위키피디아를 이용한 분류자질 선정에 관한 연구

An Experimental Study on Feature Selection Using Wikipedia for Text Categorization

김용환(Yong-Hwan Kim)*
정영미(Young-Mee Chung)**

초 록

텍스트 범주화에 있어서 일반적인 문제는 문헌을 표현하는 핵심적인 용어라도 학습문헌 집합에 나타나지 않으면 이 용어는 분류자질로 선정되지 않는다는 것과 형태가 다른 동의어들은 서로 다른 자질로 사용된다는 점이다. 이 연구에서는 위키피디아를 활용하여 문헌에 나타나는 동의어들을 하나의 분류자질로 변환하고, 학습문헌 집합에 출현하지 않은 입력문헌의 용어를 가장 유사한 학습문헌의 용어로 대체함으로써 범주화 성능을 향상시키고자 하였다. 분류자질 선정 실험에서는 (1) 비학습용어 추출 시 범주 정보의 사용여부, (2) 용어의 유사도 측정 방법(위키피디아 문서의 제목과 본문, 카테고리 정보, 링크 정보), (3) 유사도 척도(단순 공기빈도, 정규화된 공기빈도) 등 세 가지 조건을 결합하여 실험을 수행하였다. 비학습용어를 유사도 임계치 이상의 최고 유사도를 갖는 학습용어로 대체하여 kNN 분류기로 분류할 경우 모든 조건 결합에서 범주화 성능이 0.35%~1.85% 향상되었다. 실험 결과 범주화 성능이 크게 향상되지는 못하였지만 위키피디아를 활용하여 분류자질을 선정하는 방법이 효과적인 것으로 확인되었다.

ABSTRACT

In text categorization, core terms of an input document are hardly selected as classification features if they do not occur in a training document set. Besides, synonymous terms with the same concept are usually treated as different features. This study aims to improve text categorization performance by integrating synonyms into a single feature and by replacing input terms not in the training document set with the most similar term occurring in training documents using Wikipedia. For the selection of classification features, experiments were performed in various settings composed of three different conditions: the use of category information of non-training terms, the part of Wikipedia used for measuring term-term similarity, and the type of similarity measures. The categorization performance of a kNN classifier was improved by 0.35~1.85% in F_1 value in all the experimental settings when non-learning terms were replaced by the learning term with the highest similarity above the threshold value. Although the improvement ratio is not as high as expected, several semantic as well as structural devices of Wikipedia could be used for selecting more effective classification features.

키워드: 자동 분류, 자질 선정, 위키피디아, 용어 간 유사도, 비학습용어
text categorization, document classification, feature selection, Wikipedia, term similarity, non-learning term

* 연세대학교 대학원(kimyonghwan@yonsei.ac.kr) (제1저자)
** 연세대학교 명예교수(ymchung@yonsei.ac.kr) (공동저자)

■ 논문접수일자 : 2012년 5월 21일 ■ 최초심사일자 : 2012년 5월 27일 ■ 게재확정일자 : 2012년 6월 16일
■ 정보관리학회지. 29(2). 155-171, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.2.155〕

1. 서 론

웹 환경에서 정보의 양이 끊임없이 늘어남에 따라서 나타나게 된 정보과부하 문제는 효과적인 문헌 자동 분류에 대한 요구로 이어졌다. 특히 지도학습 기반의 자동분류 기법인 텍스트 범주화에 대한 연구는 2000년대 들어 매우 활발하게 진행되고 있다.

일반적인 텍스트 범주화에서 분류대상이 되는 문헌은 문헌 내에 출현한 용어만을 고려하여 색인하는 BOW(Bag Of Words) 방식에 의해 n차원 용어 벡터로 표현되고, 분류 알고리즘에서 학습문헌 집합과 입력문헌에 공통적으로 출현하는 자질을 사용하여 학습문헌 집합과 입력문헌 간의 유사도를 측정하거나 입력문헌이 특정 주제범주에 포함될 확률 값을 계산하여 입력문헌에 주제범주를 배정하게 된다.

그러나 위와 같은 텍스트 범주화 과정에는 두 가지 중요한 문제점이 있다. 첫째, 분류자질에 관한 문제이다. 특정 범주를 표현할 수 있는 용어가 입력문헌에는 있지만 학습문헌 집합에 없다면, 그리고 입력문헌의 용어가 학습문헌 집합에 없는 동의어, 유의어 또는 신조어라면, 해당 용어는 범주화 자질로 선정되지 않아 분류과정에 영향을 줄 수 없거나 해당 문헌을 적합하지 않은 범주로 배정할 가능성이 커진다. 둘째, BOW 방식의 문제점을 들 수 있다. Wang과 Domeniconi(2008)는 BOW 방식이 서로 의미는 같으나 형태가 다른 동의어들을 각각 다른 자질로 표현한다는 점을 지적하였다. 이에 대한 해결책 중의 하나로 어간/어근 추출 방법(stemming)이 제시되기도 하였지만, 이 방법에서는 서로 다른 뜻을 가지는 용어가 비슷한 형태를 가지게 된다면 동일한 자질로 간주되는 문제가 나타날 수 있다. 그리고 반대의 경우로 본래의 용어와 약어의 경우에는 어간추출을 사용하더라도 형태가 다르기 때문에 서로 다른 자질로 간주된다. 결과적으로 하나의 문헌 내에 형태가 다른 동의어가 존재하게 되면 문헌 간 유사도 측정시 상대적으로 낮은 유사도 값을 가진다. 따라서 이 두 가지 문제점들은 텍스트 범주화 과정에서 성능을 떨어뜨리는 요인이 된다.

본 연구의 목적은 텍스트 범주화 과정에서 나타나는 한계점을 극복하기 위하여 동일 문헌에 출현하는 동의어들을 단일 자질로 통합시키고, 학습문헌 집합에 출현하지 않은 입력문헌의 용어를 가장 유사한 학습문헌의 용어로 대체하여 분류자질로 사용함으로써 범주화의 성능을 향상시키는 것이다. 동의어를 통합시키거나 유사한 용어로 대체하기 위해서는 문헌에 나타나는 모든 용어들의 의미와 용어와 용어 간의 관계를 확인할 수 있어야 하므로 외부 정보자원의 활용이 필수적이다.

연구에 사용될 외부 정보자원의 요건으로는 일반 분야의 다양한 용어들을 포함하고 있어야 하고, 각각의 용어에 대한 설명이 있어야 하며, 용어 간의 관계가 표현되어 있거나, 이러한 관계를 표현할 수 있는 장치를 포함하고 있어야 한다. 위키피디아는 웹 백과사전으로 이러한 요건을 충족하고 있다. 일반 분야의 많은 용어들을 정의하고 있을 뿐만 아니라 위키피디아가 가지고 있는 고유의 의미적, 구조적 장치들(본문, 카테고리, 링크 정보 등)은 다양한 용어 간의 관계를 파악하기에 적합하다. 따라서 본 연구에서는 위키피디아를 외부 정보자원으로 사용하여 실험을 진행하였다.

2. 위키피디아의 구조 및 활용

위키피디아의 구조는 제목(개념)과 본문, 카테고리, 링크, 리디렉션(redirection), 동음이의어 페이지(disambiguation page), 템플릿, 토론, 역사로 구성되어 있다. 실험에서는 템플릿, 토론, 역사를 제외한 나머지 구조를 활용하였다.

제목과 본문은 위키피디아가 웹 백과사전으로서 가지는 가장 기본적인 구조라고 할 수 있다. 제목은 단일 개념으로 명사구로 되어 있으며, 제목에 나타난 개념이 여러 의미를 포함할 경우에는 주석(괄호)을 통해 식별한다. 본문은 단일 개념을 설명하는 글로서 해당 개념에 대한 정의, 역사, 어원, 같이 보기(See also) 등 관련 문서를 나타내는 내용, 그리고 참고(reference)가 단락별로 구분되어 있다.

카테고리는 2004년 이후 나타난 장치로 문서를 작성한 저자에 의해서 추가된다. 따라서 단순 카테고리의 기능뿐만 아니라 태그로서의 기능도 갖고 있으며, 복수로 설정할 수 있다. 위키피디아에서 카테고리 구조는 기본적으로 단순 계층 구조로 이루어지나 일부에서는 순환 계층 구조로 이루어진다.

링크는 내부 링크와 외부 링크로 구분된다. 내부 링크는 위키피디아 문서 본문에 포함된 용어를 설명하기 위한 것으로 해당 용어를 제목으로 하는 다른 위키피디아 문서로 연결해 주는 링크이다. 외부 링크는 설명하고자 하는 개념과 관련된 외부 문서로 연결하는 링크이다.

리디렉션과 동음이의어 페이지는 동의어와 유의어의 연결, 동음이의어 식별 등에 있어서 활용할 수 있는 위키피디아의 의미적·구조적 장치 중 하나이다. 리디렉션은 위키피디아를 통해 용어를 검색하는 경우, 해당 개념에 대해 동일한 내용을 가지는 대표적인 개념으로 연결시켜 주는 역할을 한다. 동음이의어 페이지는 검색 용어가 여러 가지 의미를 가지고 있을 때, 각각의 의미에 대한 간략한 설명과 링크를 보여주는 페이지로 용어 검색 시 보다 정확한 의미를 검색할 수 있다.

위키피디아는 사전, 온톨로지, 시소러스 등으로 많이 활용되지만, 용어 간 유사도 측정 또는 텍스트 범주화를 위해 활용한 선행 연구들은 크게 세 가지 유형으로 구분하여 살펴 볼 수 있다. 첫째는 위키피디아를 이용하여 용어 간의 연관성을 측정하기 위해 실험문헌 집단의 용어에 적합한 위키피디아 개념을 매핑하는 방법에 관한 것으로서 실험문헌 집단의 용어를 위키피디아의 제목에서 검색하여 직접적으로 개념과 매핑시킨 연구(Milne, Witten, & Nichol, 2007; Ponzetto & Strube, 2006, 2007), 하나의 용어를 하나의 개념이 아닌 다수의 개념과 매핑시킨 연구(Gabrilovich & Markovitch, 2006), 앵커텍스트의 일반성(commonness)과 연관성(relatedness)을 사용하여 실험 집단의 용어와 위키피디아의 개념을 매핑시킨 연구(Huang, Milne, Frank, & Witten, 2009; Milne & Witten, 2008), 행렬 연산 방식을 이용하여 문헌과 개념을 매핑한 연구(Minier, Bodo, & Csato, 2007) 등이 있다.

둘째는 위키피디아의 구조적 장치를 이용하여 용어 간의 유사도 또는 관련도를 측정한 연구들로 실험대상인 용어들을 위키피디아에 매핑시킨 후, 계층구조에서 두 위키피디아 문서 간의 거리로써 용어의 관련도를 측정한 연구(Ponzetto & Strube, 2007), 실험 대상 용어들과 매핑된

위키피디아 개념 벡터 간의 유사도를 산출하여 용어의 관련도를 측정한 연구(Gabrilovich & Markovitch, 2006), 매핑된 위키피디아 문서의 내부링크를 요소로 하는 벡터를 만들어 위키피디아 문서 간의 유사도를 용어의 관련도로 측정한 연구(Milne, Witten, & Nichol, 2007), 매핑된 위키피디아 문서의 링크를 사용하여 생성한 벡터에 구글 정규화 계수를 적용하여 산출한 유사도를 용어 간의 관련도로 측정한 연구(Milne & Witten, 2008) 등이 있다.

셋째는 실험문헌 집단의 자질들을 위키피디아의 개념, 또는 카테고리와 같은 특징들로 변경하거나 자질을 추가하는 방식으로 범주화 실험을 수행한 연구들이다. 문헌에 매핑된 위키피디아 개념들을 범주화 자질로 추출한 뒤 SVM 분류기를 통하여 텍스트 범주화 실험을 실시한 연구(Gabrilovich & Markovitch, 2006), 문헌을 위키피디아의 개념 벡터로 구성한 다음 차원축소기법을 사용하여 텍스트 범주화 실험을 실시한 연구(Minier, Bodo, & Csato, 2007), 각각의 용어와 매핑된 위키피디아 문서에 내재되어 있는 내부링크, 리디렉션, 카테고리 정보를 이용하여 유의어, 동의어를 추출한 다음 자질에 추가하여 범주화 실험을 실시한 연구(Wang, Hu, Zeng, Chen, & Chen, 2007) 등이 있다.

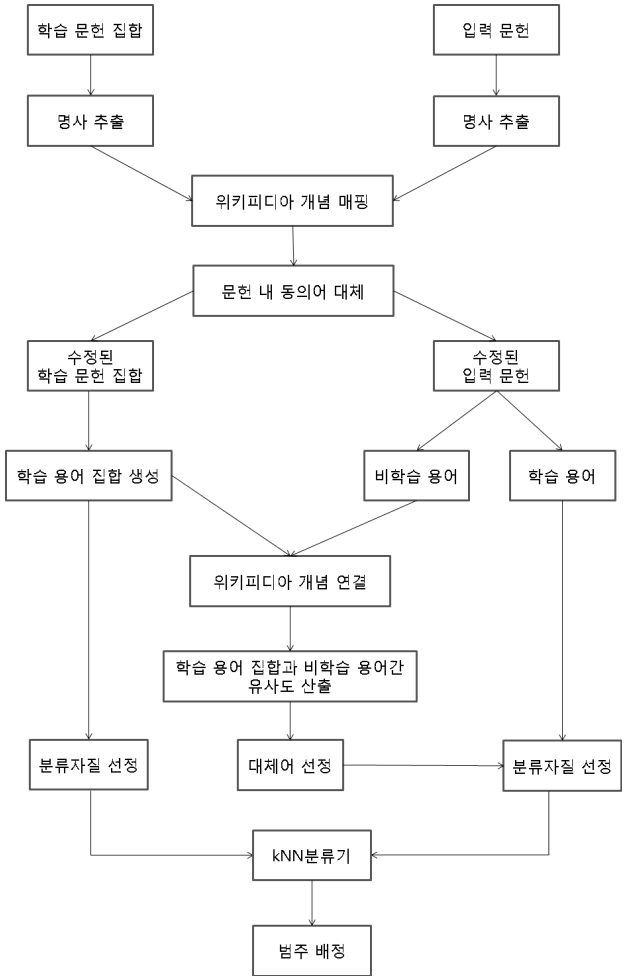
3. 실험 설계

3.1 실험 개요

본 연구에서는 Reuters-21578을 실험문헌 집단으로 사용하여 분류자질을 선정하는 실험을 실시하였다. 이 실험에서는 모두 135개의 주제범주 가운데 해당문헌이 가장 많은 상위 10개의 범주에 속하는 5,149개의 학습문헌과 2,040개의 입력문헌을 사용하였다. 제목과 본문에 대한 색인에는 파이썬 기반의 통계적 자연언어 처리 도구인 NLTK(Natural Language ToolKit)를 이용하였다(Bird, Klein, & Loper, 2009). 먼저 각 기사의 내용을 문장별로 분리한 다음 NLTK의 기능 중에서 chunk_tag 기능을 사용하여 구문 분석을 실시한 후 복합명사를 포함한 명사와 명사구만을 추출하였다. 색인 과정에서는 앞에서 언급하였듯이 어간추출과 같은 방법으로 인해 생겨나는 문제점을 방지하기 위해서 추가적인 자연언어 처리를 적용하지 않았다. 또한 용어의 고유한 의미를 보존하고 각각의 용어를 효과적으로 위키피디아 문서와 매핑하기 위해 문헌에 출현한 용어를 그대로 추출하였고 철자 오류도 수정하지 않았다.

색인 후, 텍스트 범주화 성능을 향상시키기 위하여 문헌에 나타나는 동의어들을 단일 용어로 대체하고, 분류자질로서 학습문헌에 없는 입력문헌의 용어를 추출하여 가장 유사한 학습문헌의 용어로 대체하여 사용하였다. 이 연구에서는 학습문헌에 없는 입력문헌의 용어를 '비학습용어', 학습문헌에 있는 용어들을 '학습용어'로 정의하여 사용하였다. 입력문헌과 학습문헌에 모두 출현한 용어들은 '학습용어'로 정의하였다.

〈그림 1〉은 실험의 전반적인 과정을 도식화한 것이다. 먼저 학습문헌 집합과 입력문헌 집합의 용어를 위키피디아 검색을 통하여 가장 적합한 개념으로 매핑한다. 그리고 동일 문헌에 출현한 용어들 중에서 같은 위키피디아 개



〈그림 1〉 실험 개요

념으로 매핑되는 용어들이 나타나게 되면 이들을 동일한 하나의 용어로 대체한 후 수정된 학습문헌 집합과 입력문헌 집합을 생성한다.

수정된 입력문헌과 학습문헌 집합을 사용하여 비학습용어를 추출한다. 추출 방법은 범주 정보의 활용 여부에 따라서 전체 학습문헌 집합의 용어와 비교하여 추출하는 방법과 범주별 학습문헌 집합의 용어와 비교하여 추출하는 방법 2가지로 나뉜다.

추출된 비학습용어들은 학습용어와의 유사도를 산출하게 된다. 유사도 측정 방법은 각각의 용어와 매핑된 위키피디아 문서의 의미적, 구조적 장치들인 (1) 개념을 표현하는 제목과 본문, (2) 개념의 카테고리 정보, 그리고 (3) 본문 내에서 내부와 외부로 이어지는 링크 정보들을 사용하였다. 유사도 측정에 사용한 척도로는 단순 공기빈도와 정규화된 공기빈도인 다이스 계수이다.

모든 용어 간의 유사도 측정이 완료되면 비학습용어와 가장 유사도가 높은 하나의 학습용어를 선택하여 비학습용어를 대체하고 이를 입력문헌의 분류자질로서 사용하여 분류 실험을 진행하였다. 실험에 사용된 분류기는 kNN 분류기로 사전실험 시 k=3일 때 가장 좋은 성능을 나타냈으므로 이후 모든 실험에서 k는 3으로 설정하였다. 그리고 분류기에서 문헌과 문헌 간의 유사도를 산출하기 위한 유사계수로서 코사인 유사계수를 사용하였다. 분류 성능의 평가 척도로는 정확률, 재현율, F_1 척도의 매크로 평균을 사용하였다.

3.2 자질 선정 과정

3.2.1 위키피디아 개념 매핑과 단일 용어 선택

이 연구에서는 XML 형식으로 되어 있는 위키피디아 데이터를 MySQL DB에 입력하여 개념과 용어 매핑에 사용하였다. 따라서 이 위키피디아 DB의 제목 필드에서 실험집단의 용어를 검색하여 나타난 용어와 위키피디아 문서를 매핑하였다. 제목 필드에서 용어가 검색될 경우, 해당 제목을 가지는 위키피디아 문서의 본문과 카테고리, 링크 정보를 사용하여 용어 간 유사도를 산출하게 된다. 그러나 다음의 4가지 경우 용어와 위키피디아 문서 매핑에 문제가 생긴다.

첫째, 검색된 용어에 대한 리디렉션이 있는 경우이다. 이 경우 위키피디아에서는 자동적으로 리디렉션을 통해 이동한 문서를 결과로서 보여준다. 실험에서는 검색된 레코드에 리디렉션이 있다면 이를 통해 이동한 문서의 본문 또는 카테고리, 링크 정보를 검색한 용어의 정보로 간주하여 매핑하였다.

둘째, 철자 오류가 있는 경우이다. Reuters 실험문헌 집단에서는 약 500여개 이상의 철자 오류가 발견되었다. 이를 해결하기 위해서 위키피디아의 추천어 장치를 사용하였다. 위키피디아 홈페이지에서는 용어 검색 시 철자 오류라고 여겨지는 경우 가장 유사한 용어를 추천어로 제시한다. 따라서 이 경우에는 위키피디아에서 제시해주는 추천어로 철자 오류를 수정하고 재검색하여 나온 위키피디아 문서의 정보를 해당 용어와 매핑하여 사용하였다.

셋째, 동음이의어 페이지(disambiguation page)가 검색된 경우이다. 동음이의어 페이지를 무시하게 되면 각각의 용어를 정확한 위키피디아 문서로 매핑하는 것이 어렵기 때문에 실험에 영향을 줄 수 있다. 따라서 이 실험에서는 동음이의어 페이지에 나오는 내부 링크에 의해 연결되는 문서 모두를 수집하여 동음이의어를 식별하였다. 학습문헌 집합에 있는 용어인 경우에는 해당 문헌의 용어가 포함되어 있는 범주의 모든 용어들과 동음이의어로 연결되는 모든 위키피디아 문서들을 비교하여 유사도 값이 가장 높은 위키피디아 문서를 해당 용어와 매핑시켰다. 입력문헌의 용어인 경우에는 검색한 용어가 포함되어 있는 문헌내의 용어와 비교하여 가장 유사도가 높은 위키피디아 문서를 매핑하였다. 이때 사용한 유사도 척도는 코사인 유사계수이다.

넷째, 해당 용어가 위키피디아 DB에서 검색이 되지 않는 경우로 이러한 용어들은 유사도 측정 기준이 없기 때문에 용어 간 유사도 측정 대상에서 제외하였다.

〈표 1〉은 위키피디아에 매핑된 용어의 수를 나타낸 것으로 학습문헌 집합의 용어에서는 15,399개이고 입력문헌 집합의 용어에서는 8,472

〈표 1〉 위키피디아 개념에 매핑된 용어의 수

전체 용어수		철자오류 수정		위키피디아와 매핑된 용어의 수	
학습문헌	입력문헌	학습문헌	입력문헌	학습문헌	입력문헌
27,594	13,093	3,375	1,690	15,399	8,472

개이다.

모든 용어를 위키피디아 문서와 매핑하고 나면, 동일한 문헌에 출현한 몇몇 용어들이 같은 위키피디아 문서로 매핑되는 것을 확인할 수 있다. 두 개 이상의 용어가 동일한 위키피디아 문서로 매핑된다면, 이 용어들은 동의어라고 할 수 있기 때문에 분류자질로 더 적합한 하나의 용어로 집중시킬 수 있다. 학습문헌 집합에서 이러한 경우가 나타난다면, 카이제곱 통계량이 높은 용어로 동의어를 대체하였다. 입력문헌의 경우, 해당 용어들이 학습용어인 경우에는 카이제곱 통계량에 근거하여 용어를 대체하였으나, 비학습용어인 경우에는 단일 용어로 대체하지 않았다.

3.2.2 입력문헌에서 비학습용어의 추출

입력문헌에서 학습문헌에 출현하지 않은 비학습용어를 추출하는 방법은 범주 정보의 사용여부에 따라서 2가지로 나누어 볼 수 있다. 첫째로 학습문헌 전체를 하나의 집합으로 보고 학습문헌 집합 전체의 용어와 입력문헌의 용어와 비교하여 비학습용어를 추출하는 것이다. 이것은 학습문헌의 범주 정보를 이용하지 않은 방법으로 하나의 입력문헌이 하나의 비학습용어 집합을 가지게 되는 방법이다.

둘째는 각각의 범주를 하나의 집합으로 보고 이를 입력문헌의 용어와 비교하여 비학습용어를 추출하는 것이다. 학습문헌의 범주 정보를 이용하는 방법으로 하나의 입력문헌이 각각의 범주별로 비학습용어 집합을 가지게 된다.

3.2.3 위키피디아의 의미적·구조적 장치 활용

위키피디아의 의미적·구조적 장치들 중에서 다음의 3가지 장치를 활용하여 학습용어와 비학습용어 간 유사도 값을 측정하였다.

(1) 제목과 본문의 활용

각각의 용어와 연결된 위키피디아 문서의 제목과 본문을 활용하여 두 용어 간의 유사도를 측정할 수 있다. 위키피디아 문서의 본문에는 해당 용어에 대한 정의와 설명이 포함되어 있다. 따라서 두 용어와 매핑되어 있는 위키피디아 문서 사이의 유사도를 측정하게 되면 이를 두 용어 사이의 유사도로 간주할 수 있다. 위키피디아 문서 간 비교를 위해서 본문은 실험문헌 집단과 동일하게 명사와 명사구를 추출하여 이를 요소로 하는 벡터로 표현하였고, 단순 TF (term frequency) 가중치를 사용하였다.

(2) 카테고리 정보의 활용

각각의 용어와 연결된 위키피디아 문서의 카테고리 정보를 활용하여 두 용어 간의 유사도를 측정할 수 있다. 모든 위키피디아 문서는 1개 이상의 카테고리 정보를 가지고 있기 때문에 이 연구에서는 동일하게 출현한 카테고리

수에 의해서 유사도를 측정하였다. 두 용어에 매핑된 위키피디아 문서 사이에 공통되는 카테고리 정보가 많다는 것은 용어가 주제적으로 관련되어 있다고 할 수 있기 때문이다.

(3) 링크 정보의 활용

각각의 용어와 매핑된 위키피디아 문서의 링크 정보를 활용하여 두 용어 간의 유사도를 측정한다. 위키피디아에 나타난 링크 정보는 해당문서와 관련된 개념을 표현하는 위키피디아 문서로 링크되는 것이거나 동일한 주제 또는 참고할 만한 기타 웹페이지로의 링크이기 때문에 공통적인 링크가 많다는 것은 두 문서가 주제적으로 관련되어 있다고 할 수 있다.

3.2.4 용어 간 유사도 측정

비학습용어와 학습용어 간의 유사도 측정을 하기 위해 사용된 유사도 척도는 단순 공기빈도와 정규화된 공기빈도이다. 위키피디아 문서의 경우에는 제목과 본문, 카테고리, 링크가 간략하게 구성되는 것이 아니라 최대한 다양한 관점으로 서술되어 있다. 단순 공기빈도를 사용하는 경우에 본문, 카테고리, 링크와 같은 특성이 길거나 많은 위키피디아 문서일수록 용어 비교에 있어서 상대적으로 높은 유사도 값을 나타낼 수 있다. 반면에 정규화된 유사계수를 사용하게 되면 용어 간 비교 시 단순 공기빈도를 사용하는 것과 반대로 상대적으로 낮은 유사도 값이 나타날 수 있다. 따라서 본 실험에서는 단순 공기빈도와 정규화된 공기빈도를 비교하였다. 정규화된 공기빈도를 나타내기 위해 아래와 같은 다이스 계수를 사용하였다.

$$S(X,Y)=\frac{2a}{2a+b+c}$$

a : 용어 X와 용어 Y로 검색된 위키피디아 문서 사이에 동시 출현한 특성(본문, 카테고리, 링크) 수

b : 용어 X로 검색된 위키피디아 문서에 출현한 특성 수

c : 용어 Y로 검색된 위키피디아 문서에 출현한 특성 수

이와 같이 단순 공기빈도(다이스 계수 공식에서 a값)와 다이스 계수를 이용하여 각각의 비학습용어와 학습용어들 사이의 유사도를 산출하고, 그 중 가장 큰 유사도 값을 가지는 학습용어로 비학습용어를 대체하였다.

비학습용어와 학습용어 간의 유사도 값에 상관없이, 가장 유사도가 높은 한 개의 학습용어로 비학습용어를 대체한 경우, 너무 낮은 유사도 값을 가지는 학습용어가 비학습용어를 대체하게 되면 오히려 성능이 더 떨어지는 결과를 나타낼 수 있기 때문에, 임계치를 두어 유사도 값을 제한할 필요가 있다. 즉 최고 유사도 값을 가지는 학습용어라도 임계치를 넘는 경우에만 비학습용어를 학습용어로 대체하여 자질로 선정할 필요가 있다. 실험에서 최고 유사도 값을 가지는 용어로 대체하는 실험과 임계치를 설정한 실험을 진행하였다. 임계치 설정에 있어서, 공기빈도의 경우에는 대체되는 학습용어들의 공기빈도 값을 순서대로 나열한 뒤, 약 10% 간격으로 유사도 임계치를 설정하였다. 제목과 본문을 사용한 경우의 임계치는 0, 13, 21, 32, 44, 61, 87, 126, 194, 421과 같이 설정하였고 카테고리 정보의 경우 0부터 9까지 모두 10개의

임계치를, 링크 정보의 경우에는 0, 2, 3, 4, 6, 8, 12, 18, 34, 110의 임계치를 설정하였다. 다이스 계수의 경우에는 모든 실험에서 동일하게 0.1 단위로 임계치를 설정하였다.

이러한 방식으로 대체되는 학습용어는 입력문헌에 없는 학습용어가 될 수도 있고, 입력 문헌에 이미 출현하고 있는 학습용어가 될 수 있다. 입력문헌에 없는 학습용어로 대체된 경우에는 해당 비학습용어만 학습용어로 대체하고 가중치 값은 그대로 적용한 반면, 입력 문헌에 이미 출현하고 있는 학습용어로 대체된다면, 이미 출현하고 있는 용어가 가지고 있는 가중치에 대체된 학습용어의 가중치를 더하여 사용하였다.

3.2.5 문헌 간 유사도 측정

범주별로 비학습용어를 추출하게 되면 10가지 범주에 대해서 각각 서로 다른 비학습용어가 학습용어로 대체된다. 이렇게 변경된 입력문헌은 하나의 문헌이지만 10개의 범주에 대해서 서로 다른 벡터 형태를 가지게 된다. 따라서 kNN 분류기에서 문헌 간 유사도 측정 시, 어떠한 방식으로 유사도를 측정해야 하는지가 문제가 된다. 실험에서는 유사도 측정 대상이 되는 학습문헌의 범주에 해당하는 입력문헌 벡터를 사용하여 유사도를 산출하였다. 즉, 하나의 입력문헌과 10개 범주의 학습 문헌들 사이의 유사도를 측정할 때, 각각 범주에 따라 서로 다른 입력문헌의 벡터를 이용하여 유사도를 산출하게 된다.

3.3 실험 결과 및 분석

3.3.1 베이스라인 실험 결과

자질선정 실험 결과에 대한 비교의 기준이 되는 베이스라인 실험의 경우, 용어의 대체 없이 학습문헌 집합과 입력문헌 집합에서 추출한 자질만을 사용하여 자동 분류를 실행하였다. 〈표 2〉는 베이스라인 실험 결과로 10개의 카테고리의 평균 F_1값은 0.8039로 나타났다.

그리고 실험문헌 집단의 모든 용어들을 위키피디아 문서에 매핑한 후, 문헌 내에서 동일한

〈표 2〉 베이스라인 실험 결과

범주	정확률	재현율	F_1 값
acq	0.9583	0.8187	0.8831
coffee	0.9545	1.0000	**0.9767**
crude	0.8333	0.8163	0.8247
earn	0.9103	0.9797	0.9438
interest	0.7600	0.7170	0.7379
money-fx	0.7966	0.6912	0.7402
money-supply	0.4242	0.8750	**0.5714**
ship	0.6774	0.6000	0.6364
sugar	0.9565	0.9167	0.9362
trade	0.7317	0.8571	0.7895
매크로 평균	0.8003	0.8272	**0.8039**

위키피디아 문서로 매핑된 용어들을 단일 용어로 대체하여, 수정된 학습문헌 집합과 입력문헌을 구성하였다. 〈표 3〉은 동의어들을 단일 용어로 대체한 후 분류 실험을 실시한 결과이다.

실험 결과 베이스라인 대비 F_1값 향상률은 약 0.76%로 성능이 향상된 것을 확인할 수 있다. 단순히 문헌 내의 동의어들을 단일 용어로 대체하여도 성능 향상이 이루어지기 때문에 동의어를 단일 용어로 대체한 후 자질선정 실험을 실시하는 것이 더 효과적이라 할 수 있다. 그러므로 이후 모든 실험에서는 동의어를 단일 용어로 대체하여 수정된 학습문헌 집합과 입력문헌을 구성한 후 자질 선정 실험을 진행하였다.

3.3.2 전체 비학습용어 추출 후 분류 실험 결과

(1) 제목과 본문 정보를 이용한 결과

〈표 4〉는 위키피디아의 제목과 본문을 이용하여 비학습용어와 학습용어 간의 유사도를 측정한 다음 비학습용어를 가장 유사도가 높은 하나의 학습용어로 대체한 실험과 유사도 임계치를 적용한 실험 결과를 나타낸 것이다.

결과를 살펴보면 임계치를 설정하지 않은 실험에서는 공기빈도를 사용하는 경우 베이스라인에 비해 약 1.18%, 다이스 계수를 이용한 경

〈표 3〉 동의어 대체 후 분류 실험 결과

범주	정확률	재현율	F_1 값	베이스라인 대비 F_1 향상률(%)
acq	0.9581	0.8139	0.8801	-0.34
coffee	0.9130	1.0000	0.9545	-2.27
crude	0.8511	0.8163	0.8333	1.05
earn	0.9074	0.9826	0.9435	-0.03
interest	0.7660	0.6792	0.7200	-2.43
money-fx	0.7742	0.7059	0.7385	-0.23
money-supply	0.4375	0.8750	0.5833	2.09
ship	0.7419	0.6571	0.6970	9.52
sugar	0.9565	0.9167	0.9362	0.00
trade	0.7625	0.8714	0.8133	3.02
매크로 평균	0.8068	0.8318	0.8100	**0.76**

〈표 4〉 위키피디아의 제목과 본문 정보를 이용한 실험 결과

	정확률	재현율	F_1 값	베이스라인 대비 F_1 향상률(%)
공기빈도	0.8164	0.8434	0.8134	**1.18**
다이스계수	0.8204	0.8399	0.8127	**1.09**
공기빈도 (임계치: 21.32)	0.8207	0.8437	0.8159	**1.49**
다이스계수 (임계치: 0.2)	0.8237	0.8428	0.8157	**1.46**

우에는 약 1.09% 성능 향상이 이루어졌다.

임계치를 설정한 실험에서는 공기빈도의 경우, 임계치를 21과 32로 설정하였을 때 약 1.49%로 가장 높은 성능 향상률을 보였다. 다이스 계수의 경우 임계치가 0.2인 경우 약 1.46%의 향상률을 보이는 것으로 나타났다.

향상률로 비교하였을 때, 단순 공기빈도를 사용한 경우가 다이스 계수를 사용한 경우보다 약간 높은 성능 향상률을 나타내고 있다. 따라서 제목과 본문 정보 사용 시에는 단순 공기빈도를 유사도 척도로 사용하는 것이 더 효과적이라고 볼 수 있다.

(2) 카테고리 정보를 이용한 결과

〈표 5〉는 위키피디아의 카테고리 정보를 이용하여 비학습용어를 학습용어로 대체한 실험 결과를 보여 준다.

임계치를 설정하지 않은 실험의 경우에 향상률을 살펴보면 공기빈도의 경우에는 베이스라인 대비 약 0.21%, 다이스 계수의 경우 약 0.75% 성능이 향상되는 것으로 나타났다.

임계치를 설정한 실험에서는 공기빈도의 경우, 임계치를 2로 설정하였을 때 약 1.34%의 성능 향상률을 보였다. 다이스 계수의 경우에는 임계치를 0.7 이상으로 설정하였을 때 1.35%의 향상률을 보였다. 결과적으로 위키피디아의 카테고리 정보를 이용하였을 때에는 다이스 계수를 적용하는 것이 더 높은 성능 향상률을 보였다.

(3) 링크 정보를 이용한 결과

〈표 6〉은 위키피디아의 링크 정보를 이용한 실험 결과이다. 실험 결과를 살펴보면, 임계치

〈표 5〉 위키피디아의 카테고리 정보를 이용한 실험 결과

	정확률	재현율	F_1 값	베이스라인 대비 F_1 향상률(%)
공기빈도	0.8066	0.8392	0.8056	**0.21**
다이스계수	0.8151	0.8406	0.8099	**0.75**
공기빈도 (임계치: 2)	0.8207	0.8433	0.8147	**1.34**
다이스계수 (임계치: 0.7)	0.8219	0.8434	0.8147	**1.35**

〈표 6〉 위키피디아의 링크 정보를 이용한 실험 결과

	정확률	재현율	F_1 값	베이스라인 대비 F_1 향상률(%)
공기빈도	0.8137	0.8414	0.8101	**0.77**
다이스계수	0.8199	0.8399	0.8120	**1.00**
공기빈도 (임계치: 110)	0.8217	0.8423	0.8140	**1.26**
다이스계수 (임계치: 0.3)	0.8219	0.8434	0.8147	**1.35**

를 설정하지 않은 실험의 경우, 공기빈도를 사용한 실험과 다이스계수를 사용하는 실험이 각각 베이스라인 대비 F_1값 향상률이 약 0.77%과 1.00%로 나타나 있다.

공기빈도와 다이스 계수 임계치를 적용한 실험 결과를 살펴보면, 비학습용어를 링크 공기빈도가 110 이상인 학습용어와 대체시켰을 때 베이스라인 대비하여 약 1.26% 성능 향상이 이루어졌고, 다이스 계수는 임계치가 0.3일 때 약 1.35% 향상된 분류 성능을 가져 왔다.

유사도 척도로 비교해 봤을 때, 최고 유사도를 갖는 학습용어로 대체한 실험과 유사도 임계치를 적용한 실험 모두 다이스 계수를 사용하는 것이 높은 성능을 나타냈다.

3.3.3 범주별 비학습용어 추출 후 분류 실험 결과

(1) 제목과 본문 정보를 이용한 결과

범주별로 비학습용어를 추출하고, 위키피디아의 제목과 본문을 이용하여 유사도를 측정한 후 용어를 대체하여 분류 실험을 실시한 결과가 〈표 7〉에 나와 있다.

〈표 7〉을 살펴보면 공기빈도의 경우 베이스라인보다 약 9.62% 성능이 떨어지는 것으로 나타났다. 다이스 계수를 사용한 실험에서도 공기빈도와 마찬가지로 베이스라인 성능보다 약 10.91% 하락하였다. 정확률에 비해 재현율이 크게 낮아진 것이 F_1 값이 낮아진 원인으로 분석된다.

공기빈도의 경우, 임계치가 421일 때 베이스라인보다 약 0.64% 향상되는 것을 알 수 있다. 다이스 계수의 경우, 임계치가 0.7 이상인 경우 약 0.80%의 향상률을 나타냈다. 전체 비학습용어 추출 후 진행한 실험과는 다르게 유사도 임계치가 비교적 높을 때 좋은 성능이 나타나게 된다.

유사도 임계치를 적용하지 않았을 경우 공기빈도를 사용하는 것이 더 높은 성능을 보이고 있으나, 유사도 임계치를 적용하였을 때에는 다이스 계수를 사용하는 것이 더 높은 성능을 나타내고 있다.

(2) 카테고리 정보를 이용한 결과

〈표 8〉은 위키피디아의 자질 중에서 카테고리 정보를 이용하여 실험한 결과이다.

카테고리 정보를 사용하였을 경우 제목과 본문을 사용하였을 경우와 마찬가지로 성능이 하락하여, 공기빈도의 경우 3.01%, 다이스 계수

〈표 7〉 위키피디아의 제목과 본문 정보를 이용한 실험 결과

	정확률	재현율	F_1 값	베이스라인 대비 F_1 향상률(%)
공기빈도	0.8028	0.7113	0.7266	**-9.62**
다이스계수	0.8011	0.6963	0.7162	**-10.91**
공기빈도 (임계치: 421)	0.8387	0.8157	0.8090	**0.64**
다이스계수 (임계치: 0.7)	0.8409	0.8160	0.8104	**0.80**

〈표 8〉 위키피디아의 카테고리 정보를 이용한 실험 결과

	정확률	재현율	F_1 값	베이스라인 대비 F_1 향상률(%)
공기빈도	0.8326	0.7648	0.7797	**-3.01**
다이스계수	0.8346	0.7833	0.7890	**-1.85**
공기빈도 (임계치: 9)	0.8453	0.8280	0.8188	**1.85**
다이스계수 (임계치: 0.7)	0.8383	0.8115	0.8067	**0.35**

의 경우 1.85% 성능이 낮아졌다.

실험 결과를 살펴보면 공기빈도 임계치가 9일 경우 가장 높은 성능을 나타냈는데 베이스라인에 비해 약 1.85%의 성능 향상률을 보였다. 전반적으로 임계치가 높아질수록 전체 성능이 향상되는 것으로 나타났고, 정확률보다는 재현율이 더 크게 상승하는 것을 확인할 수 있다.

다이스 계수의 경우는 임계치가 0.7 이상일 때 약 0.35%의 성능 향상을 보이나 다른 실험에 비해 상대적으로 낮은 성능 향상이 이루어지는 것을 확인할 수 있다.

유사도 척도를 비교하였을 때, 최고 유사도를 갖는 학습용어로 대체한 결과에서는 다이스 계수를 적용하는 것이 좋은 성능을 보이고 있으나, 유사도 임계치를 적용하였을 때에는 단순 공기빈도를 사용하는 것이 더 높은 성능을 나타내고 있다.

(3) 링크 정보를 이용한 결과

〈표 9〉는 위키피디아의 자질 중에서 링크 정보를 이용하여 실험한 결과이다.

실험 결과를 살펴보면 공기빈도의 경우 베이스라인 대비 F_1 값이 약 13.34% 하락하였고, 다이스 계수의 경우에도 베이스라인보다 약 8.26% 성능이 하락한 것으로 나타났다.

공기빈도의 경우, 비학습용어를 링크 공기빈도가 110 이상인 학습용어로 대체시켰을 때 베이스라인에 비해 약 0.91% 성능 향상이 이루어졌다. 다이스 계수의 경우 임계치가 0.3일 때, 약 0.80% 향상된 분류 성능을 가져왔다. 유사도 척도로 비교해 봤을 때, 최고 유사도를 갖는 학습용어로 대체한 실험과 유사도 임계치를 적

〈표 9〉 위키피디아의 링크 정보를 이용한 실험 결과

	정확률	재현율	F_1 값	베이스라인 대비 F_1 향상률(%)
공기빈도	0.7922	0.6799	0.6967	**-13.34**
다이스계수	0.8070	0.7161	0.7375	**-8.26**
공기빈도 (임계치: 110)	0.8408	0.8170	0.8112	**0.91**
다이스계수 (임계치: 0.3)	0.8409	0.8160	0.8104	**0.80**

용한 실험 모두 다이스 계수를 사용하는 것이 더 높은 성능을 나타냈다. 링크 정보를 사용할 경우 다이스 계수가 더 효과적으로 성능을 향상시킬 수 있는 것으로 볼 수 있다.

3.3.4 종합 평가

본 연구에서는 하나의 실험문헌에서 출현한 복수의 용어들이 동일한 위키피디아 문서로 매핑될 경우 이 용어들을 동의어로 보고 단일 용어로 표현하였다. 동의어들을 단일 용어로 대체한 후 분류 실험을 실행한 결과 약 0.76%의 성능 향상이 이루어졌다.

분류자질로 선정하는 실험에서는 (1) 비학습용어 추출 시 범주정보의 사용여부, (2) 용어의 유사도 측정 방법(위키피디아 문서의 제목과 본문, 카테고리 정보, 링크 정보), (3) 유사도 척도(단순 공기빈도, 정규화된 공기빈도) 등 세 가지 조건을 결합하여 실험을 수행하였다. 〈표 10〉과 〈표 11〉은 전체 실험 결과를 요약한 것이다.

〈표 10〉을 보면 범주 정보를 사용하지 않은 것이 베이스라인보다 좋은 성능을 나타내고 범주 정보를 사용한 경우에는 오히려 성능이 떨어지는 것을 확인할 수 있다. 범주 정보를 사용하

〈표 10〉 최고 유사도를 갖는 학습용어로 대체한 실험 결과

	범주 정보 미사용		범주 정보 사용	
	F_1 값	성능 향상률(%)	F_1 값	성능 향상률(%)
베이스라인	0.8039	(비교 대상)		
동의어 결합	0.8100	0.76		
제목・본문+공기빈도	0.8134	1.18	0.7266	-9.62
제목・본문+다이스계수	0.8127	1.09	0.7162	-10.91
카테고리+공기빈도	0.8056	0.21	0.7797	-3.01
카테고리+다이스계수	0.8099	0.75	0.7890	-1.85
링크+공기빈도	0.8101	0.77	0.6967	-13.34
링크+다이스계수	0.8120	1.00	0.7375	-8.26

〈표 11〉 유사도 임계치를 적용한 실험 결과

	범주 정보 미사용			범주 정보 사용		
	임계치(분포)	F_1 값	성능 향상률(%)	임계치(분포)	F_1 값	성능 향상률(%)
베이스라인		0.8039	(비교 대상)			
동의어 결합		0.8100	0.76			
제목・본문+공기빈도	21 이상(20%)	0.8159	**1.49**	421 이상(90%)	0.8090	0.64
제목・본문+다이스계수	0.2 이상	0.8157	1.46	0.7 이상	0.8104	0.80
카테고리+공기빈도	2 이상(20%)	0.8147	1.34	9 이상(90%)	0.8188	**1.85**
카테고리+다이스계수	0.7 이상	0.8147	1.35	0.7 이상	0.8067	0.35
링크+공기빈도	110 이상(90%)	0.8140	1.26	110 이상(90%)	0.8112	0.91
링크+다이스계수	0.3 이상	0.8147	1.35	0.3 이상	0.8104	0.80

는 경우에는, 비학습용어와 관련이 없는 범주에서도 그 범주에 해당하는 학습용어와 대체되기 때문에 입력문헌과 주제적으로 관련 없는 학습문헌이 더 높은 유사도를 가질 수 있게 된다. 따라서 범주 정보를 사용한 경우 베이스라인보다 성능이 떨어지는 것을 확인할 수 있는 것이다.

위키피디아의 특성 중에서 범주 정보를 사용하지 않는 경우, 즉 전체 비학습용어를 이용한 경우에는 위키피디아의 제목과 본문을 사용한 실험이 1.18%의 가장 높은 성능 향상률을 나타냈고, 범주 정보를 사용한 경우에는 카테고리 정보를 사용한 실험이 가장 낮은 성능 하락을 보였다.

공기빈도와 다이스 계수와의 성능 차이를 살펴본 결과, 범주 정보를 사용하지 않는 경우에는, 위키피디아의 본문과 제목, 링크 정보를 이용할 때 공기빈도를 사용하는 것이 더 높은 성능을 나타냈다. 그리고 범주 정보를 사용하는 경우에는 위키피디아의 제목과 본문을 사용하는 경우를 제외하고는 다이스 계수를 사용할 때 더 좋은 성능을 보였다.

〈표 11〉을 보면 모든 실험에서 베이스라인보다 높은 성능을 나타냈다. 전반적으로 범주 정보를 사용하지 않고 전체 비학습용어를 추출한 실험의 성능이 더 높게 나타나는 것을 볼 수 있다. 그러나 가장 높은 성능 향상률은 범주 정보를 사용하였을 때 나타났다. 용어 간 유사도 측정과정에서 위키피디아의 카테고리 정보를 이용하고, 유사계수로는 공기빈도, 임계치를 9로 설정한 실험이 1.85%의 가장 높은 성능향상률을 나타냈다.

실험 결과 대부분이 베이스라인보다 높은 성능을 나타냈기 때문에 BOW 방식이 안고 있는 문제점의 해결방법으로 제시한 (1) 한 문헌에 출현한 동의어를 하나의 용어로 변경하고, (2) 비학습용어를 학습용어로 변경하는 방법이 효과적이라고 할 수 있다.

이 연구에서 실험 결과 성능이 향상된 원인으로는 다음의 사항들을 살펴볼 수 있다.

우선 위키피디아를 통해 철자 오류가 수정된다는 것이다. 위키피디아의 추천어 제시 기능을 통해 가장 유사한 용어로 변환되기 때문에 자동적으로 오류가 수정되어 해당 용어가 범주화 자질로서 선정될 수 있다.

둘째로, 동의어들이 단일 용어로 통일된 것을 들 수 있다. 실험에서 모든 용어를 위키피디아 문서로 매핑하였기 때문에 약어와 같은 동의어들이 분류자질로서 더욱 적합한 단일 용어로 변경되었다. 이로 인해 문헌 간 유사도 측정 시 관련 문헌들의 유사도 값이 향상되는 결과를 가져왔다.

셋째로, 의미적으로 서로 관련된 용어로의 변경이 이루어졌다는 점이다. 실험에서 'decree'가 'law'로 변경되는 것처럼 비학습용어가 의미적으로 유사한 학습용어로 대체된다. 그리고 임계치를 설정할 경우 성능이 더욱 향상되는 것은 용어의 대체가 더 정확하게 이루어지기 때문인 것으로 볼 수 있다.

4. 결 론

본 연구에서는 텍스트 범주화 실험에서 일반적으로 발생하는 분류자질 선정과 관련된 문제점을 해결하기 위하여 외부 정보자원인 위키피디아를 활용하였다. 자질 선정 및 범주화 실험

에서 분류대상 문헌에 출현한 동의어들을 단일 자질로 표현함으로써 범주화 성능이 약 0.7% 향상되었고, 또한 학습문헌 집합에 출현하지 않은 용어들을 가장 유사한 학습문헌의 용어들로 대체함으로써 최대 1.85%의 성능 향상을 가져왔다.

실험 결과 문헌 색인에 일반적으로 사용되는 BOW 방식의 한계점을 극복하는 방안으로 이 연구에서 제시한 동의어 결합 방법과 비학습용어를 학습용어로 대체하여 분류자질로 사용하는 방법은 모두 범주화 성능 향상에 효과가 있는 것으로 분석되었다. 또한 단순히 가장 유사도 값이 큰 용어가 아니라 임계치를 설정하여 일정한 값 이상의 유사도를 갖는 용어만을 대체 용어로 선정하는 경우 더욱 성능을 높일 수 있었다.

본 연구에서는 위키피디아의 제목과 본문, 카테고리 정보, 링크 정보만을 활용하여 분류자질을 선정하였지만 위키피디아는 내부적으로 더 많은 장치들을 가지고 있다. 위키피디아의 네트워크 구조나 계층 구조 등을 온톨로지로 활용한다면 범주화 성능을 더욱 향상시킬 수 있을 것으로 기대된다.

참 고 문 헌

Bird, S., Klein, E., & Loper, E. (2007). Natural language processing in Python. O'ReillyMedia.

Gabrilovich, E., & Markovitch, S. (2005). Feature generation for text categorization using world knowledge. Proceedings of the 19th international Joint Conference on Artificial intelligence, 1048-1053.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. Proceedings of the 21st National Conference on Artificial Intelligence, 1301-1306.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proceedings of the 20th International Joint Conference on Artificial Intelligence, 1606-1611. Retrieved from http://ijcai.org/papers07/Papers/IJCAI07-259.pdf

Huang, A., Milne, D., Frank, E., & Witten, I. H. (2009). Clustering documents using a Wikipedia-based concept representation. Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (LNCS 5476/2009), 628-636. http://dx.doi.org/10.1007/978-3-642-01307-2_62

Milne, D., Witten, I. H., & Nichols, D. M. (2007). A knowledge-based search engine powered by Wikipedia. Proceedings of the 16th ACM Conference on Information and Knowledge

Management, 445-454. http://dx.doi.org/10.1145/1321440.1321504

Milne, D., & Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Proceedings of the First AAAI Workshop on Wikipedia and Artificial Intelligence, (WIKIAI 2008). Rertrieved from http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-005.pdf

Minier, Z., Bodo, Z., & Csato, L. (2007). Wikipedia-based kernels for text categorization. Proceedings of the International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 157-164. http://dx.doi.org/10.1109/SYNASC.2007.8

Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 192-199. http://dx.doi.org/10.3115/1220835.1220860

Ponzetto, S. P., & Strube, M. (2007). Knowledge derived from Wikipedia for computing semantic relatedness. Journal of Artificial Intelligence Research, 30(1), 181-212. http://dx.doi.org/10.1613/jair.2308

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. Proceedings of the 21st National Conference on Artificial Intelligence, 1419-1424.

Wang, P., Hu, J., Zeng, H., Chen, L., & Chen, Z. (2007). Improving text classification by using encyclopedia knowledge. Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, 332-341. http://dx.doi.org/10.1109/ICDM.2007.77

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. http://dx.doi.org/10.1145/1401890.1401976