특허 문헌 검색에서 복합명사 가중치 부여 방법

손기준[O] 이상조
경북대학교 언어정보 연구실
kjson[O]@sejong.knu.ac.kr sjlee@bh.knu.ac.kr

Weighting Methods for Compound Nouns in Patent Retrieval System

Kijun Son[0] Sangjo Lee
Dept. of Computer Engineering, Kyungpook National University

요 약

문서 검색 시스템에서 특정 주제에 관한 문서를 검색하기 위한 색인어의 가중치 부여 방법으로 단순빈도와 역문헌빈도에 의한 가중치 부여 방법을 주로 이용한다. 하지만 빈도 정보만을 이용한 방법은 성능 및 정확도의 향상에 한계가 있다. 이에 본 논문에서는 특허 문헌 검색 시스템의 검색 효율을 높이기 위해 자주 출현하는 복합명사의 재출현 양상과 복합명사의 역할변화에 따른 가중치 부여 방법을 제안한다. 본 연구에서 제안한 가중치 부여 방법을 이용하여 실험한 결과 단순빈도와 역문헌빈도 정보를 이용한 방법보다 더 나은 성능을 보였다.

1. 서 론

문서의 내용을 대표하는 용어를 추출하고 가중치를 부여하는 기법은 정보 검색 시스템에서 색인어를 추출하거나 문서 분류, 문서 요약 시스템을 구현하기 위한 전 단계로서 그 활용 분야가 다양하고 필요성이 증가 되고 있다. 문서의 내용을 분석하여 추출된 용어는 해당 문서의 내용을 대표하고, 문서의 내용을 요약해주는 효과가 있다. 따라서 정보검색 시스템은 사용자의 요구에 만족하는 문서를 선택하고 순위를 결정하기위한 정확한 색인어의 추출과 가중치 부여 방법은 정보검색에서 중요하다.

한국어 명사들은 비교적 자유롭게 결합하여 새로운 복합명사를 만들어낸다. 이러한 한국어의 다양한 언어 종속성의 중요한 요소 중 하나가 복합명사의 처리에 대한 필요성이다[8]. 복합명사로 인해 발생하는 문제 중 하나가 색인어의 가중치 부여 문제이다. 색인어의 가중치는 색인어의 중요도에 따라 차별적으로 부여되며 검색 성능에 직접적인 영향을 준다. 하지만, 지금까지의 연구는 복합명사의 인식과 분해 및 합성에 집중되어 왔다[4,8]. 따라서 새로운 가중치 부여 방법에 대한 연구가 미흡하였으며, 기존의 가중치 부여 방법을 그대로 사용하는 것이 일반적이었다.

복합명사와 복합명사를 이루는 구성명사는 단일명사와는 성질과 역할이 다르므로 각각의 중요도에 따른 새로운 가중치 부여 방법이 필요하다. 복합명사를 색인과정에서 따로 인식을 하지 않을 경우 복합명사가 가지는 어휘적 특성을 무시함으로, 검색시스템의 정확률과 재현율을 낮추는 요인이 된다. 이것은 복합명사가 단일 명사보다 특정성이 커서 문서를 더 잘 표현하며 색인어로서의 가치가 높기 때문이다.

본 논문에서는 복합명사의 재출현 양상과 복합명사 역할 변화를 이용하여 색인어의 가중치를 부여하는 방법을 제시하고자 한다.

2장에서는 복합명사 관련 기존의 연구들을 살펴보고, 3장에서는 특허문헌 검색을 위한 복합명사 가중치 부여기법에 대하여 설명한다. 4장에서는 실험 및 실험결과를 평가하고, 5장에서 결론을 맺는다.

2. 관련연구

복합명사를 다루는 기존의 연구에는 통계적인 기법을 이용한 모델과 자연언어 처리 기법을 이용한 연구로 나누어 볼 수 있다. 일반적인 통계에 기반 한 가중치 부여 방법은 단순빈도(term frequency)와 역문헌 빈도(inverted document frequency)에 의한 가중치 부여 방법을 사용한다. 복합명사의 경우 문서 내에서 단일명사 보다 빈도가 낮기 때문에 단일 명사보다 낮은 가중치를 가진다[6]. 통계적인 기법은 많은 양의 문서를 효율적으로 처리하기 위한 방법이며, 자연언어 처리 기법을 이용한 방법은 문서의 정교한 표현을 통한 성능의 향상을 목표로 하고 있다.

영어권 색인 연구를 살펴보면, [1]은 복잡한 자연언어 처리를 이용한 방법보다 간단한 통계적인 방법이 선호되어야 한다고 주장하고 있으며, [2]는 대규모의 문장을 처리하기 위해 구문분석기 대신 명사구 구문분석기를 이용하여 명사구 색인에 이용하였다.

국내의 연구로 [3]은 자연언어 처리 기법을 사용하지 않고, 명사사전과 복합명사 구성 패턴의 통계적인 정보를 이용하여 복합명사를 합성하였다. [4]는 문장을 의존문법에 기반 한 구문분석 기법을 이용하여 단문으로 분할한 뒤, 5가지의 패턴에 기반 하여 명사를 합성하여 명사구 후보를 생성 하였다. [5]는 복합명사를 가능한 분해 후보를 생성하고, 후보들에 대하여 가중치를 부여함으로써 최적 후보를 선택하는 방법을 이용하여 복합명사를 분할하였다.

복합명사 문제는 일본어나 중국어 등에서도 발생하는 문제로 연구 되어왔다. [9]는 사전과 지식베이스를 이용하였고, [10]는 코퍼스의 통계적 데이터를 이용한 방법이 있다.

3.특허 문헌 검색을 위한 복합명사 가중치 부여 기법

특허 문헌 검색 시스템에서 색인과 검색은 같은 처리 과정을 거치는데 이것은 검색시 색인어와 질의어의 불일치를 피하기 위해서이다. 색인과 검색과정이 분리 되어 있으면 처리 효과가 달라져 검색 성능에 차이가 나게 될 것이다. 일반적으로 복합명사 처리를 크게 4가지 복합명사를 전혀 처리 하지 않는 경우, 복합명사를 분할만 하는 경우, 복합명사를 합성만 하는 경우, 분할과 합성을 동시에 하는 경우로 나누어 볼 수 있다.

따라서 특허문헌 검색 시스템에서 색인과 검색 과정에 행해지는 복합명사의 처리 방법들이 검색 성능에 어떠한 영향을 미치는 지를 검증하기 위해 그림 1과 같은 복합명사 가중치 부여 방법을 사용한다.

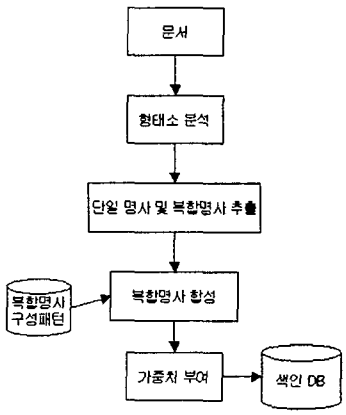

그림 1 복합명사 가중치 부여 방법

복합명사 가중치 부여 방법은 크게 복합명사 분할과 가중치 부여의 단계로 구성된다. 복합명사 분할은 한국어에서 띄어쓰기가 자유로워 발생하는 불일치의 문제를 해결하므로 재현율을 올릴 수 있는 방법으로 다음과 같은 과정을 거친다. 단어가 분할 대상으로 입력되면, 입력된 단어에 복합명사의 구성 패턴을 적용한다. 복합명사의 구성 패턴은 표1과 같다. ' 칼라 스킨층을' ,' 정수용 필터를' 등과 같이 ' n1 n2를' 의 패턴은 복합명사로 간주 될 수 있다. 이때 복합명사를 구성하는 단일 명사의 수가 3이상인 명사를 대상으로 분할을 시도한다. 그리고 분할이 실패한 경우 수동으로 분할을 한다. 이렇게 분할된 명사들은 다시 재합성 과정을 거치고, 분할 및 재합성된 단일 명사와 복합명사는 가중치 부여 과정을 거친다.

표 1 복합명사의 구성 패턴

n1에 n2하여 (예: 에지에 몰딩하여)
n1하여 n2한 (예: 가열하여 성형한)
n1과 n2의 (예: 스트립과 양단의)
n1과 n2로 (예: 백질과 회백질로)
n1및n2 (예: 단성유 및 은제올라이트)
n1 n2를 (예: 아스팔트 패드를)

특허문헌을 검색하기 위해 사용하는 자신의 정보요구를 질의어로 표현한다. 이때 사용자 질의의 예는 다음과 같다.

- (에어백 and 자동차 and 스프링 and 지연)
- (음극선관 and (길이 or 전장 or 안길이) and 단축)
- (전광기기 사용되는 모듈레이터의 제조 방법)

사용자는 자신의 정보요구를 질의어로 표현을 한다. 일반적인 검색에서 사용자는 하나 혹은 두 단어 정도의 질의어를 입력한다. 하지만 특허문헌 검색 시스템에서 사용자는 여러 개의 단일 명사를 조합하여 질의어를 표현한다. 이에 따라 색인어와 질의어의 불일치 문제를 피하기 위하여 검색과 색인 과정에 복합명사처리가 필요하다. 색인과 검색의 처리과정이 다르면 검색의 효율이 낮아져서 검색 성능의 저하를 가져올 수 있다.

웹스 검색 시스템을 이용한 사용자들이 사용한 질의어 중 11,047개의 질의어에 대하여 분석한다. 사용자가 질의한 질의어의 조합한 후 질의어를 구성하는 단일명사의 개수를 계산한 결과는 표 2와 같다. 표 3을 보면 질의어를 구성하는 구성명사의 수가 5개 이상인 것은 출현 횟수가 많이 낮다.

표 2 질의어를 구성하는 단일명사 수

	1	2	3	4	5
단일명사 수	7772	2509	584	137	37
구성비율	68.1%	22.0%	5.1%	1.2%	0.3%

3.1 가중치 부여

문서의 중요도를 나타내는 색인어의 중요도를 나타내는 값으로 사용자에게 질의와 관련이 높은 문서들을 제공하기 위해 각 문서를 대표하는 색인어를 잘 선정하고, 선정된 색인어에 가중치를 중요도에 따라 할당하는 것이 중요하다.

문헌빈도와 역문헌빈도 방법은 문서 내 빈도수만을 강조하여 문서 내 출현 빈도가 낮은 복합명사의 경우 낮은 가중치를 가진다. 복합명사는 개념적 중요도에 따라 단일명사보다 높은 가중치를 갖지만, 너무 과도하게 높은 가중치를 부여하면 오히려 검색의 성능을 저하시킬 수 있다.

가중치의 부여는 선택된 색인어들의 중요도를 결정 짓는 중요한 과정이다. 즉 색인어의 추출 방법만큼이나 색인어의 가중치 부여 방법도 중요하게 다루어져야 하는 부분이다. 이에 본 논문에서는 복합명사의 재출현 양상과 복합명사에서의 역할을 고려하여 가중치 가중치를 부여한다.

복합명사의 재출현 양상은 다시 출현 하는 어휘의 모습이 완전 일치하는 경우와 복합명사에서의 역할 변화로 구분한다. 완전 일치는 재출현 한 어휘가 완전히 일치하는 경우를 의미하며, 복합명사에서의 역할 변화는 재출현 하는 어휘가 복합명사의 중심어에 위치하는지 아니면 수식어로 일치하는 지로 나누어진다. 예를 들면, ' 자동 개폐' 는 자동으로 열리는 것으로 ' 개폐 장치' 는 개폐하는 장치로 파악할 수 있다. 두 경우 ' 개폐' 의 역할이 다르므로 구분을 하여야 한다. 이에 따른 가중치는 아래와 같다.

본 논문에서는 복합명사의 재출현 양상과 복합명사에서의 역할 변화에 따른 가중치 부여 기법을 이용한다. 복합명사와 구성명사의 특성을 고려하여 변형된 식을 사용하였고, 복합명사, 구성명사, 그리고 단일명사의 가중치를 부여하기 위해 다음과 같은 어휘의 역할 변화를 고려하였다.

본 논문에서 제안한 복합명사의 재출현 양상과 복합명사에서의 역할 변화에 따른 가중치는 다음과 같다.

복합명사 -> 복합명사 1.0
복합명사 -> 복합명사의 일부 0.7
복합명사 -> 단일명사 0.4
단일명사 -> 복합명사의 수식어 0.8
단일명사 -> 복합명사의 중심어 0.6
단일명사 -> 복합명사의 부속어 0.5

가중치 부여 방법은 일반적으로 정보검색에서 많이 사용하는 tf * idf 가중치 부여 기법을 사용한다. 복합명사의 복합명사의 재출현 양상과 복합명사 역할 변화에 따라 식 (1)에 의하여 계산되며, 독립적으로 하나의 개념을 나타내는 단일명사의 가중치는 tf*idf 가중치를 이용하여 계산된다.

$$w_{c,j} = (TF * WC_c) \cdot IDF$$
$$(w_{c,j} = f_{i,j} \cdot g_{c,j} \cdot \log \frac{N}{n_c}) \quad (1)$$

TF : **문헌 내에 서의 단어 의미**

WC = **복합명사의 가중치**

IDF : **전체문헌들에서 용어가 출현한 문헌수의 역수**

4. 실험 및 평가

본 논문에서 제안한 복합명사의 재출현 양상과 복합명사 역할 변화에 따른 가중치 부여 방법의 유효성을 검증하기 위해 제안한 가중치 부여 방법을 이용하여 검색 시스템의 성능을 평가한다. 원문의 형태소 분석을 위해서 Ham[7]을 이용한다.

4.1 실험 문서 및 환경

실험에 사용된 실험 데이터는 윕스에서 제공해준 특허 문헌 2556개의 문서를 대상으로 실험을 수행한다. 실험을 위해 벡터 공간 모델을 검색 모델로 한 정보검색 시스템을 사용하였다. 색인어와 가중치는 본 논문에서 제안한 어휘역할 변화에 따른 가중치 부여 방법을 이용하였다. 문헌벡터와 질의 벡터 간의 유사도는 코사인 계수를 사용한다.

실험에 대한 평가 방법은 일반적으로 정보검색 시스템의 평가에 사용되는 정확률을 이용한다. 정확률은 시스템이 부적합한 문헌을 검색하지 않는 능력을 나타낸다.

제안한 방법과 가중치를 부여하지 않은 방법에 대한 유효성을 검증하기 위해 2556개의 특허 문헌을 대상으로 실험을 수행한다. 2556개의 문서에서 복합명사를 추출하고, 추출된 복합명사를 세어보았다. 다음의 표 3은 실험문서에서 나타나는 복합명사의 빈도수 이다.

표 3 복합명사를 구성하는 단일명사 수

단일명사수	2	3	4	5	6	7	8	9
복합명사수	3017	7200	2013	594	320	34	1	1

4.2 복합명사 가중치 부여 방법에 대한 평가

실험 방법은 기존의 tf*idf 가중치 부여 방법을 이용하여 색인어의 가중치를 부여하는 방법과 본 논문에서 제안한 가중치 부여 방법을 이용한 계산 방법과 그 결과를 비교한다. 유사도가 매우 낮은 문서인 경우는 검색에서 제외하는 것이 바람직하므로 0.05이하의 유사도를 가지는 문서는 제외시켰다. 이 실험에서는 유사도 값이 높은 상위 10개의 문서를 대상으로 정확률을 계산한다. 복합명사의 가중치 부여 방법에 따른 성능은 표 4와 같다. 즉 제안한 방법이 tf*idf 가중치 부여방법에 비해4.5%의 성능 개선효과가 있음을 알 수 있다.

표 4 실험결과

tf*idf	제안한 방법
정확률	정확률
40.5%	45%

5. 결론 및 향후과제

본 논문에서는 특허 검색 시스템의 성능향상을 위해 복합명사의 재출현 양상과 복합명사 역할 변화에 따른 가중치 부여 방법을 제안하였다. 제안한 가중치 부여 방법을 이용하여 특허 검색 시스템의 성능을 검증한 결과 제안한 가중치 부여 방법은 복합명사의 어휘 역할 변화를 고려하지 않은 방법보다 더 나은 성능을 보였다.

향후과제로는 검색 시스템의 성능 향상을 위해 문헌집합과 질의의 특성에 따른 용어가중치의 부여 방법과 통합 가중치 부여 방법에 대한 연구가 필요하다.

참고문헌

[1]Gerard Salton, Chris Buckley, " A comparison between statistically and syntactically generated term phrases," Tr89-1027, CS department, Cornell Univ., 1989.

[2] Chengxiang Zhai, " Fast statistical parsing of noun phrases for document indexing," Fifth conference on applied natural language processing, pp.312-319, 1997.

[3] 남세진, 이지연, 신동욱, 채미옥, " 복합명사의 통계적 처리에 대한 평가," 제8회 한글 및 한국어 정보처리 학술발표논문집, pp. 36-41, 1996.

[4] 이현아, " 구문분석과 공기 정보를 이용한 개념 기반 명사구 색인 방법," 포항공대 전산과 석사 학위 논문, 1996.

[5] 강승식, " 한국어 복합명사 분해 알고리즘," 정보과학회논문지(B), 제25권, 제1호, pp.172-182, 1998.

[6] Tomek Strzakowski, " Natural Language Information Retrieval," Information Processing & Management, Vol.31, No3, pp.337-417, 1995.

[7] 강승식, " HAM v.470c: 한국어 형태소 분석기와 한국어 분석 모듈," http://nlp.kookmin.ac.kr/ham/ham.html

[8] 윤보현, 조미정, 임해창, " 통계정보와 선호 규칙을 이용한 한국어 복합명사의 분해," 정보과학회논문지(B), 제 24권, 제8호, pp.900-909 ,1997.

[9] R. Sproat, C. Shih, W. Gale & N. Chang, " A Stochastic Finite-state Word Segmentation Algorithm for Chinese, " Computational Linguistics, Vol. 22930, 1996.

[10] Y. Ogawa & T. Matsuda, " Overlapping Statistical Word Indexing: A New Indexing Method for Jjapaness Text, " Processing of ACM SIGIR, Philadelphia, PA, 1997.

[11] 원형석, 박미화, 이근배, " 복합명사 분할과 명사구 합성을 이용한 통합 색인 기법," 정보과학회논문지(B), 제27권, 제1호,pp.84-95, 2000.

[12] Spark Johnes, " Indexing Term Weighting," Information Storage and Retrieval, Vol. 9, no. 11, pp.619-633, 1973.