문서범주화 성능 향상을 위한 의미기반 자질확장에 관한 연구

A Semantic-Based Feature Expansion Approach for Improving the Effectiveness of Text Categorization by Using WordNet

정은경(Eun-Kyung Chung)*

초 록

기계학습 기반 문서범주화 기법에 있어서 최적의 자질을 구성하는 것이 성능향상에 있어서 중요하다. 본 연구는 학술지 수록 논문의 필수적 구성요소인 저자 제공 키워드와 논문제목을 대상으로 자질확장에 관한 실험을 수행하였다. 자질확장은 기본적으로 선정된 자질에 기반하여 WordNet과 같은 의미기반 사전 도구를 활용하는 것이 일반적이다. 본 연구는 키워드와 논문제목을 대상으로 WordNet 동의어 관계 용어를 활용하여 자질확장을 수행하였으며, 실험 결과 문서범주화 성능이 자질확장을 적용하지 않은 결과와 비교하여 월등히 향상됨을 보여주었다. 이러한 성능향상에 긍정적인 영향을 미치는 요소로 파악된 것은 정제된 자질 기반 및 분류어 기준의 동의어 자질확장이다. 이때 용어의 중의성 해소 적용과 비적용 모두 성능향상에 영향을 미친 것으로 파악되었다. 본 연구의 결과로 키워드와 논문제목을 활용한 분류어 기준 동의어 자질 확장은 문서 범주화 성능향상에 긍정적인 요소라는 것을 제시하였다.

ABSTRACT

Identifying optimal feature sets in Text Categorization(TC) is crucial in terms of improving the effectiveness. In this study, experiments on feature expansion were conducted using author provided keyword sets and article titles from typical scientific journal articles. The tool used for expanding feature sets is WordNet, a lexical database for English words. Given a data set and a lexical tool, this study presented that feature expansion with synonymous relationship was significantly effective on improving the results of TC. The experiment results pointed out that when expanding feature sets with synonyms using on classifier names, the effectiveness of TC was considerably improved regardless of word sense disambiguation.

키워드: 자질선정, 의미기반, 문서범주화
WordNet, text categorization, semantics, feature selection, feature expansion

* 이화여자대학교 사회과학대학 문헌정보학 조교수(echung@ewha.ac.kr)

■ 논문접수일자 : 2009년 8월 16일 ■ 최초심사일자 : 2009년 8월 20일 ■ 게재확정일자 : 2009년 8월 28일
■ 정보관리학회지, 26(3): 261-278, 2009. 〔DOI:10.3743/KOSIM.2009.26.3.261〕

1. 서 론

기계학습 기반 문서범주화(text classification 또는 text categorization)는 다양한 응용 분야를 구성하고 있으며, 문서 자동 분류, 필터링, 정보검색 등의 분야에서 활발한 연구와 실제적인 시스템이 구현되고 있다. 문서범주화는 폭발적으로 증가하는 문서의 양으로 인해, 효율적이며 정확한 문서범주화 결과에 대한 수요가 급증하는 추세이다(Forman 2003). 문서범주화 연구에 있어서 중요한 연구 과제 중의 하나가 문서범주화 성능향상에 긍정적인 영향을 미치는 자질 요소를 규명하는 것이다. 이러한 분야의 연구가 활발한 이유는 최적의 자질구성은 궁극적으로 문서범주화의 성능향상에 중요한 역할을 수행하기 때문이다. 효과적인 자질요소 규명은 두 가지 연구 분야로 구분될 수 있다. 첫째, 문서범주화의 대상이 되는 문헌 집합에서 자질을 선정 및 추출하는 기법에 관한 연구이다. 이러한 연구 동향은 주로 문헌의 전문을 다루기 때문에 최적의 자질만으로 구성된 자질요소로 효과적으로 축소하는데 관심을 둔다. 즉, 문서범주화 기법을 적용하는데 있어 문서의 용어를 있는 그대로 사용하기 보다는 문서의 용어(자질)를 효율적으로 축소시켜 최적의 자질 요소와 성능향상을 추구한다. 둘째, 문서범주화 성능향상을 위한 최적의 자질요소 규명하는데 있어서, 기존의 자질 요소에 의미와 관련된 자질들을 첨부하는 기법이다. 따라서 이러한 연구 분야는 문헌의 전문을 처리하기 보다는 자질 규모가 작은 문헌의 구성요소를 활용하는데 초점을 둔다.

본 연구는 문서범주화 성능향상을 위해서 기존 자질 요소와 의미상의 용어 관계를 활용하여 최적의 자질요소를 규명하는데 그 목적이 있다. 따라서 본 연구는 의미기반의 자질확장이 문서범주화 성능향상에 미치는 영향을 파악하고자 한다. 이를 위해서 첫째, 관련 연구를 통해서 자질확장에 영향을 미치는 관련 요소를 파악하고자 하였다. 문서범주화의 자질확장은 대상 문서의 구성요소, 확장의 기준, 다양한 용어관계, 용어의 중의성 등 실제적으로 다양한 요소로 구성되어 있다. 둘째, 규명된 자질확장의 요소에 대한 실험을 통해 실증적인 분석을 수행하고자 하였다. 이러한 연구결과는 다양한 문서 구성 요소와 의미론적 관계 용어로의 확장을 통해 문서범주화 성능향상을 기할 수 있으며, 문서범주화의 실제적인 응용 시스템 설계와 구현 시에 응용될 수 있는 이론적 배경으로 활용될 수 있을 것으로 기대된다.

2. 자질확장

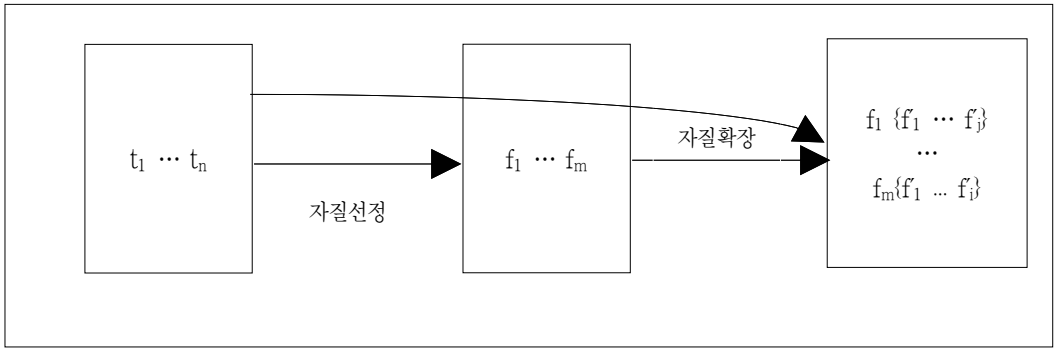
2.1 자질선정과 자질확장

문서범주화에 있어서 자질(feature)은 일반적으로 문헌 내의 용어(term) 혹은 색인어(index term)로 표현된다. 문헌 내의 모든 용어를 자질로 표현하여 문서 범주화 기법을 적용하는 것은 컴퓨터의 공간 및 계산 측면에서 효율적이지 않을 뿐만 아니라, 범주화 성능에 있어서도 긍정적인 영향을 미치지 않는다. 따라서 문서 내의 일부 특정한 용어 혹은 자질을 선정하여 문서범주화에 사용하게 된다. 문서 범주화 과정은 전체 문서에서 중요도 높은 자질을 선정하는

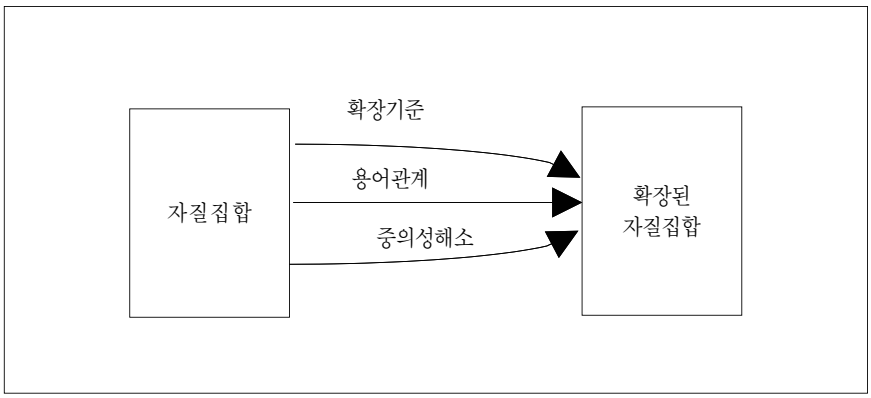
것이 문서범주화 기법의 성능, 정확성, 확장성 등에 영향을 미치는 중요한 요인이다(Chen et al. 2009). 일반적으로 자질 선정은 두 가지 방식으로 수행될 수 있는데, 분류기(classifier)를 먼저 사용하여 분류 결과의 정확성에 근거하여 자질을 선정하는 렙퍼(Wrapper)방식과 분류기와는 별개로 선정하는 필터(filter) 방식이 사용된다. 전자의 경우는 속도에 문제가 있기 때문에, 자질 선정은 일반적으로 후자인 필터 방식을 의미한다(John, Kohavi & Pfleger 1994).

문서범주화에서 사용되는 이러한 자질선정과 자질확장 과정은 〈그림 1〉에서와 같이 간략하게 도식화될 수 있다. 문서범주화의 대상 문서는 집합 $\{t_1 \cdots t_n\}$의 자질을 수록하고 있으며, 다양한 자질선정 기법에 따라 축소된 자질은 $\{f_1 \cdots f_m\}$으로 표현된다. 자질확장은 선정된 자질을 바탕으로 확장하거나, 혹은 기존의 자질을 그대로 사용하여 확장하는 두 가지 형태로 구분할 수 있다. 각 자질 $\{f_1 \cdots f_m\}$은 다양한 확장 기법에 따라 확장된 집합 $f_1 \{f'_1 \cdots f'_j\} \cdots f_m \{f'_1 \cdots f'_i\}$을 수록하게 된다.

우선 자질선정에 관한 기존의 연구를 살펴보면, 다양한 자질 선정 기법을 비교하여 최적의 자질을 제시하는 연구들이 수행되어 왔다. Forman (2003)은 12개의 자질을 비교하여 상당히 광범위한 실험을 수행하였다. 그러나 최종적으로는 지지벡터기계(Support Vector Machine: SVM)를 사용하여 다양한 자질선정 기법을 비교하였기에 문서 범주화 기법과 이에 상응하는 최적의 자질을 선정하는데 있어서는 한계가 있다. 이와 유사한 연구로서 Brank, Milic-Frayling, & Mladenic(2002)은 Naive Bayes, Perception, 지지벡터기계 등의 문서 범주화 기법과 Odds ratio, Information Gain의 자질 선정 기법을 사용하여 비교하였다. 실험 결과 지지벡터기계(SVM)를 사용하였을 때 가장 성능이 우수한 것으로 나타났다. 또한 문서 범주화의 Neural Network 기법에 특화된 자질선정 기법을 제안하여 그 성능의 증가를 나타냈다(Verikas & Bacauskiene 2002). 이재윤(2005)은 kNN(k-nearest neighbors) 문서 범주화 기법을 사용하여, 자질 선정에 있어서 중요한 시사점을 제시하였다. 기존의 연구와는 달리 저빈도 성향의 자질이 kNN 문서 범주화 기법에 있어서 적절하다고 밝혔다.



〈그림 1〉 문서범주화의 자질선정과 자질확장 과정



〈그림 2〉 자질확장에 영향을 미치는 요소

최근 문서범주화의 이러한 자질선정 분야와 함께 추출된 자질 혹은 기존의 자질에 대해서 WordNet 기반, 즉 의미(semantics) 기반 자질 확장(feature expansion)에 관한 연구가 활발하게 진행되고 있다(Manusy & Hilderman 2006). WorldNet[1]은 어휘 데이터베이스로서 영어 단어의 명사, 동사, 형용사, 부사를 동의어, 반의어, 상위어, 하위어 집합으로 연결해 구성된다(Fellbaum 1998). WordNet을 활용한 자질 확장 연구는 〈그림 2〉에서 살펴볼 수 있는 바와 같이, 크게 세 가지 방향으로 구분될 수 있다. 첫째, 확장의 기준에 근거한 연구방향이다. 확장의 기준이 분류어 혹은 본문 용어에 위치할 수 있으며, 이에 따라 해당 기준의 동의어, 상위어, 하위어, 관련어 등으로 확장하는 방식이다. 둘째, 확장의 기준을 용어간의 관계에 따르는 방식으로서 동의 관계로의 확장, 계층관계로의 확장, 유사 개념으로의 확장 등으로 세분될 수 있다. 셋째, 자질 확장 시에 용어의 중의성 해소(WSD: Word Sense Disambiguation) 기법 적용 여부에 따른 상이한 접근법이 제시될 수 있다.

2.2 자질확장의 기준

문서 범주화 연구에서 사용된 자질 확장은 기준에 따라 분류어 기준과 용어 기준으로 구분하여 살펴볼 수 있다. 첫째, 문서 범주화의 자질 확장이 분류어 기준으로 수행된 연구는 상대적으로 많이 이루지지는 않았다. de Buenaga Rodriguez, Gomez-Hidalgo, & Diaz-Agudo (2000)은 분류어(category name)를 기반으로 한 자질 확장 실험을 실시하였다. WordNet 상에서 분류어의 동의어에 해당하는 용어 집합을 해당 문헌의 용어 집합에 추가하여 확장시키는 방식이다. 분류어의 의미가 모호한 경우에는 수작업으로 의미를 구분하여, 가장 적절한 의미에 해당하는 동의어만이 추출될 수 있도록 하였다. 로치오 알고리즘과 Widrow-Hoff 알고리즘을 사용하여, 로이터-21578 데이터를 사용하여 실험하였다. 두 알고리즘 모두를 사용하여

1) WordNet(http://wordnet.princeton.edu/)

실험한 결과, WordNet을 활용하여 자질을 확장한 실험이 자질 확장을 사용하지 않은 경우에 비해 정확도에 있어서 2배 이상의 성능 향상을 보여주었다. 한편, Barak, Dagan & Shnarch (2009)는 분류어의 동의어를 사용하여 자질 집합을 확장하였다. 이러한 자질 확장을 사용하여, 성능 향상 결과를 제시하였다.

둘째, 상대적으로 단순한 분류어 기반의 자질 확장과는 반대로, 문서에서 추출된 자질을 기반으로 자질 확장 연구는 다양한 접근 방식으로 수행되었다. 본문 용어를 기준으로 다양한 용어관계로의 확장에 관한 연구가 수행되었다(Scott & Matwin 1998; 1999; Rosso, Ferretti, Jimenez & Vidal 2004; Mansuy & Hilderman 2006). 이들 연구에 대한 보다 자세한 고찰은 2.3 용어관계 기준의 자질확장에서 다루었다.

2.3 용어 관계 기준의 자질 확장

문서 범주화 연구에서 사용된 자질 확장은 WordNet이 포함하고 있는 용어의 관계를 활용하여 다양하게 수행된 선행 연구를 찾아볼 수 있다. 우선 WordNet의 명사와 형용사 중심의 용어 관계는 동의어(synonym), 상위어(hypernym), 하위어(hyponym), 전체어(holonym), 부속어(meronym) 등으로 구분될 수 있다. 동의어는 유사 개념을 나타내는 용어 관계이며, 수평적 관계이다. 이에 반해, 상위어와 하위어 관계는 계층적인 개념을 나타낸다. 부속어와 전체어 관계는 부분과 전체 개념을 나타내는 것으로서, 예를 들면, '책상'이 전체어라면, 부속어는 '책상 다리'가 해당된다. 이러한 WordNet의 여러 용어 관계를 활용하여 다양한 연구가 수행되었다.

첫째, WordNet의 용어관계 중에서 동의어 관계가 가장 활발하게 활용되었으며, 많은 연구들은 동의어 관계와 함께 다른 관계를 함께 실험하여 분석 결과를 제시하였다. 이러한 연구 중에서 동의어 관계만을 활용한 연구로는 de Buenaga Rodriguez, Gomez-Hidalgo, & Diaz-Agudo(2000)와 Rosso, Ferretti, Jimenez, & Vidal(2003)을 찾아볼 수 있다. de Buenaga Rodriguez, et al.(2000)은 앞서 언급한 바와 같이 분류어를 기준으로 자질 확장을 수행하였는데, WordNet의 다른 관계보다는 동의어 관계만을 활용하여 확장하여 성능 향상을 보였다. 또한 Rosso, Ferretti, Jimenez, & Vidal(2004) 등은 문서에서 추출된 자질 기반으로 동의어로 확장하여 실험하였다. 문서 범주화 기법 중에서는, kNN 기법을 사용하였다. WordNet의 동의어(sysnet)로 확장하였으며, 중의어 해소 기법을 활용하여 동일한 개념으로 자질 확장을 한정하였다. 그 결과 문서범주화 실험에 있어서 성능 향상을 나타냈다고 보고하였다.

둘째, 자질 확장에 있어서 동의어와 함께 다양한 용어 관계를 활용한 연구를 찾아볼 수 있다. Barak, Dagan, & Shnarch(2009)은 분류어 기준으로 자질 확장하는데 있어서 동의어 관계와 하위어 관계를 활용하였다. 하위어 관계가 존재하지 않을 경우에는 부속어 관계를 대신 사용하는 원칙을 채용하여 성능 향상을 보여주었다. 또한 Scott & Martwin(1999)은 기존의 동의어 관계로만 확장한 선행 연구와는 달리, 상위어(hypernym) 관계로의 확장을 제안하였다. 내용상으로는 상위어 관계를 활용하는데 있어, 계층의 깊이를 구분하여 높이가 0이면(h=0)이면 동의어에 해당하며, 높이가 1이

면(h=1) 차상위 계층의 상위어로의 확장을 의미한다. 한편 Mansuy & Hilderman(2006)은 Naive Bayes와 지지벡터기계 기법을 사용하여 WordNet의 의미적 구조를 활용하는 것이 성능 향상에 긍정적인 영향을 미치지 못한다는 결과를 제시하였다. 그러나 이 연구의 실험은 중의어 해소 단계를 포함하지 않았다. Mansuy & Hilderman도 WordNet을 활용한 자질 확장은 중의어 해소 단계가 매우 중요하며, 이 과정이 생략되었기 때문에 긍정적 영향을 찾을 수 없었던 것으로 그 결과를 설명하였다. Mansuy & Hilderman(2006)는 WordNet이 수록하고 있는 용어 간의 관계를 광범위하게 분석하여, 용어 관계가 자질 확장에 사용되었을 때의 영향에 관하여 결과를 제시하였다. WordNet의 동의어, 상위어(hyponym), 부속어(holonyms) 기반의 자질 확장은 문서범주화에 긍정적인 영향을 미쳤으나, 이러한 실험 결과를 일반화하기에는 제한적이라는 결론을 제시하였다. 지지벡터기계 기법은 상위어 관계 용어를 사용하여 자질 확장하였을 때만 성능향상을 보였으나, Coordinate Matching(한글용어를 쓰고 괄호에 영문을 쓰던지, 정 번역이 어려우면 ' '를 사용하고 본문중에 영어를 직접쓰지 않도록) 범주화 기법은 동의어, 상위어, 부속어 관계의 용어를 사용하여 자질 확장하였을 때 성능이 향상되었다고 보고하였다.

2.4 중의성 해소

일반적으로 용어 중의성 해소는 하나의 용어가 다중의 의미를 지니는 경우 해당 문맥에서 적합한 의미를 규명하는 것이라고 정의된다(Sebastiani 2002). 문서 범주화에 있어서 자질 선정이 중요한 것처럼, 자질 확장에 있어서는 용어의 중의성 해소가 결정적인 요인으로 작용한다(Zhang, Sun & Wang 2004). 기존 선행 연구에서 사용된 중의성 해소 접근 방식은 크게 세 가지로 구분할 수 있다(Mansuy & Hilderman 2006). 첫째, 수작업으로 용어의 의미를 구분하여 가장 적합한 의미만을 선정하는 방식이다(de Buenage Rodriguez et al., 1997; Kehagias et al., 2003; Rosso et al., 2004). 이러한 방식은 처리 속도와 분량 면에서 정확성을 높이는 데 있어서는 최상이지만, 실제적인 구현에 있어서 제한적일 수밖에 없다.

둘째, Scott & Matwin(1998)과 Bloehdorn & Hotho(2004) 등은 이러한 문제점을 지적하여, 용어의 모든 의미(sense)를 일괄적으로 모두 활용하여 자질을 확장시키는 실험을 수행하였다. 이 방식은 모든 과정을 자동적으로 처리할 수 있기 때문에 효율적인 측면이 있으나, 정확성을 높이는데 있어서는 제한적이라는 단점이 있다. 세 번째 중의성 해소 접근 방식은 상대적으로 진화한 방식으로서 전자의 두 가지 접근법의 장점만을 활용하여, 그 효율성을 높이고자 하였다. 즉, 중의성 해소를 하는데 있어서 수동 방식이 아니라, 자동화된 방식을 추구하는 동시에, 가장 적합한 의미만을 선택하는 것이다. Bloehdorn & Hotho(2004)의 연구는 하나의 용어에 대해서 WordNet상의 여러 의미 중에서 가장 적합한 의미를 선택하여 사용하는 전략을 구사하였다. 가장 적합한 의미 선택의 기준은 컨텍스트(context)에 대한 이해로 설명할 수 있는데, 문헌에 나온 주변 어휘와 어휘 사전의 정의 간에 서로 매핑되는 용어의 빈도

수에 의해 적합한 의미를 선정하는 방식이다. Kehagias, Petridis, Kaburlasos, & Fragkou (2001)는 'Brown Corpus semantic concordance' 데이터 집합을 WordNet과 함께 사용하였다. Brown 데이터 집합은 중의성 용어의 경우는 사전에 가장 적합한 의미가 부여되어 있는 데이터 집합이다. Kehagias 등은 이 데이터를 사용하여, 중의성이 해소된 용어의 사용이 문서범주화 성능 향상에 크게 기여하였음 보여주었다. 한편으로 Mansuy & Hilderman(2006)은 반대 경우의 실험을 통해서 이러한 용어의 중의성 해소가 문서범주화 성능 향상에 긍정적인 영향을 미치는 연구결과를 제시하였다. 이들 연구는 용어의 중의성 해소 과정을 포함하지 않고, 자질 확장을 한 경우에, 성능향상에 긍정적 영향을 보이지 않음을 보였다.

3. 실험설계

3.1 실험데이터

본 연구는 문서범주화의 성능향상을 목적으로 의미기반의 자질 확장 실험을 수행하기 위하여 두 가지 실험데이터를 사용하였다. 첫째, 실험 문헌집합은 전자공학, 컴퓨터학, 물리학, 통계학, 산업공학 분야의 서지정보 및 전문(full text) 데이터베이스인 인스펙(INSPEC) 데이터베이스에서 추출되었다. 문헌집합을 추출하기 위해서 사용된 주제어는 총 20[2]개이며, 각각 50문헌이 무작위로 선정되었다. 문헌의 수는 기존의 연구결과 가장 성능이 우수한 결과를 참조하여 선정되었다. 본 연구는 문헌집합에서 문서범주화 성능에 긍정적인 영향을 주는 우수한 자질을 선정하는 기법이라기보다는 선정된 자질을 기반으로 하여 자질 확장의 영향에 중점을 둔다. 따라서 각 문헌의 요소 중에서, 저자 선정 키워드와 논문 제목에서 추출된 자질과 같이 기존 정제된 자질 기반으로 자질확장 실험을 수행하였다. 〈표 1〉에서 살펴볼 수 있는 바와 같이, 각 문헌집합과 이에 속한 자질 수가 제시되어 있다. 일반적으로 키워드의 자질 수가 논문제목에서 추출된 자질 수에 비해 약 2배 정도 많은 것으로 나타났다.

〈표 1〉 문헌집합과 자질 수

	전체문헌집합	이질문헌집합	동질문헌집합
키워드	4,576	3,377	2,246
논문제목	2,264	1,551	1,165

둘째, 어휘 데이터베이스로서 WordNet[3]이 사용되었다. WordNet은 영어 어휘 목록으로서 동의어(synonym), 반의어(antonym), 상위어(hypernym), 하위어(hyponym), 부속어(holonym) 등의 다양한 관계별로 어휘를 수록하고 있다(Miller 1995). 실험에서는 동의어 관계가 자질확장에 사용되었으며, WordNet 상의 'sysnet' 용어로의 확장을 의미한다. 실험에 사

2) computer architecture, computer graphics, computer interfaces, discrete systems, information management, knowledge based systems, pattern recognition, reliability, user interfaces, software architecture, software development management, software engineering, software metrics, software portability, software prototyping, software quality, software reliability, software reusability(굵은체는 이질문헌집합, 보통체는 동질문헌집합).

3) http://wordnet.princeton.edu/

용된 세 문헌집합은 사전 데이터 전처리 과정을 거쳤다. 기본적인 처리로는 불용어(stop words)가 제거되었으며, 동일 용어의 이형처리를 위해서 용어 정규화(normalization)과정으로서 Porter(1980)의 스템밍(stemming) 기법을 사용하였다. 마지막으로 처리된 텍스트는 WEKA (Witten & Frank 2000) 시스템으로 실험하기 위하여 WEKA 입력 형태인 ARFF(.arff) 파일 포맷으로 변환되었다.

실험을 위하여서는 자바 프로그램 언어로 구현된 기계학습 어플리케이션인 WEKA(Witten & Frank 2000) 시스템이 선정되었다. WEKA 시스템은 안정성과 신뢰할 수 있는 성능 구현으로 문서 범주화 연구에 빈번히 사용되어 왔다. WEKA에서 구현된 분류기는 사전 실험을 통하여 성능이 우수한 분류기로 밝혀진 지지벡터기계(SVM)[4]과 나이브베이즈[5] 분류기가 최종 선정 및 사용되었다.

3.2 실험 평가

범주화 성능 평가를 위하여 일반적으로 세 가지 측정단위인 재현율, 정확률, F-척도가 주로 사용된다(Lewis 1995; Sebastiani 2002; Yang 1999). 재현율은 모든 적합 문헌수 중에서 선정된 적합 문헌수로 나타내며, 정확률은 옳게 부여된 문서 수와 틀리게 부여된 문서 중에서 부여된 적합 문헌수로 나타낸다. 이러한 재현율과 정확률은 서로 상충되는 부분들이 있어 이를 보완하기 일반적으로 F-척도(van Rijsbergen 1979)로서 재현율과 정확률의 평균을 나타낸다.

4) weka.classifiers.functions.SMO
5) weka.classifiers.bayes.NAIVEBAYES

〈표 2〉 문서 범주화 성능 평가표

	옳은 주제어	틀린 주제어
옳은 주제어 선정	a	b
틀린 주제어 선정	c	d

재현율(R) = a /(a +c)
정확률(P) = a /(a +b)
F 척도 = 2PR /(P +R)

4. 실험결과 분석

4.1 실험 개요

이 연구는 문서범주화 기법의 선정된 자질에 대해서 의미기반 확장을 통해 성능향상을 이루고자 하는데 그 목적이 있다. 이에 적합한 자질 확장을 위한 여러 요소를 사용하여 성능을 비교·분석하였다. 따라서 이 연구의 실험은 세 단계로 수행하였다. 첫째, 베이스라인으로 설정한 기본 실험은 자질 확장 기법을 사용하지 않고 문서범주화 실험을 수행하였다. 기본 실험은 문헌집합의 특성, 문서 요소, 문서범주화 기법과 같은 세 가지 변수에 따라서 수행되었다. 두 번째 실험은 분류어 기준의 자질 확장 실험이 수행되었다. 본 연구의 실험에서 사용된 20개의 분류어를 사용하여 동의어 확장이 실시되었다. 동의어 확장 시에는 각 용어가 지니는 중의성 해소(WSD)에 대한 처리 여부가 변수로 사용되었다. 중의성 해소를 적용하지 않은 방식(no WSD)은 WordNet에서 수록하고 있는 모든 동의어 용어를 선별과정이 없이 모두 사용하였다. 반면에 중의성 해소 과정이 사용된

경우(WSD)는 WordNet에서 수록하고 있는 모든 동의어 중에서 해당 용어와 문맥상의 동일한 의미를 수록하고 있는 용어만으로 자질 확장을 한정하였다. 세 번째 실험은 두 가지 문서범주화 기법인 지지벡터기계와 나이브베이즈를 각각 사용하는 것이다. 본 실험에서 설계한 분류어 기준 동의어 확장 실험이 특정한 문서범주화 기법에 제한적이지 않고, 상대적으로 일반화 가능한지를 파악하고자 하는 목적이다.

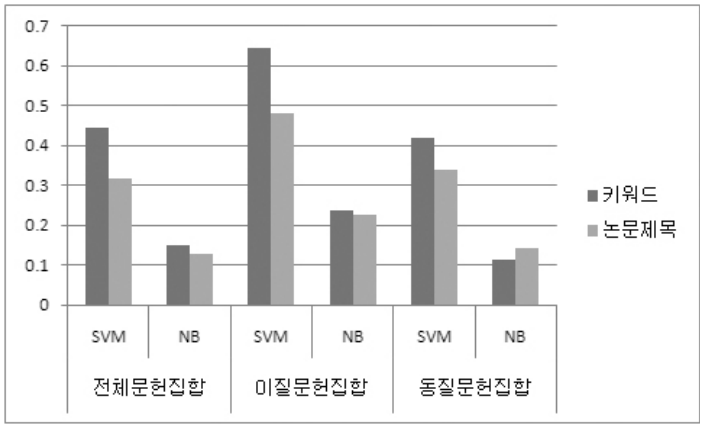
4.2 기본 실험

본 연구의 자질확장에 대한 효용성 평가를 위해서, 자질 확장 과정이 포함되지 않은 기본 실험이 수행되었다. 기본 실험 결과는 〈표 3〉에서 살펴볼 수 있는 바와 같다. 문서범주화 기법, 세 가지 문헌집합, 문헌 구성요소를 사용하여 기본 실험을 수행하였다. 기본 실험결과 매크로 평균 F-척도를 기준으로 살펴볼 수 있다.

이러한 결과는 〈그림 3〉에서 제시된 바와 같이, 세 가지 측면에서 분석될 수 있다. 첫째, 전반적으로 이질문헌집합이 전체문헌집합이나 동질문헌집합에 비교하여 우수한 성능을 나타냈다. 일반적으로 문헌집합 내에 주제 면에서 이질적인 문서를 포함하고 있는 경우와 동질적인 문서를 포함하는 경우와의 성능을 비교하면, 이질적인 문헌집합이 우수한 성능을 나타낸다. 본 실험도 이러한 관련 연구에서의 결과와 마찬가지 실험 결과를 보여주었다. 이러한 결과는 정도의 차이는 있지만 지지벡터기계와 나이

〈표 3〉 매크로 평균 F-값

	전체문헌집합		이질문헌집합		동질문헌집합	
	SVM	NB	SVM	NB	SVM	NB
키워드	0.445	0.148	0.645	0.237	0.421	0.112
논문제목	0.319	0.128	0.481	0.226	0.338	0.144



〈그림 3〉 기본실험결과

브베이즈 기법을 사용하였을 때 모두 일관된 결과를 나타냈다. 둘째, 키워드와 논문제목을 비교하였을 때는 키워드가 논문제목에 비해서 전반적으로 우수한 성능을 나타냈다. 이러한 결과는 상이한 문서범주화 기법을 사용하였을 때와 상이한 문헌집합을 사용하였을 때 모두 일관된 형태로 나타났다. 키워드를 사용하였을 때가 논문제목을 사용하였을 때와 비교하면, 저자에 의해서 선정된 키워드는 논문제목에 비해서 우수한 자질만으로 구성되었다고 볼 수 있다. 따라서 전체적으로 키워드 이질문헌집합을 사용하였을 때, 상대적으로 좋은 성능을 보였다. 셋째, 두 가지 문서범주화 기법에 있어서 성능 차이를 나타냈다. 기존 연구의 결과와 마찬가지로, 지지벡터기계 기법이 나이브베이즈 기법에 비해서 월등히 우수한 결과를 나타냈다. 전체적으로는 두 문서범주화 기법을 사용한 실험 결과가 다른 변수 적용 시에 매우 유사한 형태를 보여준다는 점에서 일관된 실험 결과를 제시하고 있다.

4.3 분류어 기준 자질확장

분류어 기준 자질확장 실험은 분류어 기준으로 하여 분류어와 동의어 관계로 자질 확장하는 것이 문서범주화 성능에 미치는 영향 파악을 위한 것이다. 이를 위해서 해당 분류어의 동의어를 WordNet에서 추출하여 자질로 확장 사용하는 실험을 수행하였다. 분류어에 대한 동의어 추출 시에는 중의성 해소 기법이 적용된 방식과 중의성 해소 기법이 적용되지 않은 방식, 이러한 두 가지 방식을 구분하여 실험이 수행되었다. 〈부록〉은 분류어의 동의어로서 중의성 해소가 적용되어 추출된 자질과 중의성 해소가 적용되지 않고 추출된 자질을 수록하고 있다. 각각 확장된 자질을 사용하여 지지벡터기계와 나이브베이즈 문서범주화 기법으로 자질확장 실험이 수행되었다.

4.3.1 지지벡터기계 기법

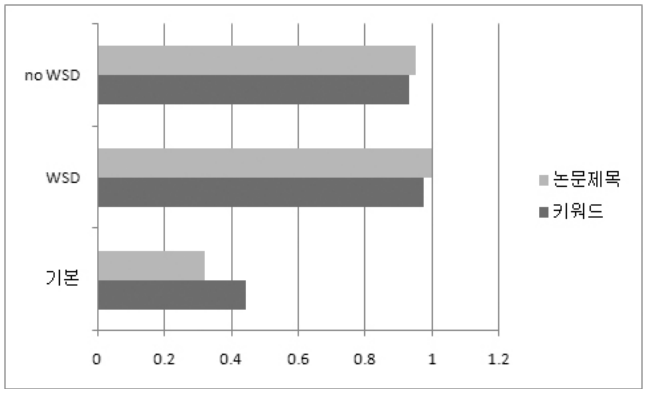
분류어 기준으로 동의어 자질 확장 실험 중 첫 번째 실험은 지지벡터기계 문서범주화 기법을 사용하였다. 전체문헌집합을 사용한 자질확장 실험결과는 〈표 4〉에서 제시하였다. 이 실험은 중의성 해소 적용 여부와 키워드 및 논문제목을 구분하여 수행하였다. 이 실험의 결과는 기본실험과의 비교를 위하여 기본, WSD, no WSD 로 구분하여 매크로 F-값으로 실험 결과가 제시하였다. 〈그림 4〉에서 살펴볼 수 있는 바와 같이 이러한 실험결과를 분류어 기준으로 동의어 자질 확장을 시행하지 않은 기준 실험결과와 비교하였을 때 성능이 월등히 향상되는 결과를 보여준다.

〈표 4〉 자질확장실험(전체문헌)

	기본	WSD	no WSD
키워드	0.445	0.974	0.932
논문제목	0.319	0.997	0.951

(지지벡터기계실험, F-값)

특히 〈표 4〉와 〈그림 4〉에서 살펴볼 수 있는 바와 같이 이 실험 결과에 따라 세 가지 시사점이 제시될 수 있다. 첫째는 기본 실험에 비교해서 월등하게 문서범주화 성능 향상이 이루어졌다. 논문제목을 활용하여 중의성 해소 기법을 적용한 경우에 기본 실험에 비교해서 최대 3배



〈그림 4〉 기본실험과 자질확장실험 비교(전체문헌집합)

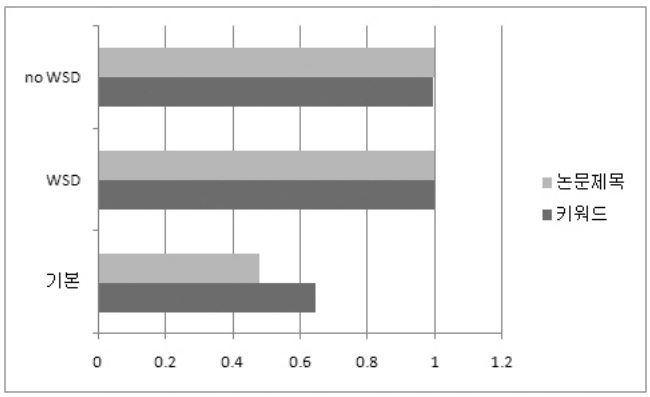
이상(0.319 -〉 0.997) 성능이 향상되었다. 용어의 중의성 해소 적용과 비적용 결과에 있어서는 매우 근소한 차이를 보이고 있다. 둘째, 자질이 확장된 문헌집합을 대상으로 한 실험은 기본 실험에 비교해서, 문서의 구성요소 간의 차이가 근소한 폭으로 줄어들었다. 셋째, 자질확장 실험 결과에서 특기할 만한 사항은 중의성 해소 기법 적용과 비적용 사이의 차이가 매우 근소하다는 점이다.

〈표 5〉 자질확장실험(이질문헌)

	기본	WSD	no WSD
키워드	0.645	0.998	0.996
논문제목	0.481	1.000	1.000

(지지벡터기계실험, F-값)

두 번째 실험은 기본실험 상에서 상대적으로 우수한 이질문헌집합을 대상으로 지지벡터기계 문서범주화 기법을 사용하여 수행하였다. 〈표 5〉와 〈그림 5〉의 실험결과에서 살펴볼 수



〈그림 5〉 기본실험과 자질확장 실험 비교(이질문헌집합)

있는 바와 같이 논문제목을 사용하여 분류어의 동의어를 사용하여 자질을 확장한 경우에는 완벽한 결과인 F-값, 1을 나타내었다. 이러한 실험 결과는 전체문헌집합 실험결과와 비교하여 매우 유사한 패턴으로 보이고 있다. 즉, 급격한 성능향상을 바탕으로 하며, 상이한 문서구성요소 및 중의성 해소 여부의 영향이 매우 미비하다는 점이다.

〈표 6〉 자질확장실험(동질문헌)

	기본	WSD	no WSD
키워드	0.421	0.974	0.936
논문제목	0.338	0.992	0.895

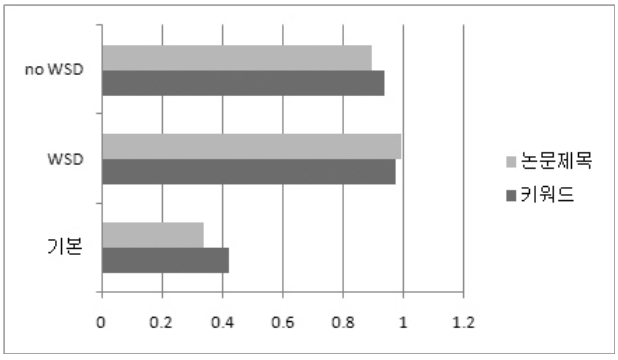
(지지벡터기계실험, F-값)

세 번째 실험은 기본실험에서 살펴볼 수 있는 바와 같이, 성능이 다소 떨어지는 동질문헌집합을 사용하여 수행되었다. 〈표 6〉과 〈그림 6〉에서 제시하는 바와 같이, 일반적으로 성능이 우수하지 못한 동질문헌집합을 사용하여 분류어의 동의어를 사용한 자질확장 결과를 제시하고 있다. 기본실험의 F-값이 상대적으로 낮은 편이었으나, 키워드와 논문제목을 활용하여 자질확장 결과 상당한 성능향상 증가 결과를 보여주고 있다. 동질문헌집합 실험에서 특기할 만한 사항은 중의성 해소 적용과 비적용의 차이가 전체문헌집합과 이질문헌집합과 비교하여 상대적으로 우세한 차이를 나타났다는 점이다.

지금까지 살펴본 바와 같이, 전반적으로 분류어 기반 동의어 자질확장 실험은 기본실험과 비교하여 상당한 성능향상을 보여준다. 이러한 성능향상은 상이한 문서의 구성요소인 키워드와 논문제목 및 용어의 중의성 해소 적용과 비적용에 따른 차이를 상쇄하는 것으로 분석된다.

4.3.2 나이브베이즈 기법

앞서 실험에서 살펴본 바와 같이 지지벡터기계 문서범주화 기법을 사용한 결과 기본실험에 비교하여 월등히 우수한 성능증가를 보여주었다. 본 절에서는 이러한 성능향상이 특정 문서범주화 기법에 제한되지 않고 일반화될 수 있는 현상인지를 파악하기 위해서, 나이브베이즈 문서범주화 기법을 사용한 분류어 기준 동의어 자질확장 실험이 수행되었다.



〈그림 6〉 기본실험과 자질확장 실험 비교(동질문헌집합)

첫째, 전체문헌집합을 대상으로 실험한 결과가 〈표 7〉과 〈그림 7〉에서 제시된 바와 같다. 나이브베이즈 기법을 사용한 기본실험의 결과는 상대적으로 성능이 낮은 편이었으나, 중의성 해소 기법이 적용된 실험에서 월등한 성능향상을 보여주었다. 중의성 해소 기법이 적용되지 않는 실험에서도 역시 성능향상을 보여주었으나, 중의성 해소 기법 적용 시와는 상당한 차이를 나타냈다는 점에서 지지벡터기계 문서범주화 기법 실험과 구별된다.

〈표 7〉 자질확장실험(전체문헌)

	기본	WSD	no WSD
키워드	0.148	0.825	0.589
논문제목	0.128	0.830	0.631

(나이브베이즈, F-값)

둘째, 나이브베이즈 문서범주화 기법을 사용한 실험이 이질문헌집합을 사용하여 수행되었다. 〈표 8〉와 〈그림 8〉에서 제시하는 바와 같이, 기본실험과 비교하여 역시 월등한 성능향상을 보여주고 있다. 논문 제목을 사용한 실험에서는 용어의 중의성 해소 적용과 비적용에 있어서 미비한 차이를 보여주고 있으나, 키워드를 사용한 실험에서는 비교적 큰 차이를 나타내고 있다.

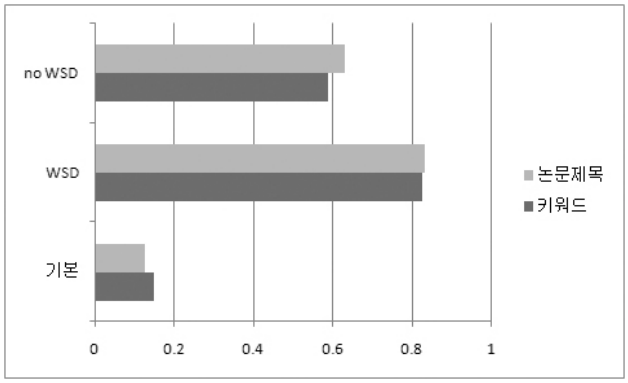
〈표 8〉 자질확장실험(이질문헌집합)

	기본	WSD	no WSD
키워드	0.237	0.992	0.775
논문제목	0.226	0.994	0.962

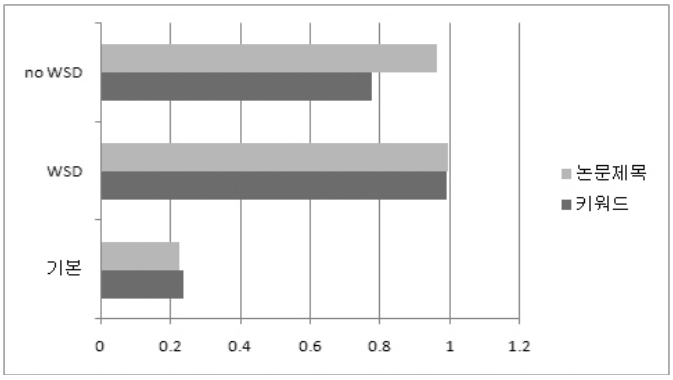
(나이브베이즈, F-값)

세 번째 실험은 동질문헌집합을 사용하여, 기본실험과 자질확장 실험을 비교하였다.

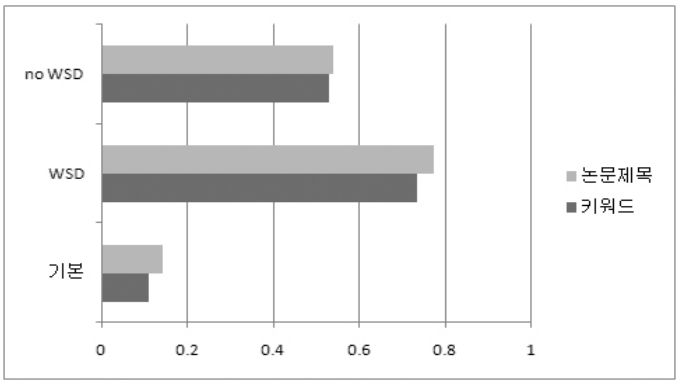
〈표 9〉과 〈그림 9〉에서 제시하는 바와 같이 기본 실험 성능에 비교하여 자질확장 실험은 일관되게 성능이 향상된 것으로 나타났다. 이러한 현상은 키워드와 논문제목 사용 시에 동일한 패턴으로 제시되었다. 동일문헌집합 실험에서는 이질 문헌집합 실험과는 달리, 중의성 해소 기법 적용과 비적용 사이에 성능 향상에 있어서 뚜렷한 차이를 보인다는 점이다.



〈그림 7〉 기본실험과 자질확장실험비교(전체문헌집합)



〈그림 8〉 기본실험과 자질확장실험비교(이질문헌집합)



〈그림 9〉 기본실험과 자질확장실험 비교(동질문헌집합)

〈표 9〉 자질확장실험(동질문헌집합)

	기본	WSD	no WSD
키워드	0.112	0.734	0.530
논문제목	0.144	0.773	0.538

(나이브베이즈, F-값)

전반적으로 나이브베이즈 문서범주화 기법을 사용한 실험 결과, 자질확장이 월등한 문서범주화 성능향상에 기여한 것으로 나타났다. 성능향상에 있어서 중의성 해소 적용 여부와 상이한 문서구성요소 사용과 같은 변수에 대해서 지지벡터기계와 비교하여, 보다 뚜렷한 차이를 보인다는 점이 특기할 만하다. 이러한 차이를 보이는 것은 지지벡터기계의 성능 실험 결과가 변수 간에 차이를 보이지 않을 만큼 증가했기 때문으로 여겨진다.

5. 결 론

기계학습 기반 문서범주화의 성능은 최적의 자질 요소를 규명하는 것에 의해서 영향을 받

는 것이 일반적이다. 본 연구는 일반적인 학술지 논문의 중요한 구성요소인 논문의 키워드(저자부여)와 논문제목을 활용하여 최적의 자질요소를 분석 및 파악하고자 하였다. 특히 논문의 키워드나 논문제목과 같이 성격상 정제된 자질로 구성된 구성요소를 활용하여, 의미관계로 연계된 동의어 용어로의 자질 확장이 문서범주화 성능에 미치는 영향을 파악하고자 하였다. 이를 위해 본 연구는 두 가지 실험을 수행하였다. 첫 번째 실험은 기본실험으로서 자질확장 없이 수행되었다. 키워드와 논문제목을 활용하였으며, 지지벡터기계와 나이브베이즈 문서범주화 기법을 각각 사용하여 실험하였다. 이 실험의 결과는 자질확장 실험 결과와 비교되었다. 두 번째 실험은 분류어를 기준으로 하여 동의어 의미관계로 자질확장이 이루어졌으며, 확장 시에는 용어에 대한 중의성 해소 기법이 적용된 자질 요소와 중의성 해소 기법이 적용되지 않은 자질 요소로 구분되었다. 이 두 가지 방식으로 확장된 자질 요소로서 지지벡터기계와 나이브베이즈 문서범주화 기법을 사용하여 실험하였다.

자질확장 실험 결과, 전반적으로 분류어를 기준으로 하여 동의어 관계의 용어로 자질확장을 수행한 실험이 자질확장을 수행하지 않은 기본실험에 비교하여 월등한 성능향상을 보여주었다. 이러한 실험 결과는 세 가지 측면에서 논의될 수 있다. 첫째는 본 연구는 두 가지 문서범주화 기법인 지지벡터기계와 나이브베이즈를 사용하여 실험하였다. 이 두 가지 상이한 문서범주화 기법을 사용한 실험 결과 정도의 차이는 있지만, 일관된 성능향상을 보여주었다. 이러한 상이한 문서범주화 기법을 사용한 경우에 일관된 성능향상 결과는 동의어 자질확장이 문서범주화 성능에 긍정적인 영향을 미친다는 연구결과의 일반화에 영향을 미친다고 볼 수 있다. 둘째, 본 연구는 두 가지 문서의 구성요소인 키워드와 논문제목을 사용하여 실험하였다. 이 두 가지 문서 구성요소는 역시 문서범주화 성능에 있어서 향상에 기여할 수 있음을 보여준다. 이러한 실험 결과는 의미는 논문제목의 활용에 있어서, 상당히 의미 있는 시사점을 제시한다. 저자가 제공하는 키워드와는 달리, 논문제목만을 활용하여 일정수준 이상의 성능향상을 이룰 수 있다는 것은 문서범주화 기법의 실제적인 시스템 적용 및 구현에 있어서 중요하다고 볼 수 있다. 그러나 구체적으로 기본 실험 상에서 키워드가 더 우수한 자질 요소로 나타났으나, 자질 확장 실험 후에는 논문제목이 더 우수한 결과를 보이기도 한다. 이러한 복합적인 결과에 대해서는 향후 후속 연구를 통해 의미 있는 결과를 밝히는 것이 바람직하다고 여겨진다. 셋째, 본 연구에서는 중의성 해소 적용과 비적용을 구분하여 실험하였는데, 성능향상에 모두 긍정적인 영향을 끼친 것으로 파악되었으며, 이 두 가지 방식이 문서범주화 성능향상에 명확한 차이를 발견하지 못하였다.

본 연구의 이러한 결과는 문서범주화 기법이 활용되는 분야인 필터링, 정보검색, 정보조직 등의 분야에서 활용될 수 있는 것으로 기대할 수 있다. 또한 향후 후속연구로서 보다 다양한 문서의 구성요소 적용, 하위어 및 상위어를 포함하는 용어의 계층적 관계 분석, 중의어 해소 기법 적용이 미치는 명확한 영향 등에 대한 보다 포괄적이며 세밀한 연구가 필요하다고 여겨진다.

참 고 문 헌

이재윤. 2005. 자질선정 기준과 가중치 할당 방식 간의 관계를 고려한 문서 자동분류의 개선에 관한 연구. 『한국문헌정보학회지』, 39(2): 123-146.

Barak, L., I. Dagan, & E. Shnarch, 2009. "Text categorization from category name via lexical reference." *Proceedings of NAACL HLT 2009: Short Papers*, 33-36.

Bloehdorn, S., & A. Hotho. 2004. Boosting for text classification with semantic features. Proceedings of the MSW 2004 Workshop at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Brank, J., M. Grobelnik, N. Milic-Frayling, & D. Mladenic. 2002. Interaction of feature selection methods and linear classification models. Proceedings of the ICML Workshop on Text Learning.

de Buenaga Rodriguez, M., J. Gomez-Hidalgo, & B. Diaz-Agudo. 1997. Using WordNet to complement training information in text categorization. In the Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing, 150-157.

Chen, J., H. Huang, S. Tian, & Y. Qu. 2009. Feature selection for text classification with Naive Bayes, Expert Systems with Applications, 36: 5432-5435.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Forman, G. 2003. "An extensive empirical study of feature selection metrics for text classification." *Journal of Machine Learning*, 3: 1289-1305.

John, G. H., R. Kohavi, & K. Pfleger. 1994. "Irrelevant features and the subset selection problem." *Proceedings of the 11th International Conference on Machine Learning*, 121-129.

Kehagias, A., V. Petridis, V. G. Kaburlasos, & P. Fragkou. 2001. *A comparison of word-and sense-based text categorization using several classification algorithms*. Journal of Intelligent Information System.

Lewis, D. D. 1995. Evaluating and optimizing autonomous text categorization systems. Unpublished Doctoral Dissertation, University of Massachusetts, Massachusetts.

Miller, G. 1995. "WordNet: A lexical database for English." *Communications of the ACM*, 38(11): 39-41.

Mansuy, T., & R. J. Hilderman. 2006. Evaluating WordNet features in Text Classification models.

Rosso, P., E. Ferretti, D. Jimenez, & V. Vidal.

2004. Text categorization and information retrieval using WordNet senses, Proceedings of GWC2004, 299-304.

Scott, S., & S. Matwin. 1998. Text classification using WordNet Hypernyms. In the Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems, 45-52.

Sebastiani, F. 2002. Hypertext categorization. In A. Zanasi(Eds.), *Text Mining and Its Applications*(109-129), Southhampton, U.K.: WIT Press.

Sebastiani, F. 2005. Text categorization. In A. Zanasi(Eds.), Text mining and its applications(109-129), Southhampton, U.K.; WIT Press.

van Rijsbergen, C. J. 1979. *Information Retrieval.* Butterworths, London.

Verikas, A., & M. Bacauskiene. 2002. "Feature selection with neural networks." *Pattern Recognition Letters*, 23: 1323-1335.

Witten, I. H., & E. Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations.* CA: San Diego, Academic Press.

Yang, Y. 1999. "An evaluation of statistical approaches to text categorization." *Information Retrieval*, 1: 69-90.

〈부록〉 분류어 기준 동의어 확장(중의성 적용 및 비적용)

분류어	WSD	no WSD
software architecture	code structure	code building discipline profession structure
software development management	code improvement administration	code improvement process change use district improvement
software libraries	code collection	code room collection depository collection building
software maintenance	code repair	code repair support support payment support wrongdoing
software metrics	code measure	code function unit of measurement measure
software portability	code movability	code movability
software prototyping	code model	code model
software quality	code degree	code attribute degree property sound property social station
software reliability	code responsibility	code responsibility
software resuability	code	code
computer architecture	machine structure	machine expert building discipline profession structure
computer graphics	machine graphic art	machine expert artwork graphic art
computer interfaces	machine program	machine expert surface program overlap computer circuit
discrete systems	separate instrumentality	separate group instrumentality method plan of action body
information management	message administration	message collection cognition information measure accusation
knowledge based systems	psychological feature instrumentality	psychological feature group instrumentality method plan of
pattern recognition	structure understanding	structure activity decoration practice exemplar plan path graph acceptance memory approval understanding organic phenomenon diplomacy credence appointment
reliability	responsibility	responsibility
software engineering	code application	code application discipline room
user interfaces	person program	person selfish person individual surface program overlap computer circuit