조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법

강승식°, †이하규, ‡손소현, ‡홍기채, ‡문병주
국민대학교 컴퓨터학부, †성공회대학교 컴퓨터정보공학부,
‡한국전자통신연구원 지식정보센터
sskang@kookmin.ac.kr, hglee@mail.skhu.ac.kr, {shson, gchong, bjmoon}@etri.re.kr

Term Weighting Method by Postposition and Compound Noun Recognition

Seung-Shik Kang°, Hagyu Lee, So-Hyun Son, Gi-Choi Hong, Byung-Joo Moon
School of Computer Science, Kookmin University
Division of Computer and Information Science, Sungkonghoe University
Knowledge Information Center, ETRI

요 약

문서의 내용을 대표하는 용어를 추출하기 위해 일반적으로 영어에서는 명사구를 색인하는 기법을 사용하지만, 주제어 추출의 관점에서 영어의 명사구가 한국어의 복합명사에 해당하기 때문에 한국어에서는 복합명사 색인 기법을 중요시하고 있다. 본 논문에서는 한글 문서에서 추출된 용어의 가중치를 결정하기 위하여 경험적인 방법에 따라 가중치를 계산하는 방법을 제안한다. 구체적인 가중치 계산 방법으로 용어 자체의 특성에 의한 가중치를 부여한 후에, 복합명사의 경계를 인식하여 띄어쓴 복합명사의 가중치를 조절하고, 다시 용어의 조사 유형에 따라 가중치를 재계산하는 방법을 제안한다. 신문기사에 대한 실험결과에 의하면 제안한 방법이 단순 출현빈도에 의한 주제어 추출 기법보다 정확도가 더 높았다.

1. 서론

문서의 내용을 대표하는 용어들을 추출하는 기법은 정보검색 시스템에서 색인어(index term)를 추출하거나 문서 분류, 문서 클러스터링, 문서 요약 시스템을 구현하기 위한 전단계로서 그 활용 분야가 매우 증가되고 있다. 문서 내용을 분석하여 추출된 용어들은 그 문서의 내용을 기술하는 디스크립터(descriptor) 역할을 하므로 문서 내용을 요약해 주는 효과가 있으며, 또한 그 문서를 특징지어 다른 문서들과 구별할 수 있게 해 준다. 따라서 문서 내용과 관련이 적은 불용어를 제거하고 있으며, 문서의 특성이 반영된 용어, 주제와 연관된 용어, 특정 정보를 담고 있는 용어들은 중요한 용어로 간주된다[1,2].

정보검색 시스템에서는 특정 주제에 관한 문서를 검색하기 위해 색인어를 단순빈도(term frequency)와 역문헌 빈도(inverse document frequency)에 의해 가중치를 계산한다. 단순빈도와 역문헌 빈도는 결과적으로 어떤 용어가 문서 내용을 대표하는 정도를 수치화하기 위한 것이다. 이와 더불어 제목, 본문, 요약 등 용어의 출현 위치와 HTML/XML 태그 등 구조 정보를 이용하기도 하지만 빈도 정보에 의한 방법은 성능 및 정확도를 향상시키는데 한계가 있다.

이러한 한계를 극복하기 위해 영어에서는 용어 가중치(term weight)를 계산하는 방법이 연구되었으며[3,4,5], 색인어를 '단어' 수준에서 '명사구(noun phrase)' 수준으로 발전시킨 명사구 색인 기법이 사용되기도 한다[6,7]. 그런데 영어의 명사구 색인에 대응되는 한국어 색인 기법은 복합어 색인으로서 문제해결 방식에 차이가 있다[8]. 즉, 영어의 경우 'artificial intelligence', 'named entity'와 같이 복합어가 N + N이 아닌 경우가 많을 뿐만 아니라 'automatic indexing'의 예처럼 명사 이외의 품사들로 구성되기도 한다. 이러한 특성 때문에 명사구 색인 기법이 사용된다.

이에 비해 한국어의 복합어는 거의 대부분이 명사로만 구성되고 있으며 단지 '학생의 날', '한글과 컴퓨터' 등 일부 예외가 있을 뿐이다. 이러한 복합어 구성에 관한 특성은 영어 정보검색에서 명사구 인식 기법이 중요한 관심사가 되어 왔던 것에 비해, 한국어에서는 복합명사 분해-결합 문제가 중요한 문제로 등장했던 이유와 밀접한 관련이 있다.

본 논문에서는 한글문서의 내용을 대표하는 주제어를 선별하는 방법의 일환으로 추출된 용어의 가중치를 계산하는 방법을 제시한다.

2. 주제어 추출 기법

한글문서에서 주제어는 문서의 내용을 대표하는 용어로서 주제어의 추출 방법은 문서의 유형에 따라 달라질 수 있다. 주로 주제어 추출의 대상이 되는 문서의 종류 및 그 특성을 살펴 보면 다음과 같다.

- 제목이 있고 두괄식인 문서 -- 신문기사 등
- 제목, 요약 등 구조화된 문서 -- 학술논문 등
- 태그가 있는 정형화된 문서 -- 웹 문서 등
- 기타 -- 시간표, 주소록 등 항목 나열식 문서

학술논문과 같이 '제목/요약/서론/결론'으로 구성된 문서는 제목과 요약에 출현한 용어가 주제어일 가능성이 높고, 신문기사와 같이 주요 내용을 본문의 앞 부분에 기술하는 문서는 제목 및 본문의 앞 문장에 출현한 용어가 주제어일 가능성이 매우 높다. 웹 문서와 같이 태그가 있는 문서는 주제어가 출현하는 필드(field) 정보를 활용하여 해당 용어들에 대한 가중치를 부여할 수 있다. 제목이나 요약 등 주제어 출현위치에

관한 단서를 찾기 어려운 일반적인 문서에 대한 주제어의 추출은 용어의 출현빈도에 의존하고 있다. 특히, 시간표와 주소록 등 명사 및 명사구가 나열되거나 개조식으로 구성된 문서에서는 주제어와 비주제어를 구분하기가 쉽지 않다.

문서 유형과 무관하게 일반적인 문서에 대해 주제어를 추출하는데 사용될 수 있는 정보로는 추출된 용어 자체의 특성(품사 정보, 격 정보 등) 혹은 문장내에서 용어의 구문론적 기능(복합어의 일부, 주절 혹은 종속절에 출현), 용어가 출현된 문장의 기능 등이 있으며 구체적인 예는 다음과 같다.

(1) 어절 단위 -- 용어의 특성 정보
- 복합명사, 미등록어
- 1음절명사, 보통명사
- 명사의 길이(음절수)
- '지금/현재/작년' 등 시간성 명사
- 조사 유형 : '은/는/이/가/을/를/의/만/도/에' 등

(2) 문장 단위 -- 용어의 구문론적 기능
- 복합어(명사구 등) 구성 여부
- 주절 혹은 종속절의 주어/목적어/보어/관형어 등

(3) 문서 단위 -- 용어가 출현한 문장의 특성
- 문장의 위치 : 제목, 앞 부분, 뒷 부분, 중간 부분
- 문장의 중요도
  접속부사 등 수사 어구에 의한 문장의 중요도

(4) 기타
- 용어의 출현빈도
- Coreference 관계에 의한 용어의 중요도

문서에서 추출된 용어의 중요도는 '용어가 추출된 어절의 특성', '용어가 출현한 구/절 특성', '용어가 출현한 문장 특성'에 의해 계산된다. 어절단위의 '용어 특성'은 복합명사/미등록어/보통명사/시간성명사 등 용어의 특성에 따라 가중치를 부여하며, 격조사에 따라 가중치를 다르게 부여한다.

문장 단위의 '구/절 특성'은 구문분석에 의해 주절과 종속절이 구분되는 경우에 주절 혹은 종속절에 출현한 용어의 가중치를 조절할 수 있다. 문서 단위의 '문장 특성'은 접속부사 등 수사 어구에 따라 주요 문장인지, 보조 설명 문장인지를 판단할 수 있으며, 주요 문장을 추출해 주는 문서요약 시스템에서 계산된 문장의 중요도 정보를 활용할 수 있다.

일반적으로 문서 내용을 대표하는 용어는 출현빈도가 높기 때문에 정보검색 시스템에서는 용어의 가중치를 출현빈도와 문헌빈도에 의해 결정하고 있다. 그런데 단순히 용어의 출현빈도만으로 주제어를 판별할 경우 일상적인 표현에서 자주 사용되는 명사가 주제어로 추출되는 오류가 발생하게 된다. 특히, 문서에서 주제어는 유사한 표현이나 관련 어휘들이 출현하는 용어이다. 즉, 동일한 개념이나 유사한 개념, 혹은 관련 어휘들이 출현하는 용어가 문서 내용을 대표하는 용어일 가능성이 매우 높다.

고유명사가 주제어인 경우는 이를 지칭하는 대명사가 반복되기도 하며, 기타 명칭(기관명, 지명, 제품명 등)인 경우에도 대용어가 사용된다. 따라서 대명사 혹은 대용어 관계가 파악되면 대명사가 지칭하는 용어의 가중치를 높일 수 있다. 용어들 간의 Coreference 관계는 명칭(named entity)의 대용어 관계로서 용어간의 유사도 및 용어의 가중치를 계산하는데 활용될 수 있다.

3. 경험적 방법에 의한 가중치 부여

문서에서 주제어를 추출하려면 문서의 내용을 분석하여 어떤 주제에 관한 문서인지를 파악해야 한다. 그러기 위해서는 중요 문장을 선택하거나 수식어 및 종속절을 제거하기 위한 자연언어 분석 기법으로 구문분석 및 의미분석이 선행되어야 한다. 그런데 한국어의 구문분석 및 의미분석 기술은 정확도가 높지 않으므로 본 논문에서는 형태소 분석 결과를 기반으로 명사구와 조사 정보를 이용하여 용어의 가중치를 계산하는 방법을 취한다.

용어의 유형 및 조사 정보에 의해 '어절 단위' 가중치와 '문장 단위' 가중치를 부여하는 과정은 그림 1과 같다.

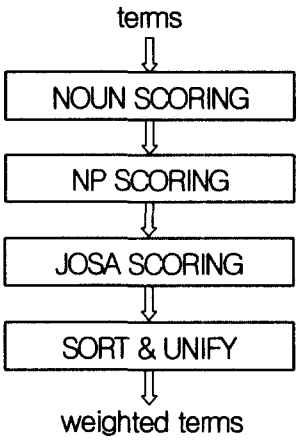

그림 1. 가중치 부여 과정

어절 단위의 '용어 유형 가중치'는 추출된 용어 유형에 따라 각 용어의 가중치를 초기화하는 단계이다. 복합명사의 가중치를 1로 하고 이를 기준으로 미등록어, 보통명사 등의 가중치를 부여하는데 각 유형에 대한 가중치는 한글 문서에 복합명사가 다수 포함되어 있을 경우에 주로 복합명사와 미등록 명사인 고유명사가 주제어일 가능성이 높기 때문이다. 즉, 아래 각 용어의 유형에 대한 가중치는 경험적인 방법으로 결정한 것이다. 여기서 보통명사는 사전에 수록된 명사 중에서 1음절 명사와 시간성 명사를 제외한 명사를 의미하며, 보통명사는 음절수에 따라 가중치를 달리하여 4음절 이상인 것은 0.8, 3음절 명사는 0.45를 부여하였다.

- 복합명사 : 1.0
- 미등록어, 영문자 포함 용어 : 0.8
- 보통명사 : 0.3
- 1음절 명사 : 0.1
- 시간성 명사 : 0.1
- 숫자가 포함된 용어 : 0.2

'단어 유형 가중치'에서는 복합명사와 미등록의 가중치를 높이고 보통명사의 가중치를 낮게 하였다. 그런데 '정보 검색 시스템'의 경우 '정보검색시스템'과 동일한 의미를 갖게 되지만 '정보', '검색', '시스템'에 대한 가중치가 낮으므로 띄어쓴 복합명사를 구성하는 단위명사들은 가중치를 높여 주어야 한다. 띄어쓴 복합명사는 주로 조사없는 명사들로 구성되는 N + N + ... + N 유형이다.

그런데 '학생의 날', '정보의 검색' 등과 같이 'N의 N' 패턴은 복합어로 간주될 수 있으므로 이 유형에 대해서도 "띄어쓴 복합명사 가중치"를 적용한다. "띄어쓴 복합명사 가중치"는 각 단위명사들에 대해 원래 가중치에 음절수별 가중치를 곱하여 복합명사 가중치인 1.0에 근접한 값으로 조절한다. 예를 들어, 2음절 보통명사는 3.0을 곱하고 미등록어는 1.3을 곱해 준다.

복합명사에 대한 가중치가 부여된 후에는 조사의 역할에 따라 용어의 가중치에 조사 가중치를 곱하여 가중치를 수정한다. 조사의 유형에 따른 가중치도 '단어 유형 가중치'와 마찬가지로 직관적으로 부여하였다.

- '은/는' : 1.0
- '이/가' : 0.8
- '을/를' : 0.8
- '만/도' : 0.5
- '의' : 0.8
- '에/에서' : 0.4
- 기타 조사 : 0.3

이 때, 띄어쓴 복합명사는 각 단위명사들에 대해 동일한 조사 가중치를 적용한다. 즉, "정보 검색 시스템을"의 예에서 '정보', '검색', '시스템'에 대해 모두 0.8을 곱하여 가중치를 조절한다.

4. 실험 및 평가

특정 문서로부터 어떤 용어가 주제어인지, 어떤 용어를 색인해야 하는지를 판단하기는 쉽지 않다. 초기의 정보검색 시스템들은 주제어 색인 방식을 취했기 때문에 불용어를 제외시켰다. 그런데 "to be or not to be"와 같이 불용어들로 구성되는 질의어가 가능하고, 때로는 불용어가 정보를 검색하는데 보조 역할을 하는 경우가 있기 때문에 최근에는 불용어를 제거하지 않는 추세이다. 또한, 문서에서 어떤 용어가 주제어인지 판단하기가 쉽지 않아서 사람마다 추출된 용어의 차이가 크다.

제안한 방법의 효용성을 검증하기 위해 색인어 가중치 부여 실험을 하였다. 실험 문서는 정보통신 관련 신문기사 20개이다. 주제어 정답(answer set)을 정의하기가 어렵기 때문에 사람 2명과 가중치 기법, 고빈도어 추출법으로 각 문서에서 주제어 10개씩을 추출하여 "공통으로 추출된 용어수"를 계산하였다. 이 때 가중치 또는 빈도가 동일한 용어가 2개 이상일 때는 출현위치가 앞인 것을 우선으로 하였다.

표 1의 실험 결과로 주제어 일치율은 수동 추출의 일치율 53.5%, 가중치 기법과 수동추출의 일치율 44.3%, 고빈도어와 수동추출의 일치율 38.5%이다. 즉, 제안된 방법은 단순 출현빈도에 의한 주제어 추출 기법에 비해 5.8% 정도 성능 개선의 효과가 있음을 알 수 있다.

표 1. 주제어 추출 비교 실험

H1-H2	가중치 부여 기법		빈도에 의한 추출	
	K-H1	K-H2	F-H1	F-H2
53.5%	43.0%	45.5%	36.0%	41.0%
	평균 44.3%		평균 38.5%	

H1: 사람1, H2: 사람2, K: 가중치, F: 고빈도어

5. 결론

한글 문서에서 문서의 내용을 대표하는 주제어들을 추출하기 위하여 어절 단위의 '단어 유형 가중치'와 '조사 유형 가중치', 그리고 띄어쓴 복합명사를 인식하여 복합명사의 가중치를 부여하는 방법을 제안하였다. 제안한 방법의 효용성을 실험하기 위하여 신문기사 20개에 대해 주제어 추출 실험을 하였다. 그 결과 제안한 방법은 수동으로 추출한 결과보다는 정확도가 낮지만, 단순 출현빈도에 의한 주제어 추출에 비해 정확도가 높았다.

참고문헌

[1] 정영미, 정보검색론, 구미무역(주), 1993.

[2] Luhn, H. P. "A Statistical Approach to Mechanized Encoding and Searching of Library Information", IBM Journal of Research and Development, vol. 1, no. 4, pp.309-317, 1957.

[3] Salton, G., "Recent Trends in Automatic Information Retrieval", Proceedings of 1986 ACM Conference on Research and Development in Information Retrieval, pp.1-10, 1986.

[4] Salton, G. and C. Buckley, "The Term-Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, vol. 24, no. 5, pp.513-523, 1988.

[5] Sparck Johnes, K., "Indexing Term Weighting", Information Storage and Retrieval, vol. 9, no. 11, p.619-633, 1973.

[6] Fagan, J. L., "The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval," Journal of the American Society for Information Science, vol. 40, no. 2, pp.115-132, 1989.

[7] Jones, L. P., E. W. Gassie, and S. Radhakrishnan, "INDEX: the Statistical Basis for an Automatic Conceptual Phrase-Indexing System", Journal of the American Society for Information Science, vol. 41, no. 2, pp.87-97, 1990.

[8] 서은경, "구문-통계적 기법을 이용한 한국어 자동색인에 관한 연구", 정보관리학회지, 10권 1호, pp.97-124, 1993.