특허 정보 검색을 위한 대체어 후보 추출 방법
(Extracting Alternative Word Candidates for Patent Information Search)

백 종 범 † 김 성 민 †
(Jongbum Baik) (Seongmin Kim)

이 수 원 ††
(Soowon Lee)

요 약 특허 정보 검색은 연구 및 기술 개발에 앞서 선행연구의 존재 여부를 확인하기 위한 사전 조사 목적으로 주로 사용된다. 이러한 특허 정보 검색에서 원하는 정보를 얻지 못하는 원인은 다양하다. 그 중에서 본 연구는 키워드 불일치에 의한 정보 누락을 최소화하기 위한 대체어 후보 추출 방법을 제안한다. 본 연구에서 제안하는 대체어 후보 추출 방법은 문장 내에서 함께 쓰이는 단어들이 비슷한 두 단어는 서로 비슷한 의미를 지닐 것이다라는 직관적 가설을 전제로 한다. 이와 같은 가설을 만족하는 대체어를 추출하기 위해서 본 연구에서는 분류별 집중도, 신뢰도를 이용한 연관단어뭉치, 연관단어 뭉치간 코사인 유사도 및 순위 보정 기법을 제안한다. 본 연구에서 제안한 대체어 후보 추출 방법의 성능은 대체어 유형별로 작성된 평가지표를 이용하여 재현율을 측정함으로써 평가하였으며, 제안 방법이 문서 벡터공간 모델의 성능보다 더 우수한 것으로 나타났다.

키워드 : 연관단어, 대체어, 유의어, 특허정보검색

Abstract Patent information search is used for checking existence of earlier works. In patent information search, there are many reasons that fails to get appropriate information. This research proposes a method extracting alternative word candidates in order to minimize search failure due to keyword mismatch. Assuming that two words have similar meaning if they have similar co-occurrence words, the proposed method uses the concept of concentration, association word set, cosine similarity between association word sets and a ranking modification technique. Performance of the proposed method is evaluated using a manually extracted alternative word candidate list. Evaluation results show that the proposed method outperforms the document vector space model in recall.

Key words : AssociationWord, AlternativeWord, Similar Word, Patent Information Search

1. 서 론

일반적으로 특허란, "지금까지 없었던 신규하고 진보한 발명에 부여하는 것"으로 정의된다[1,2]. 즉, 특허를 출원하고자 하는 기술이 이미 존재하는 기술과 동일하거나 유사한 경우에는 등록이 불가능하다는 의미이며, 이러한 유사 기술의 존재 여부를 확인하기 위해서 특허 정보 검색은 필수적인 절차이다. 특허 정보 검색에서는 특허문헌을 효율적으로 관리하기 위해서 IPC 분류(International Patent Classification)를 이용한다. 비록 IPC 분류에 의해 검색해야할 특허 문헌의 개수가 많이 줄어든다 할지라도 여전히 방대한 양의 특허 문헌을 검색해야 하므로 IPC 분류가 기존의 특허 정보 검색에 존재하는 문제의 범위를 줄여줄 수는 있지만 검색의 근본적 문제를 해결해 주지는 않는다.

일반적으로 검색을 힘들게 하는 요소는 크게 단어의 다의성과 표기의 다양성으로 나눌 수 있다. 단어의 다의성이란 하나의 단어가 다양한 의미로 쓰이는 것을 의미하며, 표기의 다양성이란 다르게 표기된 단어가 같은 의미로 쓰이는 것을 의미한다. 다행히도 특허 정보 도메인은 IPC 분류에 의해 단어의 다의성은 어느 정도 해결이 되어있는 도메인이라고 할 수 있다. 그러므로 본 연구에서는 표기의 다양성으로 인한 검색의 어려움을 해결하는데 중점을 두고자 한다.

본 논문은 위와 같은 표기의 다양성으로 인한 키워드 불일치에 의한 정보 누락을 최소화하기 위하여 특허 문헌 뭉치에서 대체어 후보를 추출하는 방법을 제안한다. 본 연구에서 정의하는 대체어란, "한 문장에서 특정 단어를 대신하여 사용해도 문장의 의미를 훼손하지 않는 단어"를 의미하며, 특허 문헌 데이터의 특성을 고려하여 대체어를 표 1과 같이 4가지 경우로 분류하여 사용한다. 특히, 본 연구에서는 표 1의 분류 중 타 분류에 비해 사용자들이 예측하기 힘든 철자변형을 찾아내는데 중점을 둔다.

· 본 연구는 숭실대학교 교내연구비 지원으로 수행되었다.
· 이 논문은 제35회 추계학술대회에서 '특허 정보 검색을 위한 대체어 후보 추출 방법 연구'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 숭실대학교 컴퓨터학과
jbb100@mining.ssu.ac.kr
mabak@mining.ssu.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학과 교수
swlee@ssu.ac.kr

논문접수 : 2009년 1월 19일
심사완료 : 2009년 2월 24일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제15권 제4호(2009.4)

표 1 대체어 분류의 정의

분류	정 의
철자변형	중심어와 동일한 대상을 다른 철자로 표기한 경우
영어	한글로 표기된 중심어에 대한 영어 표기
한글	영어로 표기된 중심어에 대한 한글 표기
유의어	중심어와 비슷한 의미를 지닌 단어

기존 연구 중 유의어 또는 연관단어를 찾는 연구는 대부분 "의미가 비슷한 단어들은 같은 문맥(context)에서 사용될 것이다"라는 가설을 전제로 하고 있다[3-5]. 이러한 연구들은 문맥을 문서 혹은 문장으로 정의하고, 특정 단어가 특정 문서에서 출현했는지 여부를 논리값 혹은 출현빈도로 기술한 문서벡터공간 모델을 이용하여 단어의 유사도를 계산한다.

그러나 기존 연구들을 특허 문헌에 적용할 경우, 본 연구에서 목표로 하는 표기의 다양성을 해결할 수 없다. 이는 문서벡터공간 모델을 이용한 유의어 발견 연구의 가장 큰 단점으로 같은 문맥에서 출현하지 않은 단어들에 대해서는 유사도 검사를 수행하지 않으므로 인해 상당 수 유사 단어들(특히 철자변형)을 발견하지 못하는 문제를 지니기 때문이다.

본 논문에서는 이와 같은 문제를 해결하기 위하여 문서벡터 모델이 아닌 확률에 기반을 두는 연관 규칙을 이용하여 대체어 후보를 추출하는 방법을 제안한다. 제안하는 방법은 먼저 집중도를 계산하여 IPC별 중요 단어를 선정하고 중요 단어로 선정된 단어들을 이용하여 연관단어 뭉치를 생성한 다음, 생성된 연관단어 뭉치의 유사도를 계산하여 대체어 후보 목록을 생성하고 마지막으로 대체어 순위를 보정하는 단계를 거친다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 본 논문에서 제안하는 대체어 후보 추출 알고리즘을 소개하며, 4장에서는 3장에서 제안하는 알고리즘의 성능을 평가한다. 마지막으로 5장에서는 결론 및 향후연구를 제시한다.

2. 관련 연구

자동으로 유의어를 찾는 시도는 크게 두 가지 방법으로 분류할 수 있다. 첫 번째 방법은 사전(dictionary)을 이용하여 유의어를 찾는 방법이고, 두 번째 방법은 문서 뭉치 속에서 유의어를 찾는 방법이다.

2.1 사전을 이용한 유의어 발견 연구

사전을 이용한 유의어 발견 연구[3,6,7]는 "유의어들은 단어의 정의에 있어서 많은 공통점을 지닐 것이고, 많은 공통적 단어들의 정의에 사용될 것이다"라는 가정하에 유의어를 탐색한다.

사전을 이용하여 유의어를 찾기 위해서는 먼저 사전 그래프(dictionary graph) G를 생성한다. 사전 그래프 G의 각 정점(vertex)은 단어로 구성되며 간선(edge) (u,v)은 임의의 정점 v가 다른 임의의 정점 u의 정의에 나타나는 경우에 연결된다.

사전 그래프 G가 생성되면 주어진 질의 단어 w에 대한 이웃 그래프(neighborhood graph) G_w를 생성한다. G_w는 G의 하위 그래프이며 정점 w가 참조하거나 참조되는 정점과 간선으로 구성된다. 즉, G_w는 단어 w와 연결된 점과 간선을 모두 포함하는 그래프이며 질의 단어 w에 대한 유의어 후보 목록이라고 할 수 있다.

질의 단어 w에 대한 이웃 그래프 G_w가 생성되면 최종적으로 유의어를 선택하게 된다. 유의어 선정 방법으로는 거리 기반 방법(distance method), ArcRank 등이 존재한다[3].

2.2 대량 문서 뭉치를 이용한 유의어 발견 연구

문서 뭉치에서 유의어를 발견하는 연구[3-5]는 "의미가 비슷한 단어들은 같은 문맥에서 사용 될 것이다"라는 가설에 기반하며, 접근 방법의 차이는 문맥을 무엇으로 정의할 것인지에 따라 달라진다. 문맥은 문서, 제목, 초록 등 다양한 기준으로 정의될 수 있다.

유의어 발견 연구에서 가장 많이 사용하는 방법은 문서벡터공간 모델을 이용하는 방법이다. 일반적으로 문서벡터공간 모델은 "단어(word)×문서(document)"의 형태로 표현되며 행과 열의 교차점에 특정 문서에서 특정 단어가 출현하는지 여부를 표시하게 된다.

문서벡터공간 모델에서 열 정보(특정 문서 내 단어 출현 여부)를 이용하여 벡터를 구성하면 문서 간 유사도 비교가 가능하며, 반대로 행 정보(특정 단어가 문서에 출현했는지 여부)를 이용하여 벡터를 구성하면 단어 간 유사도 비교가 가능하다. 문서 간 혹은 단어 간 유사도를 구하기 위한 척도로는 코사인 유사도가 가장 많이 사용된다.

3. 대체어 후보 추출

본 논문에서 제안하는 대체어 후보 추출 방법은 "특정 단어 A와 함께 자주 쓰이는 '연관단어 뭉치 A'가 다른 단어 B와 함께 자주 쓰이는 '연관단어 뭉치 B'와 비슷할 경우, 이 두 단어는 대체어일 것이다"라는 직관적 가설을 전제로 한다. 즉, 특정 단어 A와 함께 자주 쓰이는 단어가 비슷한 단어 B는 대체어일 가능성이 높다는 것이다.

앞서 언급한 가설에 따라 본 연구에서 대체어를 추출하는 방법은 IPC별 중요 단어 선정 단계[1단계], 연관단어 뭉치를 생성하는 단계[2단계], 생성된 연관단어 뭉치의 유사도를 계산하여 대체어 후보 목록을 생성하는 단계[3단계], 대체어 순위를 보정하는 필터링 단계[4단계]로 구성된다(그림 1).

3.1 IPC별 중요 단어 선정

IPC 분류 내에서 단어별 대체어를 생성하기 위해서는

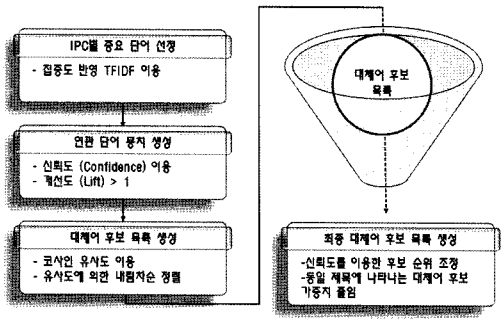

그림 1 대체어 후보 추출 시스템 구조도

해당 IPC 분류에서 중요한 단어가 무엇인지 선정하는 단계가 필요하다. 만약 이 단계를 생략한다면 알고리즘 수행 과정에 불용어가 포함되어 수행 시간이 많이 소요되는 것은 물론이고 최종 결과 역시 왜곡될 수 있다.

중요 단어를 선정하기 위해서 일반적인 TFIDF를 IPC 구분 없이 특허 문헌 뭉치에 적용할 경우 각 IPC 분류별 문서 비율에 따라 해당 분류에서 자주 출현하는 중요 단어가 낮은 IDF를 지님으로 인해 왜곡된 대표어 선정 결과를 초래할 수 있다. 이러한 문제를 해결하기 위하여 본 연구에서는 각 IPC 분류별 중요 단어를 추출하는 방법으로 [8]에서 사용한 분류별 집중도를 반영한 TFIDF를 변형하여 분류별 집중도(식 (1))를 정의하였다.

$$concentration_i(w) = \frac{(\frac{n_i}{n_i - df_i(w)})(1 - \frac{gf(w)}{N_G})}{\sqrt{\sum_{j \in G}((\frac{n_j}{n_j - df_j(w)})^2(1 - \frac{gf(w)}{N_G})^2)}}$$

n_i : 문서 분류 i에 속하는 문서의 수
$df_i(w)$: 문서 분류 i에서 단어w가 출현한 문서의 수
G : 전체 문서 분류 집합
N_G : 전체 문서 분류의 수
gf(w) : 단어 w가 출현한 문서 분류의 수

식 (1) 분류별 집중도

식 (1)에서 의미하는 $concentration_i(w)$는 분류 i에서 단어 w가 지니는 집중도로 정의된다. $n_i/(n_i-df_i(w))$는 문서 분류 i내에서 단어 w가 중요한 정도를 의미하며, $1-(gf(w)/N_G)$는 여러 문서 분류에서 출현한 단어 w의 가중치를 낮추어주는 역할을 한다.

표 2는 일반적인 TFIDF 및 분류별 집중도를 이용했을 경우의 중요 단어 선정 결과이다. 표 2(a)에 따르면 일반적인 TFIDF를 이용할 경우에는 희소한 단어(출현 빈도가 낮은 단어)가 상위에 위치한다. TFIDF가 모두 같은 이유는 H04L(디지틀 정보의 전송)분류에서 한번만 출현한 단어들이 가장 높은 TFIDF를 지니고 있기 때문이다. 반면에 표 2(b)에 따르면 H04L분류에서 중요하다고 판단되는 단어들이 상위에 위치하는 것으로 나타난다.

표 2 H04L에서의 중요 단어

순위	단어	TFIDF	단어	집중도
1	네트워크형	14.7651	패킷	0.3507
2	환경기반	14.7651	인터넷	0.3489
3	컴플리언트	14.7651	프로토콜	0.3480
4	비정규	14.7651	서비스	0.3446
5	유아이	14.7651	에이티엠	0.3407
...	...	...	...	...

(a) 일반적인 TFIDF 이용 (b) 분류별 집중도 이용

또한 표 3은 H04L분류에서 TFIDF가 가장 낮은 최하위 5단어를 추출한 것으로 표 2(b)에서 중요하다고 판단한 단어 대부분이 이에 속하는 것으로 나타났다.

표 3 TFIDF 최하위 5단어

단어	TFIDF
이용	2.8979
네트워크	2.8852
장치	1.7567
시스템	1.5987
방법	0.7570

3.2 연관단어 뭉치 추출

본 연구에서는 연관단어 뭉치를 추출하기 위해서 연관 규칙[8,9]을 이용한다. 본 연구에서 정의하는 연관단어 뭉치란 "연관규칙 'X→Y'가 있을 때 최소 신뢰도(minimum confidence) α와 최소 개선도(minimum lift) β를 만족하는 연관단어들의 집합"을 의미한다. 본 연구에서 연관단어 뭉치는 3.1절에서 선정한 분류별 중요 단어들만을 이용하여 생성한다. 최소 지지도를 선정하지 않은 이유는 이를 설정함으로 인해 희소하게 등장하는 철자변형 단어들이 다른 단어들과 유사도를 비교할 기회조차 갖지 못한 채 알고리즘 수행 과정에서 배제되는 문제를 완화하기 위해서이다.

표 4는 철자변형에 속하는 단어인 '콘텐츠', '콘텐트'의 연관단어 뭉치의 일부이다. 표 4를 보면 철자변형이라고 생각한 두 단어의 연관단어 뭉치 대부분이 일치하는 것으로 나타난다.

3.3 대체어 후보 목록 생성

대체어 후보 목록은 전 단계에서 생성된 연관단어 뭉치들 간에 코사인 유사도를 계산함으로써 생성한다. 즉, 각 단어별로 생성된 연관단어 뭉치가 특정 단어를 설명해주는 기술자라고 가정하고, 두 단어의 연관단어 뭉치를 모두 포함하는 하나의 벡터 공간(vector space)을 생성하여 코사인 유사도를 계산한 후 내림차 순 정렬을 함으로써 대체어 후보 목록을 생성한다.

표 4 '콘텐츠', '콘텐트'의 연관단어 뭉치 일치 정도

중심어	연관어	신뢰도	중심어	연관어	신뢰도
콘텐츠	프로그램	0.0789	콘텐트	[illegible]	0.1364
콘텐츠	서버	0.0789	콘텐트	서비스	0.0909
콘텐츠	멀티미디어	0.0702	콘텐트	라우팅	0.0909
콘텐츠	컴퓨터	0.0702	콘텐트	가입	0.0909
콘텐츠	[illegible]	0.0526	콘텐트	[illegible]	0.0909
콘텐츠	[illegible]	0.0526	콘텐트	[illegible]	0.0909
콘텐츠	[illegible]	0.0439	콘텐트	분배	0.0909
콘텐츠	[illegible]	0.0439	콘텐트	[illegible]	0.0909
...	...	...	...	...	...

3.4 대체어 순위 보정

대체어 후보 목록을 코사인 유사도에만 의존하여 내림차 순 정렬을 수행할 경우, 같은 문장에서 자주 같이 등장한 단어(연관단어)들이 대체어 후보 목록의 상위에 위치하는 문제가 종종 발생한다. 본 논문은 이런 현상을 개선하기 위하여 신뢰도를 이용한 순위 보정 방법을 제안한다.

신뢰도를 이용한 순위 보정 방법이란 두 단어 X, Y에 대한 연관규칙에서 Confidence(X, Y)와 Confidence(Y, X)의 평균을 취한 결과의 역을 이용하여 코사인 유사도 값에 곱해줌으로써 동시 출현할 확률이 낮은 대체어 후보의 순위를 상승시켜주는 방법이다(식 (2)).

$$simScore(X,Y) = sim(A_X,A_Y)\left(1-\frac{Confidence(X,Y)+Confidence(Y,X)}{2}\right)$$

식 (2) 순위 보정 식

표 5에서 왼쪽(a)은 3.3절에서 생성한 대체어 후보 목록이며 오른쪽(b)은 신뢰도를 이용하여 대체어 후보를 보정한 결과이다. 결과 중 평가지표(표 6)에 존재하는 단어로는 '컨텐츠', '콘텐트', '컨텐트' 등이 있으며, 표 5(a)를 식 (2)를 이용하여 순위 보정을 수행한 결과 표 5(b)와 같이 상위 10순위에서 벗어나 있던 '콘텐츠'의 대체어 '컨텐트'가 8순위로 상승하는 것을 확인할 수 있었다.

표 5 철자변형 단어 '콘텐츠'의 대체어 후보 목록

순위	중심어	대체어	유사도	중심어	대체어	보정값
1	콘텐츠	[illegible]	0.6726	콘텐츠	[illegible]	0.8363
2	콘텐츠	[illegible]	0.5780	콘텐츠	[illegible]	0.7890
3	콘텐츠	서버	0.5658	콘텐츠	애플리케이션	0.7772
...	...	...	...	...	...	...
7	콘텐츠	기억	0.5462	콘텐츠	파일	0.7609
8	콘텐츠	클라이언트	0.5451	콘텐츠	[illegible]	0.7600
...	...	...	...	...	...	...
13	콘텐츠	[illegible]	0.5200	콘텐츠	응용	0.7516
...	...	...	...	...	...	...

(a) 코사인 유사도에 의한 대체어 후보 목록 (b) 순위 보정 결과

4. 실험 및 평가

4.1 자료 수집

본 연구에서는 실험을 위하여 '네이버 특허 서비스'에 존재하는 특허 문헌들을 수집하였다. 수집된 특허 문헌은 총 172,458건으로 본 실험에서는 IPC 분류 중 기술 용어의 의미를 판별하기 쉬운 '디지틀정보의 전송(H04L)' 분류에 속한 27,845건의 특허 문헌에 대하여 실험을 수행하였다.

또한 본 실험의 대체어 추출 성능 평가를 위한 최대한 객관적인 평가지표 생성을 위해서, 3.1절에서 선정한 중요 단어를 '엠파스 IT 용어 사전'에서 검색하고 '영어-한글' 관계에 있는 단어들을 수집하여 기본 평가지표를 생성하였다.

4.2 평가지표

본 연구에서는 제안한 알고리즘의 세부적 성능을 평가하기 위하여 대체어를 '철자변형', '영어', '한글', '유의어' 등 총 4가지 경우로 분류하였으며 각각의 정의는 표 1과 같다.

평가지표에서 중심어란 "대체어를 찾고자 하는 기준 단어"를 의미하며, 대체어란 "한 문장에서 특정 단어를 대신하여 사용해도 문장의 의미를 훼손하지 않는 단어"를 의미한다. 본 연구에서 정의하는 철자변형, 영어, 한글, 유의어 등은 대체어의 유형별 지표이며 각 유형별 정의는 표 1과 같다. 표 6은 표 1의 정의에 따라 데이터베이스에 존재하는 단어들의 관계를 분류한 평가지표의 일부이다.

표 6 대체어 평가지표 (H04L)

중심어	대체어	유형
컨텐츠	콘텐트	철자변형
컨텐츠	콘텐츠	철자변형
컨텐츠	컨텐트	철자변형
가상사설망	VPN	영어
근거리통신망	LAN	영어
ACCESS	접근	한글
AP	액세스포인트	한글
가입자선	가입자라인	유의어
감지	인식	유의어
...	...	...

4.3 성능 평가

본 연구에서 집중도 임계치는 '집중도에 의한 단어의 분포'를 고려하여 설정하였다. 실험 결과에 따르면 집중도(식 (1))를 계산한 결과, H04L 분류에 속한 단어들이 0.3333을 중심으로 정규분포를 이루는 것으로 나타난다. 본 연구에서는 가장 많은 단어들이 모여 있는 지점인 0.3333을 H04L분류에서 특징을 지니지 못하는 일반적인 단어들이 위치하는 지점으로 가정하고, 이를 집중도 임계치로 설정하여 IPC분류별 중요단어를 선정하였다.

또한 연관단어생성 과정에서는 최소 신뢰도 α를 0으로 설정하여 특정 단어를 기준으로 신뢰도를 지니는 모

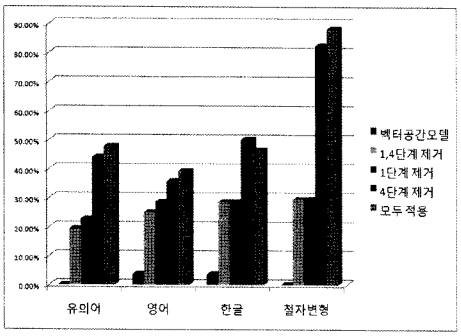

그림 2 벡터공간 모델 및 본 알고리즘 단계별 성능 비교 (α=0, β=1)

든 연관단어들을 추출한 후, 최소 지지도 β를 1로 설정하여 양의 상관관계(Positive Correlation)를 지니는 연관단어들만 취함으로써 연관단어 뭉치를 생성하였다.

실험 과정은 3장에서 제안한 절차에 따라 수행하였으며, 비교 평가를 위하여 벡터공간 모델을 이용한 실험도 수행하였다.

평가는 4.2절에서 생성한 평가지표를 이용하여 각 단어의 대체어 후보 목록 상위 10개에서의 재현율(Recall)을 측정하는 방식으로 진행되었다.

그림 2는 디지틀 정보의 전송(H04L) 분류의 특허 문헌 뭉치 내에서 기존의 유의어 추출 연구에서 사용하는 문서 벡터 공간을 이용한 대체어 추출 알고리즘과 본 논문에서 제안하는 알고리즘의 성능을 단계별로 비교한 자료이다. 제안하는 알고리즘의 단계별 성능 평가는 첫 단계인 "중요 단어 선정" 단계와 마지막 단계인 "대체어 순위 보정" 단계가 대체어 추출 성능에 미치는 영향을 알아보기 위한 것으로서 각 단계를 제거한 상태에서의 재현율을 비교하였다.

그림 2에 따르면 특허 문헌의 제목만 이용하여 대체어 추출을 시도할 경우 전체적으로 본 논문에서 제안하는 알고리즘이 기존의 문서벡터공간모델을 이용하는 것보다 월등한 성능을 보이는 것을 확인할 수 있다. 또한 본 알고리즘에서 중요하게 다루는 중요단어 선정 단계(1단계)와 대체어 순위 보정 단계(4단계)를 모두 적용하는 것이 그렇지 않은 경우보다 성능이 뛰어남을 확인할 수 있었다. 다만 영어, 한글의 경우 대체어 순위 보정 단계(4단계)를 적용할 경우 재현율이 떨어지는 것을 확인할 수 있는데, 이는 제목에서 특정 외래어를 한글로 표기한 후 괄호에 영어를 표기하여 부연 설명을 하거나 그 반대의 경우가 존재하기 때문이다.

5. 결론 및 향후 연구

본 논문에서는 특허 문헌 뭉치 속에서 대체어 후보를 추출하는 방법을 제안하였다. "의미가 비슷한 단어들은 같은 문맥에서 사용 될 것이다"는 기존 연구들과 동일한 가설로 접근하였으나, 중요 단어 선정 절차를 거치지 않고 문서 단어 동시 출현 빈도에 의존하여 분석하던 기존 연구들과 달리 특허 문헌의 IPC 분류를 이용하여 각 분류별 중요 단어를 선정하였다는 점과 그 단어들을 이용하여 문서벡터공간 모델 대신에 확률에 기반을 둔 연관 규칙을 이용하여 각 단어의 특징 벡터를 생성하였다는 점에서 선행 연구들과 차별된다. 또한 유사도를 비교하는 최종 단계에서 신뢰도를 이용하여 같은 문장에서 함께 나타날 가능성이 높은 단어들의 순위를 낮추어 줌으로써 대체어가 상위 10개 목록에 속할 확률을 높여주는 방법을 제안하였다.

본 연구에서는 연관단어 뭉치 간 유사도를 비교하는 척도로 코사인 유사도를 이용하였으나, 코사인 유사도 외에 다양한 척도를 이용한 비교 실험을 수행하여 대체어 추출 성능을 향상시키는 방법에 대한 연구가 필요하다. 또한, 평가 단계에서도 나타났듯이 영어, 한글 유형의 경우 괄호에 포함된 부연 설명으로 인해 대체어 순위 보정 단계에서 성능이 떨어지는 문제를 개선하기 위하여 이 경우를 따로 분리하여 대체어 순위 보정을 수행하는 방법에 대한 추가 연구가 필요하다.

참 고 문 헌

[1] 장백국제특허법률사무소, "선행기술 검색안내," http://www.k8.co.kr/htm/8-2_1.htm/.
[2] 박용준, "특허정보 검색방법", (주)아이피풀, 2005.
[3] Pierre P. Senellart and Vincent D. Blondel, "Automatic discovery of similar words," in Survey of Text Mining, Springer, 2003.
[4] Hsinchun Chen and Kevin J. Lynch, "Automatic construction of networks of concepts characterizing document databases," IEEE Transactions on Systems, Man and Cybernetics, Vol.22(5), 885-902, 1992.
[5] Magnus Sahlgren, "The Word-Space Model," Ph.D. Dissertation, Stockholm University, Stockholm, Sweden 2006.
[6] Jon M. Kleinberg, "Automatic construction of networks of concepts characterizing document databases," Journal of the ACM, Vol.46(5), 604-632, 1999.
[7] Vincent D. Blondel and Pierre P. Senellart, "Automatic extraction of synonyms in a dictionary," Presented at the TextMining Workshop, Arlington, Virginia, 2002.
[8] 이성진, "키워드 샵에서의 상품 추천을 위한 연관 키워드 그룹 추출 기법", M.S. Thesis, Soongsil University, Seoul, Korea 2003.
[9] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, 2nd ed., Morgan Kaufmann, 2006.