

■ 2009년도 학생논문 경진대회 수상작

특허 정보 검색 품질 향상을 위한 대체어 후보 자동 생성 방법

(Automatic Construction of Alternative Word Candidates to Improve Patent Information Search Quality)

백종범[†] 김성민[†] 이수원^{††}
(Jongbum Baik) (Seongmin Kim) (Soowon Lee)

요약 정보 검색에서 원하는 정보를 얻지 못하는 원인은 다양하다. 그 중에서도 표기의 다양성은 검색 시 키워드 불일치로 인한 정보 누락을 발생시키는 원인이 된다. 본 논문은 이러한 키워드 불일치에 의한 정보 누락을 최소화하기 위하여 검색 대체어 후보를 자동 생성하는 방법을 제안한다. 본 연구에서 제안하는 대체어 후보 자동 생성 방법은 문장 내에서 함께 쓰이는 단어들이 비슷한 두 단어는 서로 비슷한 의미를 지닐 것이라는 직관적 가설을 전제로 한다. 이와 같은 가설을 기반으로 하여 본 연구에서는 분류별 집중도, 신뢰도를 이용한 연관단어 뭉치, 연관단어 뭉치 간 코사인 유사도 및 신뢰도를 이용한 필터링 기법 등을 이용한 대체어 후보 자동 생성 방법을 제안한다. 본 연구에서 제안한 대체어 후보 자동 생성 방법의 성능은 대체어 유형별로 작성된 평가지표를 이용하여 정확도 및 재현율을 측정함으로써 평가되었으며, 제안 방법이 context window overlapping을 이용한 대체어 추출 방법보다 더 우수한 것으로 나타났다.

키워드 : 대체어, 유의어, 연관단어, 정보 검색

Abstract There are many reasons that fail to get appropriate information in information retrieval. Allomorph is one of the reasons for search failure due to keyword mismatch. This research proposes a method to construct alternative word candidates automatically 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 filtering technique using confidence. Performance of the proposed method is evaluated using a manually extracted alternative list. Evaluation results show that the proposed method outperforms the context window overlapping in precision and recall.

Key words : Allomorph, Synonym, Associated Word, Information Retrieval

· 본 연구는 숭실대학교 교내연구비 지원으로 수행되었다.

[†] 학생회원 : 숭실대학교 컴퓨터학과
jbb100@mining.ssu.ac.kr
mabak@mining.ssu.ac.kr

^{††} 종신회원 : 숭실대학교 컴퓨터학부 교수
swlee@ssu.ac.kr

논문접수 : 2009년 5월 22일

심사완료 : 2009년 8월 14일

Copyright©2009 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 소프트웨어 및 응용 제36권 제10호(2009.10)

1. 서 론

일반적으로 검색을 힘들게 하는 요소는 크게 단어의 다의성과 단어표기의 다양성으로 나눌 수 있다. 단어의 다의성이란 하나의 단어가 다양한 의미로 쓰이는 것을 의미하며, 단어표기의 다양성이란 다르게 표기된 단어들이 같은 의미로 쓰이는 것을 의미한다[1]. 단어표기의 다양성은 외래어 표현, 축약형 표현, 의도적 비표준 표현 등을 포함한다[2].

단어표기의 다양성 중에서도 특히 외래어를 한글로 표기하는 경우(외래어 표현)에 다양한 표현이 많이 나타난다. 예를 들어 ‘contents’라는 영어 단어를 한글 문서

에 기술하는 경우 ‘컨텐츠’, ‘컨텐트’, ‘콘텐트’, ‘콘텐츠’ 등 다양한 방법으로 표기가 가능하다. 반대로 그대로 영어로 기술하더라도 ‘TV’와 같은 줄임말인 경우(축약형 표현)에는 ‘T.V’와 같이 중간에 ‘.’을 추가하거나 ‘television’과 같이 줄임말을 풀어서 기술하는 등 다양한 표현이 사용될 수 있다. 또한 특허 문헌의 경우 검색결과에 특허문헌이 노출되는 것을 회피하기 위하여 “텔레비죤”과 같이 잘 사용되지 않는 용어(의도적 비표준 표현)를 사용하기도 한다.

단어표기의 다양성을 해결하기 위한 연구는 크게 ‘유의어 발견 연구’와 ‘대체어 발견 연구’로 나누어진다. 유의어 발견 연구는 대부분 “의미가 비슷한 단어들은 같은 문맥(Context)에서 사용될 것이다”라는 가설을 전제로 하고 있다[3-7]. 이러한 연구들은 문맥을 문서 혹은 문장으로 정의하고, 특정 단어가 특정 문맥에서 출현했는지 여부를 논리값(Boolean) 또는 출현빈도로 기술한 벡터공간모델을 이용하거나 벡터공간모델에서 단어와 문맥 간의 Pairwise Mutual Information(PMI)을 추출함으로써 문맥 간의 의미적 유사도를 계산한다.

그러나 이러한 유의어 발견 연구들은 같은 문맥에서 출현하지 않은 단어들에 대해서는 유사도 검사를 수행하지 않으므로 인하여 유사 단어들을 충분히 발견하지 못하는 동시출현문제(Co-occurrence Problem)를 지니고 있다.

최근 캐나다 및 스페인 등에서 이러한 동시출현문제를 해결하기 위한 시도로서 “문장 내에서 함께 쓰이는 단어들이 비슷한 두 단어는 서로 비슷한 의미를 지닐 것이다”라는 가설을 기반으로 하는 대체어 발견 연구 [8,9]가 제안되었다. 그러나 이러한 연구들은 주로 제한된 실험용 데이터와 단어들만 이용하여 평가를 수행하였으므로 그 실제 성능은 미지수이며, 주로 영어권에서 이루어진 연구들이므로 한글 문서 검색에 적용하였을 때의 성능 또한 보장할 수 없다.

본 연구에서는 이러한 기존의 유의어 및 대체어 발견 연구들을 한글 문서 검색, 특히 특허 검색에 적용할 때의 문제점을 분석 및 보완하여 한글 특허 검색에 적합한 대체어 후보 자동 생성 방법을 제안한다. 본 논문에서 정의하는 대체어란, “한 문장에서 특정 단어를 대신하여 사용해도 문장의 의미를 훼손하지 않는 단어”를 의미하며, 특히 문헌 데이터의 특성을 고려하여 대체어를 표 1과 같이 3가지 경우로 분류하여 사용한다. 특히, 본 연구에서는 표 1의 분류 중 타 분류에 비해 사용자들이 예측하기 힘든 이형어를 찾아내는데 중점을 둔다.

본 연구에서 제안하는 방법은 먼저 IPC(International Patent Classification)별 중요 단어를 선정하고, 이를 이용하여 연관단어 뭉치를 생성하는 과정을 수행한다.

표 1 대체어 분류의 정의

분류	정의
이형어	기준단어와 동일한 대상을 다른 철자로 표기한 경우
대역어	영어로 표기된 기준단어에 대한 한글 표기 혹은 그 반대의 경우
유의어	기준단어와 비슷한 의미를 지닌 단어

그 후, 생성된 연관단어 뭉치의 유사도를 계산하여 대체어 후보 목록을 생성하고, 마지막으로 대체어 후보 목록 내에 존재하는 연관단어를 필터링하는 과정을 수행한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 기존의 유의어 및 대체어 발견 연구들을 소개하고, 3장에서는 본 연구에서 제안하는 대체어 후보 자동 생성 알고리즘을 소개한다. 4장에서는 3장에서 제안하는 알고리즘의 성능을 평가하며, 마지막으로 5장에서는 결론 및 향후연구를 제시한다.

2. 관련 연구

본 연구에서는 표기의 다양성을 해결하기 위한 기존 연구들을 동시출현문제의 여부에 따라 ‘유의어 발견 연구’와 ‘대체어 발견 연구’로 분류한다. 본 연구에서 언급하는 ‘유의어 발견 연구’는 그림 1과 같이 대체어의 일부 유형(표 1)으로서의 유의어를 발견하기 위한 연구를 의미하며, ‘대체어 발견 연구’란 이러한 유의어 뿐 아니라 ‘유의어 발견 연구’에서 동시출현문제로 인해 발견할 수 없었던 이형어, 대역어 등의 유형까지 포함한 대체어를 발견하기 위한 연구를 의미한다.

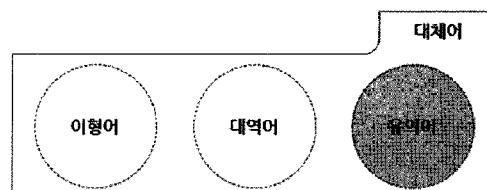
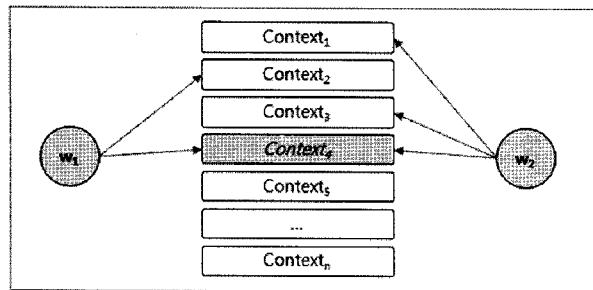


그림 1 대체어와 유의어의 관계

2.1 대량 문서 뭉치를 이용한 유의어 발견 연구

2.1.1 Edmonton의 유의어 발견 연구

Edmonton의 연구[6]에서는 “두 단어가 각각 쓰이는 문맥이 비슷하다면, 이 두 단어의 의미 또한 비슷할 것이다”라는 가설을 전제로 한다. 이를 그림으로 표현해보면 그림 2와 같이 나타낼 수 있다. 그림 2에서 w_1 은 $Context_2$, $Context_4$ 에서 등장한 단어이며, w_2 는 $Context_1$, $Context_3$, $Context_4$ 에서 등장한 단어이다. 두 단어 w_1 , w_2 가 각각 출현한 문맥 중 같이 출현한 문맥은 $Context_4$ 하나이며, 두 단어 사이에 이러한 동시 출현 문맥이 많을수록 두 단어 간의 유사도가 높아진다.



이러한 가설에 따라 [6]에서는 단어 간 유사도를 구하기 위하여 먼저 식 (1)과 같이 단어와 문맥간의 PMI를 계산한 다음, 식 (2)와 같이 이를 정규화 하는 과정을 거침으로써 단어별 특징 벡터를 구성한다. 이렇게 생성된 특징 벡터를 이용하여 최종적으로 두 단어의 특징 벡터 간의 코사인 유사도를 계산하여 두 단어 간의 유사도를 구한다.

$$mi_{w,c} = \frac{F_c(w)}{\sum_i F_i(w) \sum_j F_c(j)}$$

$mi_{w,c}$: 단어 w 와 문맥 c 사이의 상호정보량
 $F_c(w)$: 문맥 c 에서 단어 w 가 출현한 빈도

식 (1) 단어와 문맥 간 PMI 계산식

$$normMi_{w,c} = mi_{w,c} \times \frac{F_c(w)}{F_c(w)+1} \times \frac{\min(\sum_i F_i(w), \sum_j F_c(j))}{\min(\sum_i F_i(w), \sum_j F_c(j))+1}$$

$normMi_{w,c}$: 정규화 된 상호정보량
 $mi_{w,c}$: 단어 w 와 문맥 c 사이의 상호정보량
 $F_c(w)$: 문맥 c 에서 단어 w 가 출현한 빈도

식 (2) PMI 정규화 수식

2.2 대량 문서 둥치를 이용한 대체어 발견 연구

2.2.1 SOC PMI

SOC PMI(Second Order Co-occurrence PMI)[8]는 PMI-IR[7]에 존재하는 동시출현문제를 해결하기 위하여 제안된 방법이다. SOC PMI에서는 단어 간의 유사도를 구하기 위하여 먼저 단어 간의 PMI를 계산함으로써 각 단어별로 특징벡터를 구성한다. 그 다음, 식 (3)을 통하여 각 단어별 특징벡터의 상위 β_i 개만 취하여 단어 간 유사도를 비교한다.

$$\beta_i = (\log(f^*(w_i)))^2 \frac{(\log_2(n))}{\delta}$$

β_i : 단어 i 에 대한 중요/ PMI 선정 개수 (상위 β_i 개 선정)
 $f^*(w_i)$: 단어 w_i 의 문서 둥치 내 출현 빈도

n : 중복을 제거한 단어의 개수
 δ : 문서 길이에 따라 설정하는 임계치

식 (3) 중요 PMI 선정을 위한 수식

SOC PMI에서는 단어 간 유사도를 계산하기 위하여 식 (4)와 식 (5)를 제안하였다. 식 (4)는 두 단어가 공통으로 지니고 있는 단어들들의 PMI 값을 서로 교환하여 합치는 수식이다. 식 (5)는 식 (4)에서 구한 두 단어 w_1 과 w_2 의 공통 단어들의 PMI 합을 각각 중요 PMI 선정 개수인 β_1 과 β_2 로 나눈 후, 합쳐줌으로써 최종적인 단어 간 유사도 점수를 산출한다.

$$f^\beta(w_1) = \sum_{i=1}^{\beta_1} (f^{pmi}(X_i, w_2))^\gamma$$

$$f^\beta(w_2) = \sum_{i=1}^{\beta_2} (f^{pmi}(Y_i, w_1))^\gamma$$

$f^\beta(w_1)$: 일치한 PMI의 합
 X_i : w_1 과 PMI로 연결된 단어들의 집합
 $f^{pmi}(X_i, w_2)$: X_i 와 w_2 사이의 PMI 값

식 (4) 일치한 단어들의 PMI 합

$$sim(w_1, w_2) = \frac{f^\beta(w_1) + f^\beta(w_2)}{\beta_1 + \beta_2}$$

$sim(w_1, w_2)$: 두 단어 w_1, w_2 간의 유사도 점수
 $f^\beta(w_1), f^\beta(w_2)$: 일치한 PMI의 합
 β_1, β_2 : 중요 PMI 선정 개수 (상위 β_i 개 선정)

식 (5) 유사도 점수

2.2.2 Context-Window Overlapping

Context-Window Overlapping(이하 CWO)[9]은 기존의 통계 정보에 의존한 알고리즘이 회소 출현 단어를 찾아내지 못하는 문제점을 개선하기 위하여 제안된 방법이다. CWO에서는 두 단어 w_1, w_2 간의 대체 가능성인 $count(w_1, w_2)$ 를 계산하기 위하여 먼저 w_1 으로 초기 검색을 수행한 후, 미리 정해진 길이(Window Length)의 Context를 추출하는 단계를 거친다. 그 다음, 추출된 각 Context에 대하여 w_1 을 w_2 로 대체하여 재검색을 수행한다. 재검색 결과에서 검색된 결과가 한 개라도 존재한다면 이 두 단어는 대체어일 가능성이 있다고 판단하여 카운트 개수를 1 증가 시킨다. 이와 같은 과정을 다음 Context에 대해서도 반복 수행하며 카운트 개수를 누적함으로서 두 단어가 대체 가능한 단어인지 여부를 수치화한다.

예를 들어 ‘텔레비전’으로 검색된 결과 중, Window Length가 3인 문맥이 총 두 개 존재하는 경우, 이를 각각에서 ‘텔레비전’을 ‘TV’로 바꾸어 검색한 결과가 각각 1개 이상씩 존재한다면 최종 대체어 점수를 2로 산출한

다. CWO에서는 이와 같은 방식으로 대체어를 찾아내는 방법을 one-way_similarity라고 명칭하며, $\text{count}(w_1, w_2) + \text{count}(w_2, w_1)$ 과 같이 두 단어를 서로 바꾸어 카운트한 결과를 합친 것을 two-way_similarity라고 명칭한다.

3. 대체어 후보 자동 생성

본 연구에서 제안하는 ‘대체어 후보 자동 생성 시스템’은 “문장 내에서 함께 쓰이는 단어들이 비슷한 두 단어는 서로 비슷한 의미를 지닐 것이다”라는 직관적 가설을 전제로 한다. 제안 시스템의 구조도는 그림 3과 같으며, 수집된 특허 문헌에 대한 IPC별 중요 단어 선정 단계[1단계], 연관단어 뭉치를 생성하는 단계[2단계], 생성된 연관단어 뭉치의 유사도를 계산하여 대체어 후보 목록을 생성하는 단계[3단계], 대체어 내에 존재하는 연관단어를 걸러내는 필터링 단계[4단계]로 구성된다.

본 장에서는 총 네 단계의 주요 프로세스를 각 절에서 상세히 설명하며, 특히 문헌 수집 모듈은 4장에서 언급한다.

3.1 IPC별 중요 단어 선정

단어별 대체어를 생성하기 위해서는 해당 IPC 분류에서 중요한 단어가 무엇인지 선정하는 단계가 필요하다. 만약 이 단계를 생략한다면 알고리즘 수행 과정에 키워드가 아닌 단어들이 다수 포함되어 수행 시간이 많이 소요되는 것은 물론이고 최종 결과 역시 왜곡될 가능성에 존재한다.

본 연구에서의 IPC별 중요 단어 선정 단계는 기수집된 특허문헌 내에서 IPC별 색인어를 추출한 후, ‘분류별 집중도’를 반영한 TFIDF[10,11]’를 변형하여 재정의한 ‘분류별 집중도’(식 (6))를 이용하여 수행된다. 집중도의 기본 아이디어는 “특정 단어 w 가 특정 문서 분류 G 에서는 자주 출현하지만 타 문서 분류에서는 상대적으로

적게 출현한다면 w 는 문서 분류 G 에 있어서 중요한 키워드일 것이다”라는 가설을 전제로 한다.

식 (6)은 이러한 가설을 이용하여 문서 분류 i 내에서 단어 w 의 중요도를 (문서 분류 i 에서 단어 w 의 중요도)/(모든 문서 분류에서의 단어 w 의 평균적인 중요도)와 같이 상대 비율로 나타낸 수식이다. 식 (6)에서 concentration $_i(w)$ 는 분류 i 에서 단어 w 가 지니는 집중도로 정의된다. $n_i/(n_i - df_i(w))$ 는 문서 분류 i 내에서 단어 w 가 중요한 정도를 의미하며, $1 - gf(w)/N_G$ 는 여러 문서 분류에서 출현한 단어 w 의 가중치를 낮추어주는 역할을 한다.

$$\text{concentration}_i(w) = \frac{\left(\frac{n_i}{n_i - df_i(w)}\right)\left(1 - \frac{gf(w)}{N_G}\right)}{\left(\sum_{j \in G} \left(\frac{n_j}{n_j - df_j(w)}\right)\left(1 - \frac{gf(w)}{N_G}\right)\right) / G}$$

n_i : 문서 분류 i 에 속하는 문서의 수

$df_i(w)$: 문서 분류 i 에서 단어 w 가 출현한 문서의 수

G : 전체 문서 분류 집합

N_G : 전체 문서 분류의 수

$gf(w)$: 단어 w 가 출현한 문서 분류의 수

식 (6) 분류별 집중도 계산식

일반적으로 식 (6)에 의해 계산된 집중도가 1이 상인 단어, 즉, 모든 분류 내 중요도 평균치를 초과하는 단어를 해당 분류의 중요 단어로 선정할 수 있으나, 이는 사용자의 의도에 따라 달라질 수 있다.

표 2는 IPC 분류 H04N(화상 통신), G06F(전기에 위한 디지털 데이터 처리)에서 각각 집중도를 계산한 후, 각 분류별로 상위 5개 단어와 하위 5개 단어를 추출한 결과이다. 표 2에 따르면 상위 5개 단어에서 H04N 분류와 G06F 분류 간에 같은 단어가 존재하지 않는 것으로 나타난다. 반면에 각 IPC 분류별 집중도 하위 5개

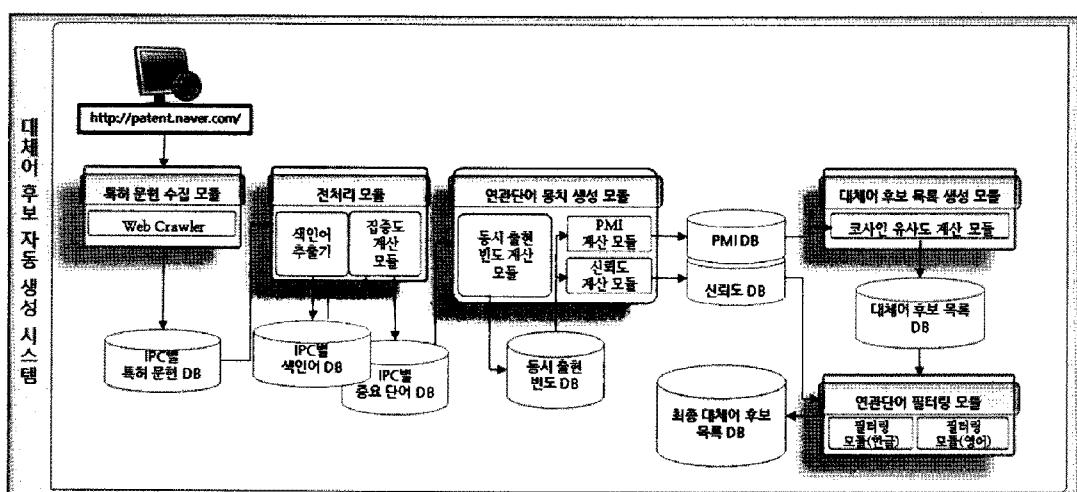


그림 3 대체어 후보 자동 생성 시스템 구조도

표 2 IPC별 집중도 계산 결과 (상위 5개, 하위 5개)

IPC	단어	집중도	IPC	단어	집중도
H04N	텔레비전	1.0619	G06F	컴퓨터	1.1100
H04N	비디오	1.0578	G06F	인터넷	1.0661
H04N	디지털	1.0560	G06F	이용	1.0473
H04N	영상	1.0548	G06F	메모리	1.0371
H04N	텔레비전	1.0539	G06F	모니터	1.0219
...
H04N	산출	0.9661	G06F	신호	0.9568
H04N	통신	0.9541	G06F	회로	0.9495
H04N	회로	0.9495	G06F	장치	0.9369
H04N	시스템	0.0000	G06F	시스템	0.0000
H04N	방법	0.0000	G06F	방법	0.0000

단어에서는 ‘회로’, ‘시스템’, ‘방법’ 등 많은 단어들이 일치하는 것으로 나타난다. 특히 ‘시스템’, ‘방법’의 경우 모든 IPC에서 등장하는 단어이므로 집중도가 0이 되는 것으로 나타난다. 이러한 결과를 볼 때, 집중도가 높을 수록 각 분류를 대표할 수 있는 단어이며 반대의 경우에는 여러 분류에서 자주 등장하는 일반적인 단어라는 것을 알 수 있다.

3.2 연관단어 뭉치 생성

본 연구에서는 연관단어 뭉치를 추출하는 방법으로 PMI[6-8]와 신뢰도(Confidence)[10,11]를 이용한다. PMI를 이용한 연관단어 뭉치는 “특정 단어 w 를 중심으로 양의 상관관계($PMI > 0$)를 지니는 단어들의 집합”으로 정의되며 이는 3.3절에서 대체어 후보를 추출하기 위한 특정 벡터로 이용된다. 또한 신뢰도를 이용한 연관단어 뭉치는 “특정 단어 w 와 함께 쓰일 확률이 높은 단어들의 집합”으로 정의되며 이는 3.4절에서 대체어 후보 목록 내에 존재하는 연관단어들을 제거하기 위해 사용된다.

신뢰도를 이용한 연관단어 뭉치는 표 3과 같이 하나의 기준단어를 기준으로 해당 IPC분류 내에서 자신을 제외한 모든 단어에 대하여 식 (7)을 계산함으로써 생성된다.

$$conf(w_1 \Rightarrow w_2) = \frac{p(w_1 \cap w_2)}{p(w_1)}$$

$conf(w_1 \Rightarrow w_2)$: 단어 w_1 이 출현한 문맥에 단어 w_2 가 출현할 확률

$p(w_1 \cap w_2)$: 두 단어 w_1 과 w_2 가 같은 문맥에 동시 출현할 확률

$p(w_1)$: 단어 w_1 이 출현할 확률

식 (7) 신뢰도 계산식

PMI를 이용한 연관단어 뭉치는 표 4와 같이 하나의 기준단어를 기준으로 해당 IPC분류 내에서 자신을 제외한 모든 단어에 대하여 식 (8)을 계산함으로써 생성된다.

표 3 신뢰도를 이용한 연관단어 뭉치 상위 5개 목록

기준단어	빈도	연관어	빈도	동시 출현 빈도	신뢰도
텔레비전	3215	수상기	1498	541	0.1683
텔레비전	3215	디지털	4416	267	0.0830
텔레비전	3215	기능	1491	205	0.0638
텔레비전	3215	자동	2257	199	0.0619
텔레비전	3215	채널	1670	166	0.0516

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1 \cap w_2)}{p(w_1)p(w_2)}$$

식 (8) PMI 계산식

표 4 PMI를 이용한 연관단어 뭉치 상위 5개 목록

기준단어	빈도	연관어	빈도	동시 출현 빈도	PMI
텔레비전	3215	음량제어장치	15	12	3.8738
텔레비전	3215	팩스기능	13	10	3.8166
텔레비전	3215	다화	69	51	3.7592
텔레비전	3215	2-튜너	25	18	3.7214
텔레비전	3215	방지기능	14	10	3.7104

표 3에 따르면 신뢰도를 이용하여 추출한 연관단어 뭉치들은 대부분 고빈도 단어들이 높은 신뢰도를 지니는 것으로 나타나며, 반대로 PMI를 이용하여 추출한 연관단어 뭉치들인 표 4에 따르면 대부분 상대적으로 저빈도인 단어들이 높은 PMI를 지니는 것으로 나타난다.

일반적으로 고빈도 단어들은 특정 한 단어가 아닌 여러 단어들과 연관 관계를 지닌다. 이에 반해 저빈도 단어들은 하나 혹은 소수의 특정 단어와 연관 관계를 지닌다.

이러한 신뢰도와 PMI의 특징을 고려할 때, 신뢰도는 대체어 후보 목록 내에 존재하는 연관단어를 제거하는데 이점을 지니며, PMI는 각 단어를 특징지어 줄 수 있는 특징벡터를 구성하는데 이점을 지닌다는 것을 알 수 있다.

3.3 대체어 후보 목록 생성

본 연구에서는 대체어 후보 목록을 생성하기 위하여 두 특징 벡터(PMI를 이용한 연관단어 뭉치) 간의 유사도를 측정하는 척도로서 코사인 유사도(식 (9))를 이용하였다. 코사인 유사도는 일반적으로 정보검색(Information Retrieval) 분야에서 자주 쓰이는 방법으로, 계산 시 두 기준단어의 연관단어 뭉치 내에서 일치한 PMI값의 차이가 적을수록 유사도 값이 커지는 특징을 지닌다.

$$sim(A_{w_1}, A_{w_2}) = \frac{\vec{A}_{w_1} \cdot \vec{A}_{w_2}}{|A_{w_1}| \cdot |A_{w_2}|}$$

식 (9) 코사인 유사도 계산식

표 5는 두 단어 ‘플래시’와 ‘플래쉬’ 간의 코사인유사도를 계산하는 예이다. 두 단어 간의 유사도를 구하기 위해서는 먼저 비교하고자 하는 두 단어의 연관단어 풍차에 존재하는 연관단어들로 하나의 벡터 공간을 구성한 후, 벡터 공간 내에 각 두 단어와 연관단어 사이의 PMI 값을 기술한다. 그 다음, 식 (9)를 이용하여 두 단어의 벡터 공간 간 유사도를 계산함으로써 두 비교 단어의 유사도를 측정할 수 있다.

표 6은 이와 같은 방법으로 계산된 ‘플래시’라는 단어의 대체어 후보 목록으로, ‘플래시’의 대체어인 ‘플래쉬’가 대체어 후보 목록 내 2순위에 위치하는 것으로 나타난다. 이러한 결과는 ‘플래시’와 ‘플래쉬’가 높은 유사도를 지녔다는 점에 있어서는 좋은 결과이다. 그러나 ‘플래시’의 대체어로서 ‘플래쉬’ 외에도 ‘플레쉬’라는 대체어도 존재할 수 있으므로 이는 결코 만족스럽지 못한 결과이다. 다음 절에서는 이러한 결과를 개선하기 위한 방법을 소개한다.

표 5 ‘플래시’와 ‘플래쉬’ 간의 코사인 유사도 계산 예

비교 단어	플래시	플래쉬
메모리소자	0.0000	8.3312
메모리셀	0.0000	8.5536
등급	5.5993	0.0000
능력	5.9212	0.0000
생신	3.9439	3.5765
업그레이드	4.0635	4.6954
메모리	4.1027	3.9430
마이크로콘트롤러	4.8339	5.4660
프로그래밍	5.0144	4.0618
이이피룸	6.0144	5.6467
...

표 6 ‘플래시’의 대체어 후보 목록

순번	기준단어	대체어	유사도
1	플래시	메모리	0.2939
2	플래시	플래쉬	0.2664
3	플래시	기입	0.2659
4	플래시	마이크로컴퓨터	0.2474
5	플래시	레지스터	0.2193
6	플래시	시리얼	0.2172
7	플래시	불휘발성	0.2147
8	플래시	데이터	0.2037
9	플래시	이이피룸	0.2001
10	플래시	그것	0.1997

3.4 대체어 후보 목록 필터링

대체어 후보 목록을 코사인 유사도에만 의존하여 생성하는 경우, 같은 문맥에서 같이 자주 등장한 단어들(연관단어)이 대체어 후보 목록 내에서 높은 유사도를 지니는 문제가 종종 발생한다. 표 7에 따르면 이러한 문제로 인해 ‘플래시’의 대체어인 ‘플래쉬’가 대체어 후보 목록 내 22순위에 위치하는 것으로 나타난다. 본 절에서는 이러한 문제를 해결하기 위하여 신뢰도를 이용한 대체어 후보 목록 필터링 방법을 제안한다.

신뢰도를 이용한 대체어 후보 목록 필터링 방법은 단어 쌍의 종류에 따라 두 가지 경우로 나누어 적용된다. 만약 단어 쌍이 둘 다 한글이라면, 이 둘은 표 7에 나타나는 바와 같이 서로 연관단어일 가능성이 높으므로 그 둘 간의 유사도를 신뢰도를 이용하여 낮추어주어야 한다(식 (10)).

그러나 만약 단어 쌍에 영어가 포함되어 있다면, 이 둘은 서로 부연 설명일 가능성이 높으므로 신뢰도를 이용하여 그 둘 간의 유사도를 높여주어야 한다. 즉, 이는 한글로 작성된 문헌들에서 자주 등장하는 티비이(TV), 퍼스널 컴퓨터(PC) 등과 같은 표기 정보를 이용하여 ‘대역어’ 유형의 대체어 발견 확률을 높이기 위한 것이다(식 (11)).

표 7 ‘플래시’의 대체어 후보 목록 및 신뢰도

순번	기준 단어	대체어	유사도 (X=>Y)	CONF (X=>Y)	CONF (Y=>X)
1	플래시	메모리	0.2939	0.7258	0.0354
2	플래시	플래쉬	0.2664	0.0000	0.0000
3	플래시	기입	0.2659	0.0161	0.0357
4	플래시	마이크로 컴퓨터	0.2474	0.0645	0.0455
5	플래시	레지스터	0.2193	0.0161	0.0079
...
20	플래시	디스크	0.1819	0.0161	0.0046
21	플래시	EEPROM	0.1780	0.0323	0.2000
22	플래시	플래쉬	0.1774	0.0000	0.0000

식 (10)과 식 (11)은 기준단어를 중심으로 Max 정규화 된 코사인 유사도와 신뢰도를 이용하여 최종 유사도 점수를 산출하는 식이다.

$$simScore(w_1, w_2) = normSim(A_{w_1}, A_{w_2}) - \gamma(normConf(w_1 \Rightarrow w_2) + normConf(w_2 \Rightarrow w_1))$$

simScore(w₁, w₂) : 두 단어 w₁, w₂ 간의 유사도 점수
normSim(A_{w1}, A_{w2}) : 두 연관 단어 풍차 A_{w1}, A_{w2}의 유사도를 정규화한 값

normConf(w₁=>w₂) : w₁이 등장한 문맥에 w₂가 같이 등장할 확률을 정규화한 값
γ : 필터링 정도 조절 임계치

식 (10) 단어 쌍이 둘 다 한글일 경우의 유사도 점수

$$\begin{aligned} simScore(w_1, w_2) &= normSim(A_{w_1}, A_{w_2}) \\ &+ \lambda (normConf(w_1 \Rightarrow w_2) + normConf(w_2 \Rightarrow w_1)) \end{aligned}$$

λ : 필터링 정도 조절 임계치

식 (11) 단어 쌍 중 하나가 영어일 경우

표 8과 표 9는 각각 코사인 유사도(식 (9))와 신뢰도(식 (7))를 Max 정규화하는 예이다. 만약 이러한 정규화 과정이 생략된 채 대체어 필터링 과정을 수행한다면, 유사도 계수에 따라 혹은 데이터양에 따라 필터링되는 정도가 일정하지 않게 된다. Max 정규화 외에도 다양한 정규화 방법들이 존재하지만 본 연구에서는 중심단어에 대한 상대적 중요도를 계산하기 위하여 Max 정규화가 가장 적합하다고 판단하여 이를 이용하여 정규화 과정을 수행하였다.

표 8 코사인 유사도 Max 정규화 예

(a) 코사인 유사도			(b) 정규화 결과			
순위	기준 단어	대체어	유사도	기준 단어	대체어	정규값
1	플래시	메모리	0.2939	플래시	메모리	1.0000
2	플래시	플래쉬	0.2664	플래시	플래쉬	0.9064
3	플래시	기입	0.2659	플래시	기입	0.9047
4	플래시	마이크로컴퓨터	0.2474	플래시	마이크로컴퓨터	0.8418
5	플래시	레지스터	0.2193	플래시	레지스터	0.7462

표 9 신뢰도 Max 정규화 예

(a) 연관단어 봉치			(b) 정규화 결과			
순위	기준 단어	연관어	신뢰도	기준 단어	연관어	정규값
1	플래시	메모리	0.7258	플래시	메모리	1.0000
2	플래시	이용	0.2097	플래시	이용	0.2889
3	플래시	데이터	0.1290	플래시	데이터	0.1777
4	플래시	프로그램	0.1129	플래시	프로그램	0.1556
5	플래시	제어	0.0806	플래시	제어	0.1110

표 10은 단어 ‘플래시’에 대한 대체어 후보 목록의 필터링 전과 후를 비교한 결과이다. 표 10(a)에 따르면 ‘플래시’의 대체어로 ‘플래쉬’만 나타나지만, 필터링 과정을 거친 표 10(b)에 따르면 ‘플래쉬’라는 숨겨져 있던 대체어가 나타나는 것을 확인할 수 있다.

표 11은 영어 단어 ‘PC’에 대한 대체어 후보 목록의 필터링 전과 후를 비교한 결과이다. 표 11(a)에 따르면 ‘PC’의 대체어가 나타나지 않지만, 필터링 과정을 거친 표 11(b)에 따르면 ‘파시’, ‘퍼스널컴퓨터’라는 숨겨져 있던 대체어가 나타나는 것을 확인할 수 있다.

표 10 단어 ‘플래시’의 대체어 후보 목록 변화

(a) 대체어 후보 목록 (b) 필터링 결과

순위	기준 단어	대체어	유사도	기준 단어	대체어	보정값
1	플래시	메모리	0.2939	플래시	EEPROM	1.1279
2	플래시	플래쉬	0.2664	플래시	플래쉬	0.9064
3	플래시	기입	0.2659	플래시	기입	0.7826
4	플래시	마이크로컴퓨터	0.2474	플래시	블휘발성	0.7305
5	플래시	레지스터	0.2193	플래시	다수개의	0.6764
6	플래시	시리얼	0.2172	플래시	레지스터	0.6650
7	플래시	블휘발성	0.2147	플래시	원침	0.6563
8	플래시	데이터	0.2037	플래시	하드웨어	0.6322
9	플래시	이이피롬	0.2001	플래시	그것	0.6098
10	플래시	그것	0.1997	플래시	플래쉬	0.6036

표 11 영어 단어 ‘PC’의 대체어 후보 목록 변화

(a) 대체어 후보 목록 (b) 필터링 결과

순위	기준 단어	대체어	유사도	기준 단어	대체어	보정값
1	PC	전원	0.2646	PC	전원	1.2377
2	PC	32비트	0.2311	PC	파시	1.0050
3	PC	컴퓨터	0.2111	PC	노트북	0.9905
4	PC	포트	0.1992	PC	컴퓨터	0.9587
5	PC	키보드	0.1858	PC	인터페이스	0.9268
6	PC	컴퓨터용	0.1848	PC	32비트	0.8734
7	PC	노트북	0.1779	PC	키보드	0.8518
8	PC	인터페이스	0.1750	PC	포트	0.8347
9	PC	퍼스널컴퓨터	0.1734	PC	퍼스널컴퓨터	0.8269
10	PC	파드	0.1713	PC	모니터	0.8119

4. 실험 및 평가

4.1 실험 데이터

3장에서 제안한 방법의 평가를 위하여 ‘네이버 특허서비스’에 존재하는 특허 문헌 중 단어의 의미를 판별하기 쉬운 ‘화상 통신(H04N, 58,901건)’, ‘전기에 의한 디지털 데이터 처리(G06F, 30,059건)’ 분류에서 특허 문헌을 수집하여 실험 및 평가를 수행하였다. 또한 본 연구에서는 특허문헌의 제목과 초록 중 제목만 이용하여 실험 및 평가를 수행하였다. 그 이유는 4.4.4절에서 상세히 다룬다.

4.2 임계치 설정

본 실험에 있어서 필요한 임계치는 분류별 집중도(식 (6))를 계산한 후, 중요단어를 선정하기 위한 ‘중요 단어 선정 임계치’와 대체어 후보 목록 필터링 기법(식 (10)과 식 (11))에서 ‘필터링 정도를 조절하기 위한 임계치’

등이 존재한다.

먼저 중요 단어 선정 임계치의 결정은 IPC분류 H04N, G06F에서 각각 계산된 집중도 분포를 이용하여 결정하였다. 그림 4와 그림 5에 따르면 해당 문서 분류에서 별다른 특징을 지니지 못하는 단어들이 1 주변으로 모여드는 현상이 나타난다. 이러한 현상은 집중도 수식(식 (6))에서 의도하는 바가 잘 반영된 결과이다. 본 연구에서는 집중도가 1.0001을 초과하는 단어들을 해당 IPC의 중요단어라고 정의하고 이를 만족하는 단어들을 이용하여 실험을 수행하였다.

또한 집중도를 이용하여 중요 단어들만 추출한 결과, 단어 및 PMI 단어 쌍의 개수를 대폭 줄임으로 인해 알고리즘의 수행 속도에 있어서 큰 이점을 지니는 것으로 나타났다(표 12, 표 13).

필터링 정도 조절 임계치는 단어 쌍이 모두 한글일 경우에는 1을 설정하여 신뢰도를 강하게 반영하였다. 그 이유는 본 연구가 특허 문헌의 제목만 이용하여 실험 및 평가를 수행하였으므로 제목 내에서 특정 용어를 일

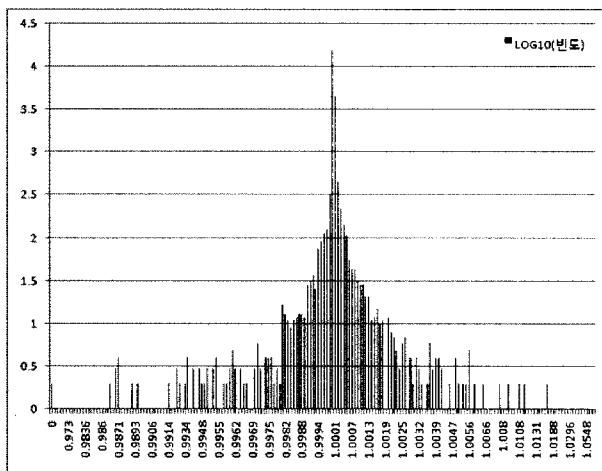


그림 4 집중도에 의한 단어의 분포(H04N)

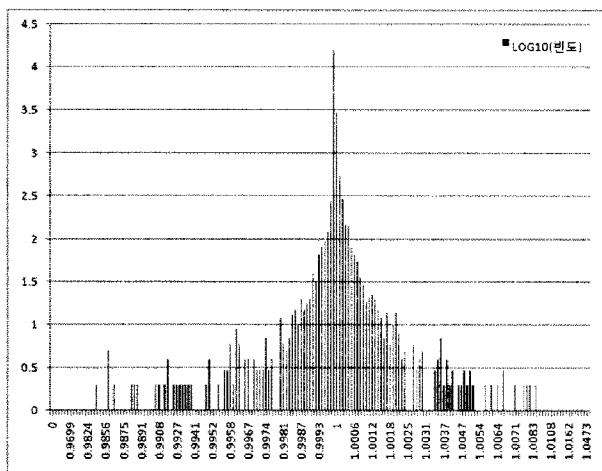


그림 5 집중도에 의한 단어의 분포(G06F)

표 12 집중도 적용에 따른 단어 및 PMI 쌍의 개수 변화
(H04N)

	단어 개수	PMI 쌍 개수
집중도 > 0	23,160	493,706
집중도 > 1.0001	1,416	96,638

표 13 집중도 적용에 따른 단어 및 PMI 쌍의 개수 변화
(G06F)

	단어 개수	PMI 쌍 개수
집중도 > 0	21,481	402,756
집중도 > 1.0001	1,704	104,272

관성 없이 기술하는 경우는 거의 존재하지 않기 때문이다. 만약 이러한 사례가 전혀 존재하지 않는다면 동시출현빈도가 0인 단어들만 대체어라고 판단하면 되겠지만, 이런 사례가 희소하게 발생하므로 본 연구에서는 신뢰도를 반영하여 필터링하는 방법을 선택하였다.

반대로 단어 쌍 중 하나가 영어일 경우에는 0.5를 설정하여 신뢰도를 상대적으로 약하게 반영하였다. 그 이유는 영어로 표기된 단어가 중심 단어가 되었을 때 필터링 정도를 과하게 설정하면, 대체어 후보 목록 중 한글로 된 상당수의 연관단어들이 높은 대체어 점수를 얻게 되는 부작용이 나타나기 때문이다.

4.3 평가 지표

본 연구에서는 표 1과 같이 정의된 ‘이형어’, ‘대역어’, ‘유의어’ 3가지 유형에 대하여 표 14와 같은 평가지표를 구축하여 평가를 수행하였다.

이와 같은 평가지표를 구축하기 위하여 평가지표 구축 사이트를 구현하여 평가지표 후보 단어에 대한 평가를 진행하였으며, 평가지표 구축 참여자가 기준단어와 기준단어의 대체어 후보 간의 관계를 각각 ‘보름’, ‘무관’, ‘이형어’, ‘유의어’, ‘대역어’ 등으로 판단하여 선택할 수 있도록 구축되었다.

평가지표 구축 사이트에서 연구 참여자들에게 제시한 대체어 후보로 이용한 데이터는 여러 유사 척도들에 의

표 14 대체어 평가지표

기준단어	대체어	유형
텔레비전	텔레비죤	이형어
텔레텍스트	텔리텍스트	이형어
파워	파우어	이형어
CD-ROM	씨디롬	대역어
LCD	엘씨디	대역어
퍼스널컴퓨터	PC	대역어
결제	지불	유의어
곱셈기	승산기	유의어
...

해 생성된 대체어 후보 목록들(Cosine Similarity, Jaccard, Tanimoto) 중, ‘텔레비전(4개)’, ‘액정(3개)’ 등 대체어 출현 사례가 많은 단어들 위주로 선정하였다. 선정된 중심 단어의 개수는 H04N 분류에서 110개, G06F에서는 114개, 계 224개이다.

4.4 성능 평가

제안 방법의 성능을 평가하기 위한 척도로서 정확도(Precision)와 재현율(Recall)을 이용하였다.

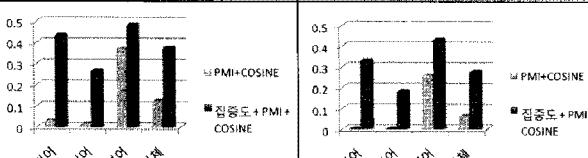
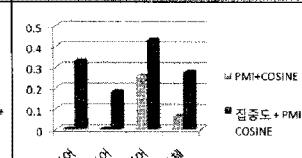
정확도는 (대체어 출현 개수)/(대체어 추출 개수)로 계산하였으며, 대체어 추출 개수는 중심단어가 지니고 있는 대체어 개수에 따라 유동적으로 바꾸어 계산하였다. 예를 들어 ‘텔레비전’의 정확도를 측정하고자 할 경우, 평가지표 상에 ‘텔레비전’의 대체어로서 ‘텔레비전’, ‘텔레비죤’, ‘텔리비전’, ‘티브이’ 등 총 4개의 대체어가 존재하므로 대체어 추출 개수는 4가 된다. 즉, 대체어 후보 목록에서 상위 4개만 추출하여 그 안에 ‘텔레비전’의 대체어가 몇 개나 존재하는지 확인함으로써 그 정확도를 측정하였다.

재현율 역시 (대체어 출현 개수)/(대체어 추출 개수)로 측정하였으나, 대체어 추출 개수를 정확도와는 달리 고정 값을 정한 후, 대체어 후보 목록 상위 5개부터 5개씩 대체어 추출 개수를 늘려가며 최종적으로 대체어 후보 목록 상위 50개까지의 재현율을 측정하였다.

4.4.1 집중도의 성능 평가

표 15는 PMI와 코사인 유사도를 이용하여 추출한 대체어 후보 목록에 집중도를 적용하지 않았을 때의 정확도와 적용한 후의 정확도를 비교한 결과이다. 표 15에서는 집중도를 이용하여 각 IPC의 핵심 키워드만 추출하여 대체어 후보 목록을 생성한 결과가 그렇지 않은 것 보다 정확도에 있어서 H04N분류와 G06F분류에서 각각 약 25% 포인트 및 21% 포인트 정도 향상되는 것으로 나타났다.

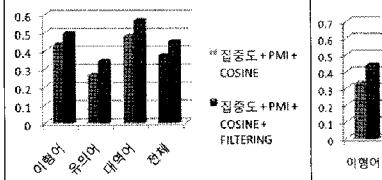
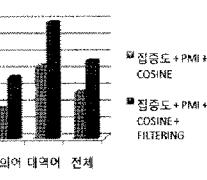
표 15 집중도 반영에 따른 정확도 향상 결과

유형	PMI+Cosine	집중도+PMI+Cosine	유형	PMI+Cosine	집중도+PMI+Cosine
이형어	0.0312	0.4375	이형어	0.0000	0.3333
유의어	0.0192	0.2692	유의어	0.0000	0.1836
대역어	0.3714	0.4857	대역어	0.2608	0.4347
전체	0.1261	0.3782	전체	0.0667	0.2778
					
화상 통신(H04N)			전기에 의한 디지털 데이터 처리(G06F)		

4.4.2 대체어 필터링 기법의 성능 평가

표 16은 앞서 평가한 집중도를 적용한 대체어 후보 목록에 필터링 기법을 적용하여 대체어 후보 목록 내의 연관단어를 필터링 했을 경우의 정확도를 비교한 결과이다. 표 16에 따르면 필터링 기법을 적용하는 것이 그렇지 않은 것보다 정확도에 있어서 H04N분류와 G06F분류에서 각각 약 8% 포인트 및 20% 포인트 정도 향상되는 것으로 나타났다.

표 16 대체어 필터링 기법 적용에 따른 정확도 향상 결과

유형	집중도+PMI+Cosine	집중도+PMI+Cosine+Filtering	유형	집중도+PMI+Cosine	집중도+PMI+Cosine+Filtering
이형어	0.4375	0.5000	이형어	0.3333	0.4444
유의어	0.2692	0.3461	유의어	0.1836	0.3673
대역어	0.4857	0.5714	대역어	0.4347	0.6956
전체	0.3782	0.4538	전체	0.2778	0.4667
					
화상 통신(H04N)			전기에 의한 디지털 데이터 처리(G06F)		

4.4.3 기존 연구와의 성능 비교

표 17은 본 연구에서 제안하는 방법 중 집중도를 이용한 IPC별 중요 단어 선정 단계(1단계)와 연관단어 필터링 단계(4단계)를 적용하지 않은 상태에서 CWO(Context-window Overlapping)[9]와의 정확도를 비교한 결과이다. 표 17에 따르면 CWO가 본 연구에서 제안하는 방법보다 높은 정확도를 지니는 것으로 나타났다.

표 17 집중도 적용 이전의 정확도 비교 결과

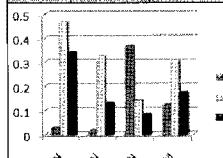
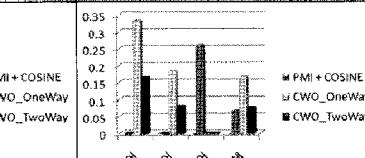
유형	PMI+Cosine	CWO_OneWay	CWO_TwoWay	유형	PMI+Cosine	CWO_OneWay	CWO_TwoWay
이형어	0.0312	0.4687	0.3437	이형어	0.0000	0.3333	0.1666
유의어	0.0192	0.3269	0.1346	유의어	0.0000	0.1836	0.0816
대역어	0.3714	0.1428	0.0857	대역어	0.2608	0.0000	0.0000
전체	0.1261	0.3109	0.1765	전체	0.0667	0.1667	0.0778
							
화상 통신(H04N)				전기에 의한 디지털 데이터 처리(G06F)			

표 18 집중도 적용 이후의 정확도 비교 결과

유형	집중도+ PMI+ Cosine	CWO_ One Way	CWO_ Two Way	유형	집중도+ PMI+ Cosine	CWO_ One Way	CWO_ Two Way
이형어	0.4375	0.4687	0.3437	이형어	0.3333	0.3333	0.1666
유의어	0.2692	0.3269	0.1346	유의어	0.1836	0.1836	0.0816
대역어	0.4857	0.1428	0.0857	대역어	0.4347	0.0000	0.0000
전체	0.3782	0.3109	0.1765	전체	0.2778	0.1667	0.0778

화상 통신(H04N)

전기에 의한 디지털 데이터 처리(G06F)

위 실험은 본 연구에서 제안한 집중도 및 필터링을 적용하지 않고 순수한 대체어 추출 성능만을 비교하기 위한 실험이었으나 CWO는 Context-window라는 개념에 의해 이미 전처리가 수행된 결과이기 때문에 본 연구의 결과물에도 집중도를 이용하여 전처리를 수행한 후 재실험을 하였다. 그 결과 표 18과 같이 전체적인 정확도에 있어서 본 연구에서 제안하는 알고리즘이 CWO의 성능보다 H04N분류와 G06F분류에서 각각 약 6% 포인트 및 11% 포인트 정도 더 우수한 것으로 나타났다.

마지막으로 제안하는 방법의 마지막 단계인 필터링 단계를 추가로 적용한 후, CWO와의 정확도를 비교해보았다. 그 결과 역시 표 19와 같이 CWO보다 H04N분류와 G06F분류에서 각각 약 14% 포인트 및 30% 포인트 정도 더 우수한 것으로 나타났다.

표 19 필터링 적용 이후의 정확도 비교 결과

유형	집중도+ PMI+ Cosine+ Filtering	CWO_ One Way	CWO_ Two Way	유형	집중도+ PMI+ Cosine+ Filtering	CWO_ One Way	CWO_ Two Way
이형어	0.5000	0.4687	0.3437	이형어	0.4444	0.3333	0.1666
유의어	0.3461	0.3269	0.1346	유의어	0.3673	0.1836	0.0816
대역어	0.5714	0.1428	0.0857	대역어	0.6956	0.0000	0.0000
전체	0.4538	0.3109	0.1765	전체	0.4667	0.1667	0.0778

화상 통신(H04N)

전기에 의한 디지털 데이터 처리(G06F)

4.4.4 제목 및 초록 이용 시 성능 비교

특히 문헌의 구조를 살펴보면 제목(발명의 명칭)에는 해당 기술과 그 용도가 핵심 키워드로 간략하게 표기되며, 초록은 핵심 키워드와 이를 설명하기 위한 부가적인 단어들이 포함된다.

이러한 특허 문헌의 특징을 이용하면 초록을 이용할 때보다 문서 분류 내에서 중요 단어를 더 잘 찾아낼 수 있으며, 결과적으로 대체어 추출 성능을 더 향상시킬 수 있다. 실제 실험 결과(표 20)에서도 제목을 이용하는 것이 초록을 이용하는 것보다 더 좋은 성능을 보이는 것으로 나타났다. 이러한 이유로 본 연구에서는 특허 문헌의 제목만 이용하여 실험 및 평가를 수행하였다.

표 20 제목 및 초록을 이용한 대체어 추출 정확도 비교 결과

유형	집중도+ PMI+ Cosine (제목)	집중도+ PMI+ Cosine+ Filtering (제목)	집중도+ PMI+ Cosine (초록)	집중도+ PMI+ Cosine+ Filtering (초록)	유형	집중도+ PMI+ Cosine (제목)	집중도+ PMI+ Cosine+ Filtering (제목)	집중도+ PMI+ Cosine (초록)
	집중도+ PMI+ COSINE (제목)	집중도+ PMI+ COSINE+ FILTERING (제목)	집중도+ PMI+ COSINE (초록)	집중도+ PMI+ COSINE+ FILTERING (초록)		집중도+ PMI+ COSINE (제목)	집중도+ PMI+ COSINE+ FILTERING (제목)	집중도+ PMI+ COSINE (초록)
이형어	0.4375	0.5000	0.1250	0.0937	이형어	0.3333	0.4444	0.0000
유의어	0.2692	0.3461	0.1730	0.1153	유의어	0.1836	0.3673	0.0816
대역어	0.4857	0.5714	0.2857	0.4000	대역어	0.4347	0.6956	0.3043
전체	0.3782	0.4538	0.1933	0.1933	전체	0.2778	0.4667	0.1222

화상 통신(H04N)

전기에 의한 디지털 데이터 처리(G06F)

4.4.5 대체어 추출 개수 변화에 따른 재현율

표 21은 대체어 유형 구별 없이 전체적인 재현율의 변화를 측정한 결과이다. 결과에 따르면 본 연구에서 최종적으로 제안하는 알고리즘인 ‘집중도 + PMI + Cosine + Filtering’이 가장 우수한 성능을 보이는 것으로 나타

났다. 이러한 현상은 이형어(표 22), 유의어(표 23), 대역어(표 24)에 있어서도 비슷하게 나타났다.

5. 결론 및 향후연구

본 논문에서는 기존의 검색 대체어 발견 연구들을 바

표 21 전체 대체어 재현율 비교 결과

IPC	대체어 추출 개수 대체어 추출 방법										
		5	10	15	20	25	30	35	40	45	50
H04N	집중도 + PMI + Cosine	0.6302	0.7647	0.8067	0.8319	0.8487	0.8655	0.8655	0.8655	0.8739	0.8739
	집중도 + PMI + Cosine + Filtering	0.7058	0.8235	0.8487	0.8739	0.8823	0.9075	0.9159	0.9243	0.9243	0.9327
	CWO_ONE WAY	0.4033	0.4705	0.5042	0.5378	0.563	0.5714	0.5882	0.605	0.6134	0.6218
	CWO_TWO WAY	0.2941	0.4285	0.4957	0.521	0.5546	0.5798	0.5798	0.5966	0.605	0.6134
G06F	집중도 + PMI + Cosine	0.5666	0.6333	0.7333	0.7555	0.8000	0.8333	0.8666	0.8777	0.8777	0.8777
	집중도 + PMI + Cosine + Filtering	0.7000	0.8111	0.8666	0.9111	0.9333	0.9333	0.9333	0.9333	0.9333	0.9333
	CWO_ONE WAY	0.2666	0.2888	0.3333	0.3555	0.3555	0.3888	0.3888	0.4111	0.4222	0.4222
	CWO_TWO WAY	0.2111	0.300	0.3222	0.3666	0.3888	0.3888	0.4000	0.4000	0.4222	0.4444
화상 통신(H04N)						전기에 의한 디지털 데이터 처리(G06F)					

표 22 이형어 유형의 대체어 재현율 비교 결과

IPC	대체어 추출 개수 대체어 추출 방법										
		5	10	15	20	25	30	35	40	45	50
H04N	집중도 + PMI + Cosine	0.5000	0.7500	0.7812	0.7812	0.7812	0.7812	0.7812	0.7812	0.7812	0.7812
	집중도 + PMI + Cosine + Filtering	0.6562	0.7812	0.7812	0.7812	0.7812	0.8125	0.8125	0.8125	0.8125	0.8437
	CWO_ONE WAY	0.5937	0.6562	0.6562	0.6562	0.6875	0.7187	0.7187	0.7187	0.7500	0.7500
	CWO_TWO WAY	0.375	0.6250	0.6562	0.6875	0.6875	0.6875	0.6875	0.7187	0.7187	0.7500
G06F	집중도 + PMI + Cosine	0.6666	0.7222	0.7777	0.7777	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333
	집중도 + PMI + Cosine + Filtering	0.7777	0.7777	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333
	CWO_ONE WAY	0.4444	0.4444	0.5000	0.5000	0.5555	0.5555	0.5555	0.6111	0.6111	0.6111
	CWO_TWO WAY	0.3888	0.4444	0.4444	0.4444	0.5555	0.5555	0.5555	0.5555	0.6111	0.6111
화상 통신(H04N)						전기에 의한 디지털 데이터 처리(G06F)					

표 23 유의어 유형의 대체어 재현율 비교 결과

IPC	대체어 추출 개수 대체어 추출 방법										
		5	10	15	20	25	30	35	40	45	50
H04N	집중도 + PMI + Cosine	0.6153	0.7500	0.8076	0.8269	0.8461	0.8846	0.8846	0.8846	0.9038	0.9038
	집중도 + PMI + Cosine + Filtering	0.7115	0.7884	0.8269	0.8461	0.8653	0.9038	0.923	0.9423	0.9423	0.9423
	CWO_ONE WAY	0.4423	0.5384	0.5769	0.5961	0.6346	0.6538	0.6923	0.7307	0.7307	0.7307
	CWO_TWO WAY	0.3653	0.4615	0.5576	0.5961	0.6346	0.6923	0.6923	0.7115	0.7307	0.7307
G06F	집중도 + PMI + Cosine	0.5306	0.6122	0.7346	0.7755	0.8163	0.8775	0.9183	0.9387	0.9387	0.9387
	집중도 + PMI + Cosine + Filtering	0.6122	0.7959	0.8571	0.8979	0.9387	0.9387	0.9387	0.9387	0.9387	0.9387
	CWO_ONE WAY	0.3265	0.3469	0.4285	0.4489	0.4489	0.4897	0.4897	0.5102	0.5102	0.5306
	CWO_TWO WAY	0.2653	0.3877	0.4285	0.4693	0.4693	0.4693	0.4897	0.5306	0.5306	0.5714
화상 통신(H04N)						전기에 의한 디지털 데이터 처리(G06F)					

표 24 대역어 유형의 대체어 재현율 비교 결과

IPC	대체어 추출 개수 대체어 추출 방법										
		5	10	15	20	25	30	35	40	45	50
H04N	집중도 + PMI + Cosine	0.7714	0.8000	0.8285	0.8857	0.9142	0.9142	0.9142	0.9142	0.9142	0.9142
	집중도 + PMI + Cosine + Filtering	0.7428	0.9142	0.9428	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	CWO_ONE WAY	0.1714	0.200	0.2571	0.3142	0.3142	0.3142	0.3142	0.3142	0.3142	0.3142
	CWO_TWO WAY	0.0857	0.1714	0.2571	0.3142	0.3142	0.3142	0.3142	0.3142	0.3142	0.3142
G06F	집중도 + PMI + Cosine	0.5652	0.6086	0.6956	0.6956	0.7391	0.7391	0.7826	0.7826	0.7826	0.7826
	집중도 + PMI + Cosine + Filtering	0.826	0.8695	0.913	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	CWO_ONE WAY	0.0000	0.0000	0.0434	0.0434	0.0434	0.0434	0.0434	0.0434	0.0434	0.0434
	CWO_TWO WAY	0.0000	0.0434	0.0434	0.0434	0.0434	0.0434	0.0434	0.0434	0.0434	0.0434
화상 통신(H04N)						전기에 의한 디지털 데이터 처리(G06F)					

교분석하고, 이를 한글 특허문헌뭉치에 적용하는데 있어 문제가 되는 점들을 보완하여 한글 특허문헌뭉치 속에서 대체어 후보를 추출하는데 적합한 방법을 제안하였다.

제안 방법은 “문장 내에서 함께 쓰이는 단어들이 비슷한 두 단어는 서로 비슷한 의미를 지닐 것이다”라는

기존 대체어 발견 연구들과 동일한 가설로 접근하였으나, 특허 문헌의 IPC 분류를 이용하여 각 분류별 중요 단어를 선정하였다는 점과 PMI로 계산된 연관단어 뭉치(특징 벡터)를 코사인 유사도를 이용하여 비교한 후, 신뢰도를 이용한 필터링 기법을 적용하여 대체어 후보

목록의 품질을 향상시켰다는 점에 있어서 기존 연구들과 차별된다.

평가 결과, 집중도와 펠터링 기법은 정확도에 있어서 기본 대체어 추출 모델(PMI + Cosine)보다 각각 평균적으로 약 23% 포인트 및 14% 포인트 정도 성능이 향상된 것으로 나타났으며, 추출 개수에 따른 재현율 변화 그래프에서도 제안 방법이 전체적으로 상위 곡선을 유지하는 것으로 나타났다. 또한 기존의 대체어 발견 연구인 Context-window Overlapping 기법과 비교한 결과 역시 본 연구에서 제안한 알고리즘이 정확도에 있어서는 평균적으로 약 22% 정도 더 우수한 것으로 나타났으며, 재현율에 있어서는 최소 30% 포인트에서 최대 50% 포인트 정도 더 우수한 것으로 나타났다.

본 연구에서 제안된 방법의 평가는 한글이라는 언어와 특허문헌이라는 도메인에서 제한되어 수행되었으나, 향후에는 타 언어 및 일반적인 정보검색 도메인으로 확장하여 그 적용 가능성을 검증해 볼 필요가 있다. 또한, 본 연구의 확장을 위해서는 '상/하위어' 등의 관계를 포함하는 온톨로지를 활용하여 검색의 정확도를 더욱 향상시킬 수 있는 방법에 대한 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), pp.885-902, 1992.
- [5] Magnus Sahlgren, "The Word-Space Model," Ph.D. Dissertation, Stockholm University, Stockholm, Sweden, 2006.
- [6] Patrick Pantel and Dekang Lin, Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp.613-619, Edmonton, Canada, 2002.
- [7] P. D. Turney, Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning*, 2001.
- [8] Islam, A. and Inkpen, D., "Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words," In *Proceedings of the International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.
- [9] Ruiz-Casado, M. and Alfonseca, E. and Castells, P., "Using context-window overlapping in synonym

discovery and ontology extension," In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP-2005, 2005.

- [10] 이성진, "키워드 셋에서의 상품 추천을 위한 연관 키워드 그룹 추출 기법", M.S. Thesis, Soongsil University, Seoul, Korea 2003.
- [11] J. Baik and S. Kim and S. Lee, "Extracting Alternative Word Candidates for Patent Information Search," *Journal of KIISE : Computing Practices and Letters*, vol.15, no.4, pp.299-303, Apr. 2009. (in Korean)



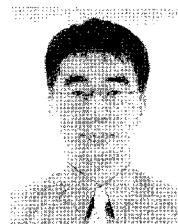
백종범

2006년 학점은행제(한국IT전문학교) 전자계산학 학사. 2009년 숭실대학교 컴퓨터학과 석사. 2009년~현재 숭실대학교 컴퓨터학과 박사과정. 관심분야는 데이터마이닝, 정보검색, 패턴인식, 인공지능



김성민

2004년 나사렛대학교 전산정보학과 학사 2006년 숭실대학교 컴퓨터학과 석사. 2006년~현재 숭실대학교 컴퓨터학과 박사과정. 관심분야는 텍스트마이닝, 시멘틱웹, 자연어처리, 인공지능



이수원

1982년 서울대학교 자연과학대학 계산통계학과 학사. 1984년 한국과학기술원 전산학과 석사. 1994년 University of Southern California 전산학과 박사. 1995년~현재 숭실대학교 컴퓨터학부 교수 관심분야는 데이터마이닝, 정보검색, 인공지능