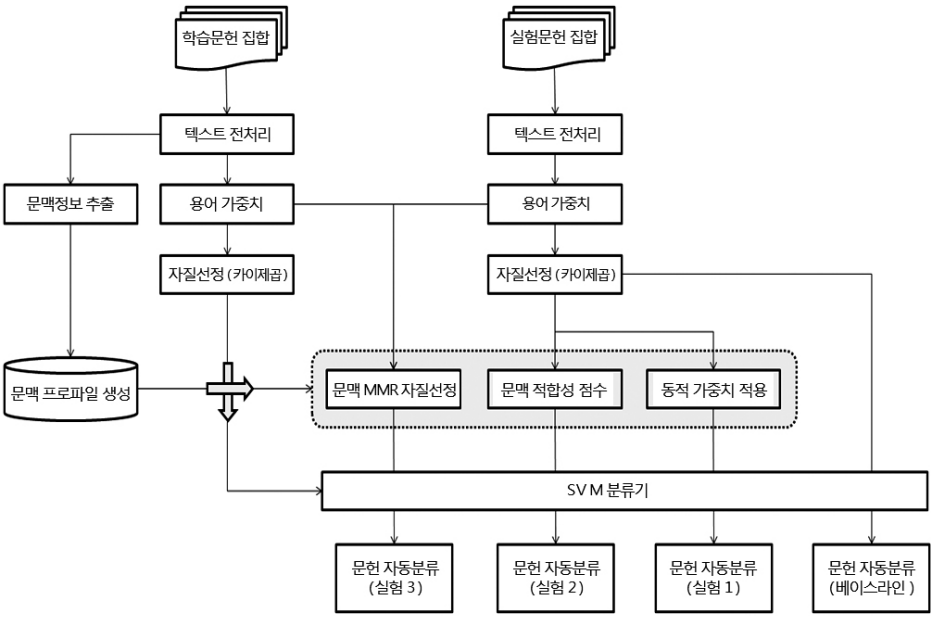
첫 번째 단계는 학습문헌 집합을 통해 주제에 따른 문맥 프로파일을 구성하는 것으로 특정 용어에 대해 각 주제범주에 따라 문맥적으로 함께 출현하는 용어들의 종류를 파악해 프로파일 형태로 정의하게 된다.

두 번째 단계는 비교대상이 되는 베이스라인의 실험으로 기존의 자동 분류 과정을 적용했을 때의 분류 성능을 파악한다. 베이스라인은 가장 좋은 성능을 보여주는 것을 적용할 필요가 있기 때문에 각종 용어 가중치 기법에 따라 그 분류 성능을 측정하고, 그 중 가장 성능이 좋은 것을 베이스라인으로 설정하게 된다.

세 번째 단계는 동적 가중치를 적용한 분류 실험이다. 첫 번째 단계에서 사전에 구성한 문맥 프로파일을 활용해 문맥에 따라 문헌의 용어 가중치 값을 조정하게 된다. 그리고 조정된 가중치를 바탕으로 자동 분류를 수행하여 산출된 분류 성능은 베이스라인과 비교하여 동적 가중치를 적용했을 때의 성능 변화를 확인한다.

네 번째 단계는 문맥 적합성 점수를 적용한 분류 실험이다. 문맥적 요소를 자동 분류에 반영하는 두 번째 방법으로 베이스라인의 분류 결과에 문맥 적합성 점수를 합산하여 최종적으로 문헌을 분류하고, 그 때의 성능을 베이스라인의 성능과 비교한다.

마지막 단계는 문맥 상이성을 통한 자질선정을 적용한 분류 실험이다. 이 단계는 자질 선정 시, 문맥적 요소를 반영하는 것으로 기존의 자질선정 방법과 문맥 프로파일 간 유사도를



〈그림 1〉 실험 개요

MMR(Maximal Maginal Relevance)로 결합해 최종 자질을 선정하게 된다. 이러한 과정을 통해 얻은 각 실험의 분류 성능은 베이스라인의 분류 성능과 비교하여 그 효용성을 평가하게 되는데 평가척도로는 정확률, 재현율, F_1 척도를 적용하게 된다.

본 연구에서 사용한 실험집단은 Reuters-21578(David, 2004)로 문헌 수가 많은 상위 10개의 주제범주를 대상으로 하였다. 분류자질은 카이제곱(chi-square) 기법을 통해 20%로 축소하여 총 3,422개의 분류자질을 가지고 실험을 진행하였고, 분류기는 선형 SVM 분류기인 SVM^{light}[1]을 이용하여 문헌을 분류하였다. Reuters-21578은 하나의 문헌이 2개 이상의 주제범주에 할당될 수 있는 형태(multi-label)이기 때문에 각 주제범주마다 이원판정을 할 수 있는 분류기를 각각 생성해 그 성능을 측정하였다. 단어의 어간 추출을 위해 Porter 스태머(Porter, 1980)를 사용하였고, 불용어 제거는 ProQuest에서 제공하는 불용어 리스트를 적용하였다.

3.2 실험집단

본 실험에서 사용한 Reuters-21578은 자동분류 분야에서 사용되는 대표적인 컬렉션으로 1987년도 로이터 뉴스 데이터 중 일부이다(David, 2004). 전체 21,578개의 문헌 중 문헌 수가 많은 상위 10개의 주제범주에 속하는 문헌들이 주로 자동 분류 실험에 활용되는데 본 연구에서도 상위 10개의 범주만을 대상으로 하였다. 그보다 하위 범주는 실험문헌의 수가 50개도 채 되지 않을 정도로 적어서 실험결과의 신뢰성을 주지 못하기 때문이다. 10개 주제범주와 문헌 수는 〈표 1〉과 같다.

각 문헌은 SGML 형태로 표현되어 여러 종류의 내용을 담고 있는데 본 연구에서 문헌의 텍스트는 TITLE과 BODY 태그에 해당하는 내용만을 대상으로 하였다. 학습/실험문헌의 구분은 ModApte 분할 방식을 통해 7,193개의 학습문헌과 2,787개의 실험문헌으로 분할하였다.

〈표 1〉 Reuters-21578 상위 10개 범주와 문헌 수

범주명	학습문헌 수	실험문헌 수	총 문헌 수
acq(acquisition)	1,650	719	2,369
corn	181	56	237
crude	389	189	578
earn	2,877	1,087	3,964
grain	433	149	582
interest	347	131	478
money-fx(money foreign exchange)	538	179	717
ship	197	89	286
trade	369	117	486
wheat	212	71	283
합계	7,193	2,787	9,980

1) SVM^{light} (http://svmlight.joachims.org/).

3.3 주제범주별 문맥 프로파일 구성

문맥을 통해 특정 용어 t_k가 어떠한 주제적 맥락에서 사용된 것인지 파악하기 위해서는 먼저, 해당 용어 t_k의 문맥이 주제적 상황에 따라 가지는 특징을 알고 있어야 한다. 이 문맥적 특징을 알고 있을 때, 특정 문헌에서 나타나는 문맥과 비교가 가능하기 때문이다. 이를 위해 각 용어에 대해 주제에 따른 문맥 프로파일(context profile)을 구성하였다. 문맥 프로파일은 주제에 따라 주로 함께 사용되는 문맥용어들의 집합(pool)으로 주제에 따라 이 문맥용어의 종류가 달라지는 특징을 이용해 특정 용어 t_k의 주제적 상황을 파악하게 된다.

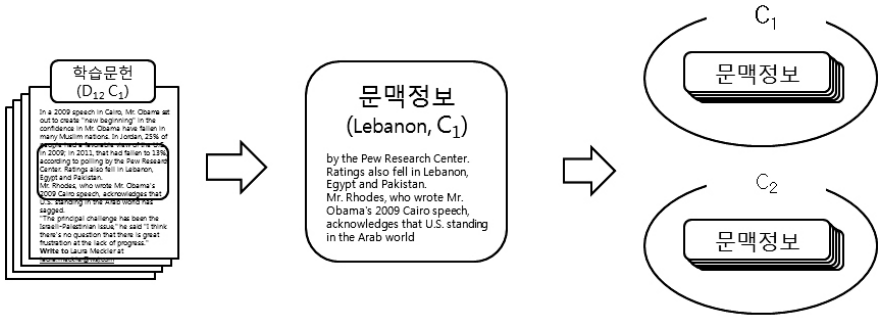
용어의 주제별 문맥 프로파일은 용어에 따라 주제범주별로 생성되기 때문에 하나의 용어에 대해 주제범주의 수만큼 문맥 프로파일이 생성되게 된다. 문맥 프로파일은 다음과 같은 문맥정보 수집 및 분류, 문맥용어 가중치 부여, 상위 n개 문맥용어 선정의 3가지 과정을 통해 생성된다.

3.3.1 문맥정보 수집 및 분류

문맥정보는 특정 용어가 사용된 문맥을 개별 문헌단위로 추출한 것으로 학습문헌 중 용어 t_k가 출현한 문헌들에서 이 문맥정보를 추출하게 된다. 문맥정보를 추출할 때, 학습문헌의 주제범주 정보 역시 함께 추출되는데 이는 해당 문맥정보를 주제에 따라 구분하기 위함이다. 가령, 학습문헌이 속하는 주제범주가 'money-fx' 범주라면, 해당 학습문헌에서 추출된 문맥정보 역시 'money-fx'의 주제적 상황에서 사용되는 문맥이라고 규정하는 것으로 각 문맥정보에 주제정보를 부여하는 효과적인 방법이 된다. 이렇게 수집된 문맥정보는 주제범주에 따라 분류된다. 이와 같은 과정을 그림으로 표현하면 〈그림 2〉와 같다.

3.3.2 문맥용어의 가중치 부여

문맥용어는 문맥정보에 출현한 용어들을 지칭하는 것으로 특정 용어 t_k와 이에 대한 문맥으로 사용되는 용어 사이를 구분하기 위해 문맥용어라는 명칭을 사용하였다. 각 주제범주별로 수집된 문맥정보를 통해 해당 주제범주를 대표하는 문맥 프로파일을 생성하기 위해서는 문맥정보에서 주로 출현하는 문맥용어 위주로 작성되어야 한다. 따라서 수집된 문맥정보에 출현한 문



〈그림 2〉 문맥정보 추출과정

맥용어마다 동시출현빈도($co-occurrence(t_k,t_i)$)를 해당 주제범주의 문헌 수(n_c)로 정규화 한 다음 공식의 가중치(w_i)를 부여하였다.

$$w_i = \frac{co-occurrence(t_k,t_i)}{n_c}$$

$co-occurrence(t_k,t_i)$: 용어 t_k의 문맥정보 중 용어 t_i가 출현한 횟수

n_c: 해당 주제범주의 총 문헌 수

$co-occurrence(t_k,t_i)$는 문맥 프로파일을 구성하고자 하는 특정 용어인 t_k와 이에 대한 문맥용어인 t_i가 동시에 출현한 문맥정보의 수로 모든 문맥정보에는 t_k가 출현하기 때문에 단순히 수집된 문맥정보에 t_i가 출현한 문헌빈도와 같다. n_c는 해당 주제범주의 총 문헌 수로 문헌 수가 많을수록 동시출현 빈도는 증가하기 때문에 주제범주의 문헌 수로 나누어 정규화 한 것이다.

3.3.3 상위 n개 문맥용어 선정

위의 가중치 부여를 통해 해당 주제범주에서 특정 용어 t_k와 문맥적으로 관련이 깊은 정도를 수치로 표현한다. 이 가중치 값을 기준으로 문맥용어를 순위화 하고, 상위 n개만 선정하여 문맥 프로파일을 구성하게 된다. 상위 n개만 선정하는 이유는 주제범주에 따라 문맥용어의 수도 차이가 있을 수 있기 때문으로 문맥용어의 수는 학습문헌 수나 문헌길이의 따라 주제범주 간 차이가 발생할 수 있다. 상위 n개를 선정하지 않고, 모든 문맥용어를 문맥 프로파일에 사용한다면 학습문헌 수가 많은 주제범주의 문맥 프로파일이 문맥용어의 수가 더 많기 때문에 무조건적으로 더 문맥적으로 유사하다고 판정될 확률이 더 높아지는 문제가 발생한다.

3.4 동적 용어 가중치 적용

동적 용어 가중치는 문맥에 따라 동일한 용어라 하더라도 가중치 값이 다르게 적용하고자 하는 것으로 문맥적 비교를 통해 분류 과정에서 잡음으로 작용할 여지가 큰 용어는 그 값을 작게 설정하여 상대적인 영향력을 감소시키고, 문맥적으로 좋은 식별력을 가지는 용어는 그 값을 상대적으로 크게 하는 효과가 있다.

동적 가중치는 다음의 3가지 과정을 통해 가중치가 부여된다. (1) 실험문헌에서 특정 용어 t_k에 대한 문맥을 추출한다. 이때 문맥이란 해당 문헌에서 특정 용어 t_k 주변에 출현한 다른 용어들을 의미한다. (2) 실험문헌에서 추출한 문맥과 문맥 프로파일 간 유사도를 산출해 문맥적으로 가장 근접한 주제범주를 선정하게 된다. 본 실험에서는 내적계수를 통해 유사도를 산출하였다. (3) 문맥적으로 근접한 주제범주를 찾은 후, 그 주제범주에서 해당 용어가 가지는 분포를 가중치로 표현해 기존의 용어 가중치와 결합하는 형태로 아래의 공식을 적용한다.

$$w_j = (w_o + k) \cdot \frac{df(c,t_i)}{n_c}$$

w_o: 기존 용어 가중치

k: 기존 가중치의 보정 값

$df(c,t_i)$: 주제범주 c에서 용어 t_i가 출현한 문헌 수

n_c: 주제범주 c의 총 문헌 수

위 공식은 기본적으로 특정 문헌에서의 용어

t_i의 문맥과 사전에 정의한 용어 t_i의 문맥 프로파일 간 유사도 비교를 통해 가장 높은 유사도를 보인 문맥 프로파일의 주제범주인 c를 기준으로 한다. $df(c,t_i)$는 이 주제범주 c에서의 용어 t_i의 문헌빈도로 이 수치는 용어 t_i가 주제 c의 문맥적 상황에서 얼마나 자주 출현하는지를 나타낸다. 이 수치의 값이 높다는 것은 용어 t_i가 해당 문맥에서는 자주 출현하는 중요한 용어이고, 반대로 이 문헌빈도 수치가 낮다면 문헌에 이 용어 t_i가 출현하기는 하였지만 해당 문맥에서는 일반적으로 사용이 많지 않는 중요도가 상대적으로 낮은 용어이다. 특정 용어의 문헌빈도는 문헌 수가 많으면 많을수록 더 큰 값을 가지기 때문에 각 주제범주의 문헌 수인 n_c로 나누어 정규화 하였다.

3.5 문맥 적합성 점수 결합

문맥적 요소를 자동 범주화에 반영하는 방법으로는 앞에서와 같이 동적 가중치를 사용하여 기존의 분류기에 바로 적용하는 방법이 있을 수 있지만 하나의 독립적 분류 방법으로 생각해서 다른 분류기의 분류 점수와 결합할 수도 있다. 즉, 하나의 문헌에 대해 각각의 주제범주와의 문맥 적합성 점수를 산출하고, 이를 정규화 하여 기존의 분류기와 결합하는 형태로도 적용하는 것도 한 방법이 되는 것이다. 문맥 프로파일은 모든 주제범주에 대해 구성되었기 때문에 특정 문헌에 대해 주제범주별 문맥적 점수를 산출할 수 있다.

3.5.1 범주별 문맥 적합성 점수

앞서 동적 가중치를 부여할 때와 같이 실험문헌의 용어에 따른 문맥을 구성하고, 구성된 문맥을 문맥 프로파일과 비교하는 과정을 통해 점수가 산출된다. 실험문헌의 용어 t_i에 문맥정보와 주제범주 c의 문맥 프로파일의 간 적합성 점수 S_i를 산출하는 공식은 다음과 같다.

$$S_i = p(t_i|c) \cdot \sum_{\substack{j \in d \\ j \neq i}} w_{ij}$$

$p(t_i|c)$: 주제범주 c에서 용어 i가 출현할 확률

w_{ij}: 용어 i의 주제범주 c 문맥 프로파일 내 용어 j의 가중치

위 공식은 실험문헌의 주제범주가 c라고 가정하고, 용어 t_i가 출현한 조건에서 실험문헌의 나머지 용어들과 주제범주 c의 문맥 프로파일의 문맥용어와 비교하는 것으로 서로 일치하는 용어들에 대해서 문맥 프로파일의 가중치를 합산하였다. 또한 실험문헌에 출현한 용어 t_i에 대해서도 해당 범주에서 출현할 확률($p(t_i|c)$)을 적용하였는데 이는 해당 주제범주에서 주로 사용되는 용어에 더 큰 가중치를 주기 위함이다. 주제범주 c와 실험문헌 간 적합성 점수는 최종적으로 위의 공식으로 산출한 각 용어의 문맥 적합성 점수를 모두 합산한 수치 $rel(d|c)$를 적용하였다.

$$rel(d|c) = \sum_{i \in d} S_i.$$

3.5.2 SVM 분류기와의 결합

주제범주에 따른 문맥 적합성 점수는 모두 양수라는 특징이 있다. 반면, SVM 분류기를 통한 점수는 음수와 양수로 구분된다. 음의 값

을 가질 때는 해당 주제범주에 부적합하다는 판정이고, 양의 값을 가질 때는 적합으로 판정한다. 이러한 특성에 맞춰 실험문헌의 문맥 적합성 점수도 적은 값일수록 음의 값을 갖는 방식으로 점수를 조정해서 결합해야 한다.

문맥 적합성 점수는 모든 주제범주에 대해 산출이 가능하기 때문에 이 주제범주에 따른 문맥 적합성 점수의 분포를 바탕으로 상대적으로 작은 값을 가지는 주제범주는 음의 값을 갖고, 큰 값을 갖는 주제범주는 양의 값을 갖는 방법을 적용하였다. 또한 값의 크고 작은 정도에 따라서도 그 값의 편차를 두게 하였다. SVM 분류기와 결합은 주제범주별로 구한 문맥 적합성 점수를 아래의 공식으로 결합 가능한 수치($r-score$)로 변형하여 SVM 분류기의 분류 점수와 더하는 형태로 이루어졌다.

$$r-score(d,c)=\alpha(\frac{rel(d|c)}{\max(rel)}-\gamma)$$

α: 값의 범위를 조정하는 매개변수
$rel(d|c)$: 주제범주 c의 문맥 적합성 점수
$\max(rel)$: 모든 주제범주 문맥 적합성 점수 중 가장 큰 값
γ: 음수와 양수를 결정하는 기준 매개변수

위 공식에서 $rel(d|c)$는 특정 주제범주 c와 실험문헌 간 문맥 적합성 점수이고, $\max(rel)$는 모든 주제범주의 문맥 적합성 점수 중 가장 큰 값이다. 그리고 이 둘의 비율인 $\frac{rel(d|c)}{\max(rel)}$를 통해 실험문헌의 길이나 용어의 종류의 수에 따라 발생할 수 있는 문맥 적합성 점수의 차이를 주제범주 간 상대적 수치로 정규화 하였다. 이 값이 클수록 문맥적으로 다른 주제범주보다 해당 주제범주가 상대적으로 유사하다는 의미이로 0~1 사이의 값을 가진다. 이 수치 중 음수의 값으로 판정하는 기준은 γ의 매개변수로 그 값을 0.4~0.9의 각기 다른 수치를 대입해 봄으로써 최적화 하였다. 두 분류기를 통한 점수를 더한 최종 점수로 분류 판정을 하는데 이 때 분류 판정은 SVM 분류기의 경우와 동일하게 양수일 때는 해당 주제범주로 배정하고, 음수일 때는 배정하지 않는 형태로 분류 판정을 하였다.

3.6 문맥활용 자질선정

기존의 자질선정 기법은 용어 t_i가 특정 주제범주와 그 이외의 주제범주에서 출현한 분포를 기반으로 한다. 즉, 특정 주제범주에서만 주로 출현하는 용어가 분별력이 큰 것으로 판단하는 것이다. 하지만 대부분의 자질선정 기법 역시 용어를 독립적으로 취급하기 때문에 다른 용어들과의 관계는 반영하지 못한다. 하지만 문맥을 활용하면, 주변 용어들 간의 관계가 반영되기 때문에 더 좋은 자질을 선정할 수 있을 것이다.

문맥을 자질선정에 활용할 때는 주제범주에 따른 문맥의 상이성을 고려할 수 있다. 주제범주에 따라 문맥이 서로 다르다는 것은 주제범주에 따라 해당 용어 t_i가 출현한 문헌은 중복되는 용어의 수가 적다는 것을 의미하기 때문이다. 그렇기 때문에 기존의 자질선정 기법에 의한 자질 값과 주제범주에 따라 문맥이 상이한 정도를 측정한 값을 MMR 형태로 결합하는 방법은 보다 분별력이 큰 분류자질을 선정할 수 있다. 실험에서 적용할 기존의 자질선정 기법으로는 베이스라인에 적용했던 카이제곱을 적용하였고,

총 분류자질의 수는 베이스라인과 동일하게 하였다. 용어 가중치나 분류기 등의 여타 다른 요소도 동일하게 적용하였는데 이는 자질선정 부분에서만 발생한 성능 변화를 파악하기 위함이다. 구체적인 자질선정 방법은 아래와 같다.

$$t-score(t,c_i)=\lambda \cdot \frac{\chi^2(t,c_i)}{\max(\chi^2)}- \frac{(1-\lambda)}{\max(\sum Sim_t(p_i,p_j))} \cdot \sum_{\substack{j\in C \\ j\neq i}} Sim_t(p_i,p_j)$$

$\chi^2(t,c_i)$: 용어 t에 대한 카이제곱 값

$\max(\chi^2)$: 용어들의 카이제곱 값 중 가장 큰 값

$\sum Sim_t(p_i,p_j)$: 용어 t의 문맥 프로파일 i,j 간 유사도 합산

λ: 매개변수

4. 실험 결과 분석 및 평가

본 연구는 Reuters-21578의 상위 10개의 주제범주를 대상으로 문맥 프로파일을 이용해 동적 가중치를 적용한 실험과 문맥 적합성 점수를 산출해 SVM 분류기와 결합한 실험, 그리고 문맥 프로파일의 주제범주 간 유사도를 활용한 문맥활용 자질선정의 3가지 실험을 진행하였다. 하나의 문헌은 여러 주제범주에 속할 수 있는 형태이기 때문에 주제범주별로 이원 판정을 할 수 있는 SVM 분류기를 각각 구성하여 성능을 측정하였다.

4.1 베이스라인 실험 결과

베이스라인 설정을 위해 단어빈도(TF)와 역문헌빈도(IDF)를 이용한 8개의 가중치 기법을 달리 적용해 보며 그 분류 성능을 측정하였다. 분류 성능이 가장 좋은 가중치 기법은 *okapi_tf・idf*로 F_1 값이 0.8971로 나타났다. 반면, 단순 *tf* 가중치는 F_1 값이 0.8275로 가장 낮은 성능을 보여주었다. 단순 *tf* 가중치는 재현율에 있어서 다른 가중치들 보다 현저히 떨어지기 때문에 전체적인 성능이 떨어지는 것으로 나타났는데 이러한 점은 *tf・idf*에서도 나타났다.

본 연구에서는 비교실험 결과에 대한 견고성(robustness)을 확보하기 위해 8개 가중치 기법 중 가장 좋은 성능을 보여준 *okapi_tf・idf*의 분류 결과를 베이스라인으로 설정하였다.

〈표 2〉 용어 가중치 기법에 따른 분류 성능

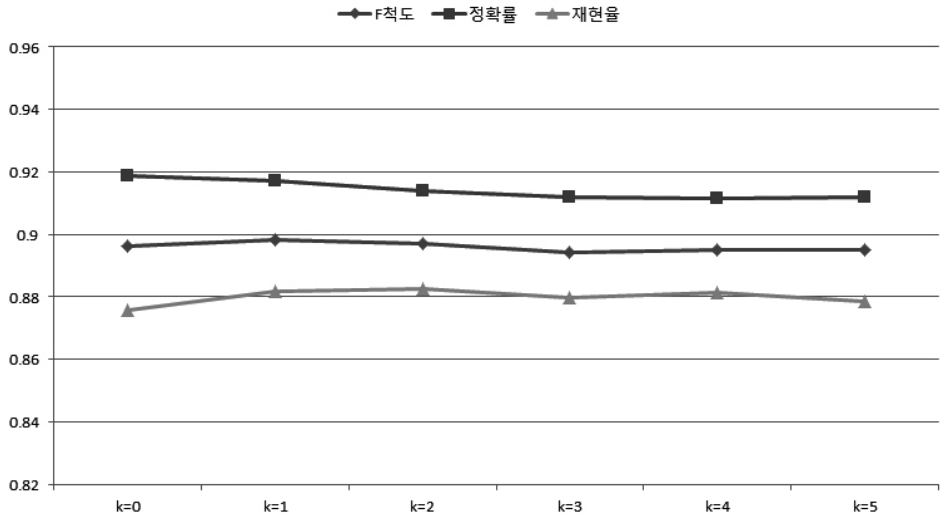
	정확률	재현율	F_1
tf	0.9108	0.7635	0.8275
tf・idf	0.9102	0.7796	0.8370
bi_tf	0.9060	0.8469	0.8723
bi_tf・idf	0.9091	0.8526	0.8769
log_tf	0.9230	0.8383	0.8769
log_tf・idf	0.9130	0.8309	0.8680
okapi_tf	0.9240	0.8188	0.8653
okapi_tf・idf	0.9256	0.8404	0.8791
평균	0.9152	0.8214	0.8629

4.2 동적 가중치 실험 결과

이 실험은 분류 자질의 수나 종류는 베이스라인과 동일하고, 단지 용어의 가중치만 변경해 준 것으로 파라미터 k의 값을 0~5로 바꿔가며 그 분류 성능을 측정하였다.

파라미터 k값에 따른 분류 성능의 변화를 살펴보면, k값이 커질수록 정확률은 다소 하락하는 경향이 있는 것으로 나타났고, 이에 반해 재현율은 k=3일 때까지는 상승하는 것으로 나타났다. 정확률과 재현율을 모두 반영한 F_1 척도는 k값이 0~1 사이에는 지속적으로 상승하는 추세를 보여주었고, k=3 이후부터는 그 값이 고정되는 모습이다. 파라미터 k값의 변화에 따른 변화추이는 이와 같았지만 전체적으로 F_1 값이 0.89~0.9 사이에 존재할 정도로 변화 폭이 그리 크지는 않았다. 또한 k값을 0~5까지 적용한 6개의 모든 경우에서 베이스라인보다 분류 성능이 더 좋은 것으로 나타났다.

파라미터 k값에 대한 분류 성능의 구체적인 수치는 〈표 3〉과 같다. 분류 성능이 가장 높은 경우는 k값이 1일 때로 베이스라인보다 2.16% 향상되는 것으로 나타났다. 정확률은 베이스라인보다 다소 하락하였지만 재현율이 상대적으로 더 많이 상승하여 전체적인 성능 향상을 가져왔다. 전체 분류 성능(F_1 값)은 향상되었지만 10개의 주제범주 중 acq, crude, earn, money-fx 주제범주에서는 다소 분류 성능이 하락하기도 하였다. acq와 earn은 각각 -0.93%와 -0.37%로 하락폭이 미미했다면, money-fx와 crude 범주는 각각 -3.48%와 -2.7%의 성능 하락을 가져온 것으로 나타났다. 그 외의 범주는 분류 성능이 향상되는 모습을 보여주었는데 corn 범주는 6.2%, ship 범주는 9.33%, wheat 범주는 8%의 큰 성능 향상을 가져왔다. 주제범주별 성능 차이는 학습문헌의 수와 관련된 모습으로 비교적 학습문헌 수가 많은 주제범주에서는 성능이 다소 하락하는 경향을 보였다.



〈그림 3〉 파라미터 k에 따른 동적 가중치 성능 변화 추이

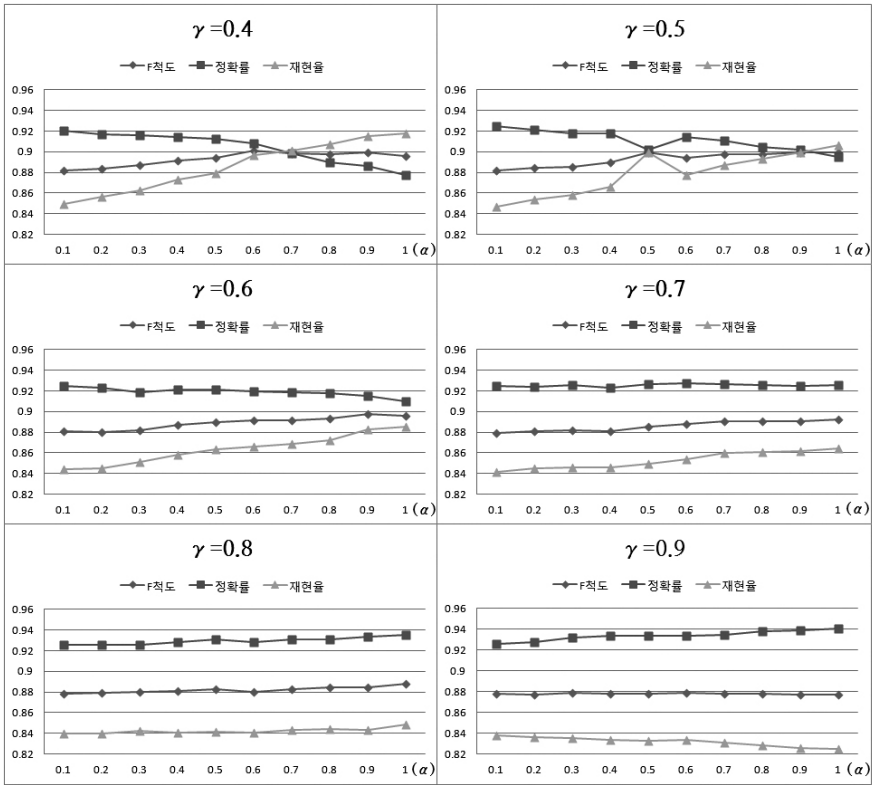
〈표 3〉 파라미터 k에 따른 동적 가중치 기법의 분류 성능

	정확률	재현율	F_1	성능향상률(F_1)
k=0	0.9189	0.8759	0.8961	+1.93%
k=1	0.9170	0.8816	0.8981	+2.16%
k=2	0.9141	0.8827	0.8969	+2.02%
k=3	0.9119	0.8797	0.8944	+1.74%
k=4	0.9117	0.8814	0.8951	+1.82%
k=5	0.9121	0.8787	0.8951	+1.82%
베이스라인	0.9257	0.8404	0.8791	(비교대상)

4.3 문맥 적합성 점수 결합 실험 결과

이 실험은 실험문헌마다 각 주제범주의 문맥 적합성 점수를 산출하여, SVM 분류기와 결합해 최종 문헌을 분류한 것으로 SVM 분류기와 결합을 위해 α와 γ의 2개의 파라미터를 사용하였다. α는 0.1~1.0, γ는 0.4~0.9 사이의 값을 대입하여 분류 성능을 살펴보았다.

〈그림 4〉는 정확률과 재현율을 함께 표시한 변화 추이로 γ 값이 클수록 정확률이 상승하고,



〈그림 4〉 파라미터에 따른 문맥 적합성 결합의 성능 변화 추이

〈표 4〉 파라미터 α, γ에 따른 문맥적합성 점수 결합의 분류 성능 (F_1 척도)

	$\gamma=0.4$	$\gamma=0.5$	$\gamma=0.6$	$\gamma=0.7$	$\gamma=0.8$	$\gamma=0.9$
$\alpha=0.1$	0.8816	0.8819	0.8808	0.8791	0.8786	0.8777
0.2	0.8839	0.8847	0.8804	0.8804	0.8788	0.8774
0.3	0.8870	0.8855	0.8820	0.8822	0.8801	**0.8791**
0.4	0.8918	0.8899	0.8873	0.8812	0.8805	0.8784
0.5	0.8941	0.8994	0.8900	0.8849	0.8825	0.8783
0.6	**0.9014**	0.8943	0.8911	0.8876	0.8805	0.8788
0.7	0.8988	0.8975	0.8919	0.8909	0.8831	0.8782
0.8	0.8973	0.8978	0.8933	0.8908	0.8842	0.8779
0.9	0.8993	0.8994	**0.8978**	0.8910	0.8845	0.8771
1.0	0.8958	**0.8995**	0.8962	**0.8924**	**0.8880**	0.8772
베이스라인	0.8791					

γ 값이 작을수록 정확률은 하락하는 것을 볼 수 있다. 반면, 재현율은 이와는 반대의 양상으로 변화하는 것으로 나타났는데 정확률의 변동 폭보다 훨씬 더 큰 것으로 나타났다. 분류 성능이 가장 좋을 때인 $\gamma=0.4$, $\alpha=0.6$의 경우를 보면, 재현율이 대폭 향상되어 F_1 수치가 증가하였는데 이 지점은 정확률과 재현율이 서로 근접한 지점이다. $\gamma=0.5$일 때도 α값에 따라 정확률과 재현율이 교차하는 모습을 보였는데 이 역시 정확률과 재현율이 가장 근접한 $\alpha=0.9$ 일 때 가장 좋은 성능을 보여주었다. 또한 $\gamma=0.4\sim0.7$ 일 때 상대적으로 다른 경우보다 좋은 성능을 보이는 것으로 나타났는데 그 이유는 이 구간에서 정확률의 하락은 크지 않고, 재현율의 상승이 상대적으로 크기 때문이다.

파라미터의 변화에 따른 분류 성능은 〈표 4〉와 같다. 전체적인 실험 결과를 살펴볼 때, γ값이 0.8와 0.9일 때를 제외하고는 모든 경우에서 베이스라인보다 더 좋은 성능을 보이는 것으로 나타났다. 또한 γ값이 커질수록 α에 의한 성능 변화가 줄어드는 모습인데 특히, γ값이 0.8 이상인 경우는 성능 변화가 거의 없었다. F_1 척도가 가장 높은 값을 가질 때 최적화 된 파라미터는 γ가 0.4이고, α가 0.6일 때로 0.9014의 수치를 기록했다. 이는 베이스라인과 비교했을 때, 2.54%의 분류 성능이 향상된 것으로 앞서 동적 가중치 조정을 통한 성능 향상률보다 더 향상된 수치로 crude 주제범주를 제외한 나머지 모든 범주에서 성능 향상이 이루어졌다.

4.4 문맥활용 자질선정 실험 결과

이 실험은 기존의 자질선정 기법과 문맥 프로파일의 주제범주 간 유사도를 MMR 형태로 결합하여 자질선정을 수행한 것으로 기존 자질선정 기법으로는 카이제곱을 적용하였고, 축소비율은 베이스라인과 동일하게 20%로 하였다. 매개변수 λ는 0.1~0.9의 값을 대입하였는데, λ의 값이 커질수록 카이제곱 기법의 영향력이 상대적으로 증가하게 된다.

실험집단인 Reuters-21578은 하나의 문헌이 여러 개의 주제범주에 속할 수 있기 때문에 각

〈표 5〉 λ 값에 따른 주제범주의 분류 성능

section 1 (acq, corn, crude, earn, garin)					
	acq	corn	crude	earn	gain
λ=0.1	0.9760	0.8191	0.8767	0.9862	0.9370
0.2	0.9781	0.8269	0.9177	0.9885	0.9296
0.3	0.9782	0.8191	0.9235	0.9885	0.9333
0.4	0.9774	0.8113	0.9259	0.9890	**0.9370**
0.5	0.9746	0.8191	**0.9291**	0.9890	0.9296
0.6	0.9797	0.8302	0.9235	0.9899	0.9333
0.7	0.9789	0.8269	0.9198	0.9899	0.9296
0.8	**0.9804**	0.8269	0.9202	0.9899	0.9296
0.9	0.9790	**0.8350**	0.9226	**0.9904**	0.9333
χ^2(λ=1)	0.9783	0.8350	0.9226	0.9894	0.9333
secton 2 (interest, money-fx, ship, trade, wheat)					
	interest	money-fx	ship	trade	wheat
λ=0.1	0.7801	0.8219	0.7848	0.8407	0.8308
0.2	0.8084	0.8409	0.7820	0.8782	**0.8372**
0.3	0.8117	0.8427	0.7662	**0.8811**	0.8308
0.4	0.8133	**0.8652**	0.7820	0.8571	0.8308
0.5	**0.8167**	0.8571	0.7662	0.8622	0.8244
0.6	0.8017	0.8483	**0.8026**	0.8622	0.8244
0.7	0.8000	0.8500	**0.8026**	0.8584	0.8244
0.8	0.8000	0.8611	0.7871	0.8584	0.8244
0.9	0.8033	0.8579	0.7949	0.8584	0.8244
χ^2(λ=1)	0.8067	0.8515	0.7949	0.8546	0.8244

주제범주마다 별도의 분류기를 생성해 그 성능을 측정하게 되는데 자질선정 역시 각 주제범주에 대한 자질들을 각각 선정하여 해당 주제범주의 자동 분류 실험에 사용되었다. λ값에 따른 분류 성능은 〈표 5〉와 같다.

λ값에 따른 성능 변화는 주제범주마다 다른 것으로 나타났다. λ값이 크고 작음에 따라 모든 주제범주에서 일관적인 성능 변화가 발생하지는 않았고, 각 주제범주에 따른 다른 양상을 보여 주었다. 주제범주별로 λ값이 최적화 되었을 때는 corn 범주를 제외한 나머지 주제범주 모두에서 기존의 카이제곱 자질선정 기법을 통한 분류 성능보다 더 좋은 성능을 보여주었다. corn 범주는 카이제곱 기법과 동일한 분류 성능을 보여주었다.

4.5 종합평가

본 연구에서는 문맥을 활용해 자동 분류의 성능을 향상시키기 위해 용어 가중치, 분류기 결합, 자질선정의 3가지 항목에 대해 실험을 진행하였다. 용어 가중치는 동적 가중치를 적용하였고, 분류기 결합에서는 문맥 적합성 점수를 SVM 분류기 점수와 결합하였다. 자질선정

부분에서는 주제범주별 문맥 프로파일의 유사도를 MMR 형태로 결합해 자질을 선정하여 그 성능을 비교하였다.

각 요소에 문맥을 활용했을 때의 F_1 값은 〈표 6〉과 같다. 표의 결과에서 볼 수 있듯이, 동적 가중치, 문맥 적합성 점수 결합, 문맥활용 자질선정의 3가지 실험 모두 베이스라인보다 성능이 향상하는 것으로 나타났다. 이들 실험 중 성능 향상률이 가장 큰 것은 문맥 적합성 점수 결합으로 2.54%의 성능 향상을 가져왔다. 그 뒤를 이어 동적 가중치를 적용한 실험이 2.16%, 자질선정에 문맥을 활용한 문맥활용 자질선정 실험이 0.96%의 성능 향상을 보였다.

분류 성능을 주제범주 별로 나누어 살펴보면, 학습문헌 수가 적은 범주에서 주로 성능이 향상되는 것으로 나타났다. 〈표 7〉은 학습문헌 수가 많은 주제범주 순으로 나열한 것인데 베이스라인의 경우, 학습문헌 수가 많은 5개의 주제범주와 상대적으로 학습문헌의 수가 적은 5개의 주제범주 간 분류 성능은 0.9350과 0.8231로 편차가 큰 것을 볼 수 있다. 반면, 문맥을 활용한 동적 가중치, 문맥 적합성 점수 결합, 문맥활용 자질선정의 분류 성능에서는 이러한 편차가 줄어드는 것을 확인할 수 있는데 하위 5개의 주제범주에서의 성능 향상이 두드러졌기 때문에 이러한 결과를 보이고 있다.

〈표 6〉 종합 성능 비교 (F_1 척도)

실험	F_1	향상률	최적 파라미터
베이스라인	0.8791	(비교대상)	
동적 가중치	0.8981	+2.16%	k=1
문맥 적합성 점수 결합	0.9014	+2.54%	γ=0.4, α= 0.6
문맥활용 자질선정	0.8875	+0.96%	주제범주별

〈표 7〉 주제범주별 종합 성능 비교

순위	범주명	베이스라인	동적 가중치	적합성 점수	문맥자질선정
1	earn	0.9894	-0.37%	+0.05%	+0.10%
2	acq	0.9783	-0.93%	+0.36%	+0.21%
3	money-fx	0.8515	-3.48%	+4.25%	+1.61%
4	grain	0.9333	+1.55%	+0.80%	+0.40%
5	crude	0.9226	-2.70%	-1.06%	+0.70%
상위 5개 평균		0.9350	-1.13%	+0.81%	+0.58%
6	trade	0.8546	+5.66%	+1.12%	+3.10%
7	interest	0.8067	+0.38%	+2.32%	+1.24%
8	wheat	0.8244	+8.00%	+6.25%	+1.55%
9	ship	0.7949	+9.33%	+9.52%	+0.97%
10	corn	0.8350	+6.20%	+3.57%	0.00%
하위 5개 평균		0.8231	+5.92%	+4.51%	+1.39%
전체 평균		0.8791	+2.16%	+2.54%	+0.96%

본 연구에서 수행한 3가지 실험은 기본적으로 분류 자질의 수나 텍스트 처리를 베이스라인과 동일하게 적용하였기 때문에 큰 성능 향상은 보이기에는 다소 제한점이 있다. 또한 베이스라인의 성능이 약 0.88로 높은 수치를 기록하였다는 것 역시 성능 향상 폭이 제한되는 점이다. 그럼에도 문맥을 활용한 3가지 실험 모두 베이스라인보다 좋은 성능을 보였다는 점은 문맥을 활용하는 것이 기존의 자동 분류에서 갖는 문제점을 어느 정도 보완할 수 있다는 것을 의미한다고 할 수 있다.

5. 결론 및 제언

본 연구는 문헌 자동 분류의 일반적인 문헌 표현 방식인 BOW가 갖는 한계점에 주목하였다. 특히, 각 용어를 독립적으로 취급하기 때문에 2개 이상의 용어를 함께 고려했을 때 가지는 문맥적 정보가 자동 분류과정에 실제적으로 반영이 되지 않는다는 점이 자동 분류의 성능을 저하시키는 한 요인이라고 보고, 문맥을 통해 얻을 수 있는 정보를 자동 분류 과정에 접목하고자 하였다.

본 연구는 2가지 구체적 목적을 가지고 수행되었다. 첫째, 문맥의 정보를 활용하는 것이 자동 분류의 성능 향상에 유용성을 가지는지 확인하는 것으로 실험을 통한 성능 비교를 통해 그 의의를 파악하고자 하였다. 둘째, 문맥적 요소를 자동 분류에 적용하는 것이 유용성이 있다면, 보다 효과적인 방법이나 부분은 무엇인지 파악하기 위한 것으로 자동 분류 성능에 영향을 미칠 수 있는 요소(가중치, 분류자질, 분류기 등)에 문맥 활용을 각각 적용하여 그 분류 성능을 비교하는 것이다.

먼저, 주변 용어들인 문맥을 활용하는 것이 자동 분류 성능 향상에 유용한가에 대한 문제에서는 실험 결과, 성능 향상의 폭이 그리 크지 않지만 전반적으로 효과가 있는 것으로 나타났다. 앞서 문맥 프로파일을 활용한 3가지 실험 모두에서 베이스라인보다 더 높은 분류 성능을 보여주었기 때문이다. 동적 가중치 실험은 최고 2.16%, 문맥 적합성 점수 결합 실험은 2.54%, 문맥 MMR 자질선정 실험은 최고 0.96% 분류 성능이 향상되는 것으로 나타났다. 또한 자동 분류 과정 중 문맥 활용의 효과가 가장 큰 방법은 분류기 결합인 것으로 나타났다. 하지만 용어 가중치 부분의 동적 가중치 실험 결과와 차이가 미미한 것으로 나타나 보다 많은 연구가 필요한 것으로 나타났다.

실험 결과와 실험을 통해 얻어진 결과를 항목별로 요약하면 다음과 같다.

첫째, 문맥을 활용해 용어 가중치를 조정하는 것은 분류 성능이 향상되는 것으로 나타났다. 분류자질, 분류기 등의 다른 요소는 베이스라인과 동일하게 적용하고 용어 가중치만 문맥 프로파일을 이용하여 재조정한 결과, 일반적 용어 가중치 중 가장 좋은 성능을 보였던 *okapi_tf・idf*의 베이스라인보다 최대 2.16% 향상되었다.

둘째, 하나의 용어 단위가 아닌 문헌 단위로 문맥의 유사성을 측정하고 이를 분류 점수에 반영한 자동 분류 실험에서도 베이스라인보다 성능이 향상되는 것으로 나타났다. 문맥 적합성 점수 결합 실험에서 최적화가 이루어졌을 때 베이스라인의 분류 성능보다 2.54% 향상되었는데 이는 동적 가중치를 적용했을 때보다

0.38% 더 높은 수치이다.

셋째, 기존 자질선정 기법인 카이제곱과 문맥 프로파일을 통해 각 주제범주별 유사도를 산출한 결과를 결합한 자질선정 실험에서 베이스라인보다 0.96%의 성능 향상이 있는 것으로 나타났다. 이 경우, 정확률과 재현율 모두 상승하는 모습을 보였다.

넷째, 베이스라인에서는 주제범주에 속하는 학습문헌의 수가 많고 적음에 따라 분류 성능의 차이가 큰 것으로 나타났으나 문맥 프로파일을 적용했을 때는 이 차이가 확연히 줄어드는 것으로 나타났다. 이는 문맥 프로파일을 적용했을 때, 학습문헌 수가 적은 하위 5개의 주제범주에서 더 큰 폭의 성능 향상을 보였기 때문으로 3가지 실험 모두에서 상위 5개의 주제범주보다 나머지 5개의 주제범주에서 성능 향상이 더 많이 이루어졌다.

이와 같이 3가지 실험 모두에서 베이스라인보다 분류 성능이 향상되었지만 이번 연구에는 한계점도 가지고 있다. 첫째로는 실험의 대상 말뭉치가 Reuters-21578 하나였다는 점이다. 두 번째로, 10개의 주제범주 중 분류 성능이 하락하는 주제범주도 있었다는 점이다. 동적 가중치 실험에서는 4개의 주제범주, 문맥 적합성 점수 실험에서는 1개의 주제범주가 베이스라인보다 하락하였다. 10개 주제범주를 모두 고려한 전체 분류 성능에서는 모두 성능이 향상되었지만 모든 주제범주에서 향상된 것은 아니었다. 세 번째는 문맥 활용의 방법이 다소 제한적이었다는 점이다. 분류 성능의 향상 폭이 최고 2.5%에 그쳤다는 점을 볼 때 보다 효과적인 적용 방법이 필요하다는 것을 보여주는 것이라 할 수 있다.

하지만 본 연구는 BOW 방식의 한계점으로 지적된 사항에 대한 대안을 제시하고 실제 분류 성능의 향상까지 이끌어 냈다는 점에서 의의를 갖는다고 할 수 있다. 특히, 용어를 추가하거나 변형하는 등의 물리적 변화 없이 주어진 문헌들을 바탕으로 부가적인 알고리즘의 적용만으로 성능 향상을 이루었기 때문에 다른 외부자원을 사용해 텍스트 자체를 변형하는 방법보다 연산이 간단하고, 적용 가능한 환경이 더 많다는 강점이 있다.

앞으로 연구를 통해서는 보다 다양한 말뭉치와 환경에서 동일한 효과를 보이는지 계속적인 실험을 통해 일반화 할 필요가 있고, 단편적으로 부수적 장치를 통해 문맥을 활용하는 것을 넘어 문헌을 표현하는 새로운 방법에 대한 연구와 분류 알고리즘 내에 여러 용어들이 함께 반영되었을 때 가지는 정보를 반영할 필요가 있을 것이다.

참 고 문 헌

김판준 (2008). 용어 가중치부여 기법을 이용한 로치오 분류기의 성능 향상에 관한 연구. 정보관리학회지, 25(1), 211-233. http://dx.doi.org/10.3743/KOSIM.2008.25.1.211

이재윤 (2005). 문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구. 정보관리학회지, 22(3), 261-287.

이지혜, 정영미 (2009). 지도적 잠재의미색인(LSI) 기법을 이용한 의견 문서 자동 분류에 관한 실험적 연구. 정보관리학회지, 26(3), 451-462. http://dx.doi.org/10.3743/KOSIM.2009.26.3.451

정은경 (2009). 문서범주화 성능 향상을 위한 의미기반 자질확장에 관한 연구. 정보관리학회지, 26(3), 261-278. http://dx.doi.org/10.3743/KOSIM.2009.26.3.261

David, D. L. (2004). Reuters-21578 text categorization test collection distribution 1.0. Retrieved from http://www.daviddlewis.com/resources/testcollections/reuters21578/

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research, 34(2009), 443-498. http://dx.doi.org/10.1613/jair.2669

Huynh, D., Tran, D., Ma, W., & Sharma, D. (2011). A new term ranking method based on relation extraction and graph model for text classification. Proceedings of the Australasian Computer Science Conference (ACSC 2011), Perth, Australia, 145-152.

Porter, M. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.

Sable, C., McKeown, K., & Church, K. W. (2002). NLP found helpful (at least for one text categorization task). Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMLNLP) 2002, 172-179.

Wang, P., & Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 713-721.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Chung, Eun-Kyung (2009). A semantic-based feature expansion approach for improving the effectiveness of text categorization by using WordNet. Journal of the Korean Society for Information Management, 26(3), 261-278. ttp://dx.doi.org/10.3743/KOSIM.2009.26.3.261

Kim, Pan-Jun (2008). A Study on the Performance Improvement of Rocchio Classifier with Term Weighting Methods. Journal of the Korean Society for Information Management, 25(1), 211-233. http://dx.doi.org/10.3743/KOSIM.2008.25.1.211

Lee, Jae Yun (2005). Improving the performance of SVM text categorization with inter-document similarities. Journal of the Korean Society for Information Management, 22(3), 261-287.

Lee, Ji-Hye & Chung, Young Mee (2009). An experimental study on opinion classification using supervised latent semantic indexing (LSI). Journal of the Korean Society for Information Management, 26(3), 451-462. http://dx.doi.org/10.3743/KOSIM.2009.26.3.451