한글 문서를 위한 효과적인 색인 방법

An Effective Indexing Method for Hangul Texts

이준호*○ 박혁로* 박현주* 안정수** 김명호**
*한국과학기술연구원 연구개발정보센터
**한국과학기술원 전산학과

J.H. Lee*○ H.R. Park* H.J. Park* J.S. Ahn** M.H. Kim**
*Korea Research and Development Information Center, KIST
**Department of Computer Science, KAIST

요 약

기존의 한글 자동 색인 방법들은 어절 단위 색인법과 형태소 단위 색인법으로 분류될 수 있다. 전자는 문서내의 어절에서 색인어의 부분으로서 가치가 없는 음절들을 제거함으로써 색인어를 추출하는 방법으로, 문서에 복합 명사들이 많이 포함되어 있을 경우 검색효과가 저하되는 문제점을 지니고 있다. 후자는 형태소 해석이나 구문 해석을 이용하여 중요한 의미를 갖는 명사나 명사구를 추출하는 방법으로, 단일 명사를 추출함으로써 복합 명사의 띄어 쓰기 문제를 극복할 수 있다. 그러나, 색인 과정에서 요구되는 많은 언어 정보를 개발하고 유지 보수해야 하는 부담을 지니고 있다. 본 논문에서는 기존의 색인 방법들의 문제점들을 완화할 수 있는 새로운 색인 방법을 제안한다. 그리고 실험을 통하여 제안하는 방법의 성능을 평가한다.

1. 서 론

정보 검색 시스템의 중요한 역할 중의 하나는 검색된 각각의 문서에 대하여 순위 결정 방법(ranking)을 적용하는 것이다. 순위 결정 방법은 문서와 질의 사이의 유사도를 나타내는 문서값(document value)을 계산하고, 계산된 유사도에 따라 문서에 순위를 부여한다. 벡터 공간 모델은 문서와 질의를 가중치가 부여된 색인어들의 벡터로 표현하고, 표현된 벡터들의 내적으로써 문서와 질의 사이의 유사도를 계산한다 [1]. 따라서 문서와 질의를 표현하는 색인어를 추출하는 색인 방법은 문서의 순위 결정에 영향을 미치는 중요한 요소이다.

기존의 한글 자동 색인 방법들은 크게 어절 단위의 색인법과 형태소 단위의 색인법으로 분류될 수 있다. 어절 단위의 색인법은 질의와 문서내의 각 어절들에 대해 조사, 어미, 접미사 등을 절단하여 색인어를 추출하는 것으로[7, 8], 색인 과정이 간단하다. 그러나 이 방법은 띄어 쓰기에 대한 부적절한 처리로 인하여 문서에 복합 명사들이 많이 포함되어 있을 경우 검색 효과가 저하되는 문제점을 지니고 있다. 형태소 단위의 색인법은 형태소 해석이나 구문 해석을 이용하여 어절 또는 문장을 분석함으로써 문서나 질의의 내용 표현에 적절한 명사 또는 명사구들을 추출한다 [9, 10, 11]. 이 방법은 복합 명사를 단일 명사들로 분리할 수 있어, 위에서 언급한 어절 단위 색인에서의 띄어 쓰기 문제를 극복한다. 그러나 형태소 분석이나 구문해석을 위한 규칙이 복잡하고, 형태소 해석 결과의 애매성 등의 이유로 부정확한 색인어가 추출될 수 있다. 또한 사전과 같은 언어 정보들을 개발하고 유지 관리해야 하는 부담이 있다.

본 논문에서는 어절 단위 색인법과 n-gram 방법 [2, 3]을 결합한 n-gram 기반의 색인 방법을 제안한다. n-gram이란 인접한 n 개의 음절을 의미한다. 제안하는 방법은 문장내의 각 어절에 대하여 어절 단위의 색인법을 적용하고, 그 결과로 생성된 어절에 n-gram 방법을 적용함으로써 색인어들을 추출한다. 이 방법은 어절 단위 색인에서의 띄어 쓰기 문제를 완화할 수 있으며, 형태소 단위 색인에서와 같은 복잡한 분석 규칙과 사전 및 언어 정보에 대한 관리를 요구하지 않는다.

2. 기존의 색인 방법들

어절 단위의 색인법은 불용어를 제외한 모든 어절들을 색인어 후보로 간주하고, 후보 어절에서 색인어의 부분으로서 가치가 없는 비색인분절(non-indexable segment)을 제거하여 나머지 색인분절(indexable segment)을 색인어로 선택한다 [7, 8]. 비색인분절이란 조사, 어미, 접미사 등과 같이 체언의 뒤에 붙여 쓰이지만

ment)을 색인어로 선택한다[7, 8]. 비색인분절이란 조사, 어미, 접미사 등과 같이 체언의 뒤에 붙여 쓰이지만 색인어에 포함시키기에는 무의미한 부분들을 말한다. 일반적으로 비색인분절의 검출을 위하여 최장 일치의 원칙(principle of the longest match)이 이용되며, 최장 일치의 원칙이란 주어진 어절내에서 검출될 수 있는 비색인분절 중에서 가장 긴 것을 선택하는 방법이다.

형태소 단위 색인법은 문서 분석의 정도에 따라 형태소 해석만을 이용하는 방법과 구문 해석을 이용하는 방법으로 구분된다. 형태소 해석만을 이용하는 방법은 문장을 구성하고 있는 최소 의미의 형태소들을 파악하고, 단어나 어절 자체의 모호함 때문에 여러개의 해석 결과가 산출되는 형태소 해석의 애매성을 제거한다. 그리고 그 결과로부터 단순 명사를 선택하고, 이들로부터 불용어를 제외한 나머지를 문서와 질의의 표현을 위한 색인어로 선정한다[9].

구문 해석을 통한 방법은 형태소 해석에서 한 단계 더 나아가 문장 단위의 구문 해석을 수행한다. 예를 들면, 색인어 추출을 위해 명사구들의 의미 역할을 설정하는 격문법을 이용한 구문해석이 시도되었다[10, 11]. 서술어가 문장 중에 반드시 지녀야 되는 격을 필수격이라 정의하고, 한국어 용언이 지닐 수 있는 격틀(case frame)을 이용하여 각 문장의 필수격을 찾는 방식으로 구문해석을 수행하였다. 그리고 필수격에 해당하는 명사나 명사구들을 색인어 후보로 채택하고 불용어를 제거한다.

어절 단위 색인법과 형태소 단위 색인법의 차이점은 최종적으로 추출되는 색인어의 단위에서 찾아볼 수 있다. 어절 단위 색인법은 여러개의 명사가 결합된 복합 명사 어절의 경우에도 단순히 조사나 접미사 등의 절단만을 수행한다. 반면에 형태소 단위 색인법은 복합 명사 어절을 최소 의미의 형태소로 분리하여 단순 명사를 색인어로 선정할 수 있다. 예를 들면, '정보검색서비스는'이라는 어절에 대하여, 전자의 방법은 '정보검색서비스'를 색인어로 선정하나, 후자의 방법은 '정보','검색','서비스'와 같은 단순 명사들을 색인어로 선정할 수 있다.

3. N-gram 기반의 색인 방법

3.1 기존 색인 방법들의 문제점

어절 단위 색인법은 음절들을 절단하는 과정에서 다음과 같은 오류를 발생시킬 수 있다. 첫째, 어절 '벨기에로서는'과 '벨기에'에 대해 각각 '로서는'과 '에'를 비색인분절로 판정하고, '벨기에'와 '벨기'라는 서로 다른 색인어를 추출한다. 둘째, 한글은 복합 명사를 구성하는 단순 명사들 사이의 띄어 쓰기에 대한 규칙을 자유롭게 규정하고 있다. 따라서 문서들이 복합 명사를 많이 포함하고 있을 경우, 어절 단위 색인법은 검색 효과를 저하시키는 다음과 같은 문제점을 지니고 있다. 예를 들면, 문서 d와 질의 q가 각각 "분산 시스템", "분산시스템"을 포함하고 있다고 가정하자. 어절 단위의 색인법은 이들 문서와 질의에 대하여 각각 {분산, 시스템}, {분산시스템}의 서로 다른 색인어를 생성한다.

위에서 언급된 복합 명사의 띄어 쓰기 문제는 형태소 단위 색인법을 사용하여 단순 명사들을 색인어로 추출한다면 극복될 수 있다. 예제의 문서d와 질의q의 각 어절에 대하여 이 방법은 동일하게 {분산, 시스템}을 색인어로 생성할 수 있다. 즉, 문서와 질의에서 복합명사의 띄어쓰기가 일치하지 않더라도, 동일한 색인어들이 생성될 수 있다.

형태소 단위의 색인법은 복합 명사 문제를 잘 처리할 수 있으며, 검색 효과도 좋은 것으로 보고되고 있다. 그러나 형태소 해석이나 구문 해석과 관련하여 다음과 같은 문제점을 지니고 있다. 첫째, 형태소 단위 색인법에서는 형태소 해석의 애매성, 사전 미등록어, 비문법적인 어절로 인한 형태소 해석의 오류로 부정확한 색인어가 추출될 수 있다. 둘째, 형태소 단위 색인법은 형태소 해석 또는 구문 해석에서 명사 사전이나 격틀 사전과 같은 많은 언어 정보(linguistic knowledge)들을 필요로 한다.

3.2 제안하는 색인 방법

본 논문에서는 어절 단위 색인법과 n-gram 방법을 결합한 새로운 색인 방법을 제안한다. n-gram이란 인접한 n개의 음절을 말한다[2, 3]. 예를 들면, '프로그래밍'이라는 단어에 대해 2-gram은 '프로', '로그', '그래', '래밍'이며, 3-gram은 '프로그', '로그래', '그래밍'이다. 제안하는 방법의 색인 과정은 다음과 같다.

a. 문서나 질의의 모든 어절들을 인식한다.
b. 불용어를 제거한다.
c. 비색인분절을 제거한다.,
d. 색인분절을 n-gram들로 분할한다.
e. 생성된 n-gram들에 가중치를 부여한다.

제안하는 n-gram 기반의 색인 방법은 검색효과의 측면에서 다음과 같은 장점을 지니고 있다.

- 복합 명사의 띄어 쓰기로 인한 문제를 완화시킨다. 예를 들면, "분산 시스템"과 "분산시스템"에 대해 2-gram 기반의 색인법은 각각 {분산, 시스, 스템}, {분산, 산시, 시스, 스템}의 색인어를 생성함으로써 3개의 색인어들을 공통적으로 포함하고 있다.
- 어절 단위 색인법을 이용했을 때의 절단 오류로 인한 검색 효과의 저하를 줄일 수 있다. 어절 단위 색인법은 "벨기에로서는", "벨기에"에 대해 각각 "벨기에", "벨기"를 색인어로 생성한다. 그러나 2-gram 방법을 적용함으로써 "벨기"라는 공통된 색인어를 얻을 수 있다.
- 한글 문서를 살펴보면 "가공력" 또는 "가공물"과 같

이 단순 명사의 뒤에 한 글자의 명사가 붙어서 형성된 명사들을 많이 발견할 수 있다. 형태소 단위 색인법은 이러한 명사들을 단순 명사로 취급하여 색인어로 추출한다. 이러한 경우 사용자가 '가공'이라는 질의를 입력하면, 관련된 많은 문서들이 검색되지 않을 수 있다. 제안하는 색인법은 이와 같은 경우에 관련 문서의 검색을 도와준다.

- 철자 오류나 일관성이 없는 외래어 표기 문제를 완화한다.

4. 성능 평가

4.1 환경 및 실험자료

본 논문에서는 제안하는 n-gram 기반의 색인 방법과 기존 색인 방법들의 성능을 비교하기 위해 SMART 시스템을 이용하였다 [4]. SMART 시스템은 문서와 질의를 색인어들의 벡터로서 표현하는 벡터 공간 모델을 기반으로 한다. 색인 과정을 통해 추출된 색인어들에 가중치를 부여함으로써 문서와 질의 표현을 위한 벡터를 생성하고, 생성된 벡터를 이용하여 문서와 질의 사이의 유사도를 계산한다. 그리고 계산된 유사도에 따라 문서들에 순위를 부여한다.

정보 검색에 관한 많은 연구들은 색인어에 가중치를 부여하기 위하여 출현 빈도(term frequency), 장서 빈도(collection frequency), 정규화(normalization)의 세 가지 요소를 고려한다 [5, 6]. 출현 빈도는 문서내에서 자주 출현하는 색인어에 보다 높은 가중치를 부여한다. 장서 빈도는 전체 문서들 중에서 적은 수의 문서에 출현하는 색인어에 보다 높은 가중치를 부여한다. 그리고 정규화 요소는 작은 크기의 문서들이 문서값 계산에 있어 불공정하게 취급되는 것을 피하도록 한다. (표 1)은 각각의 구성 요소에 대해 잘 알려진 공식들을 보여준다.

성능 평가를 위한 실험 자료로서 KT 자료 집합을 사용하였다 [12]. KT 자료 집합에는 정보과학회논문지, 한국정보과학회 학술발표대회논문집, 정보관리학회지에 수록된 1,053개의 논문과 30개의 질의로 구성되어 있다. 입력된 문서는 저자, 서명, 초록, 분류 번호, 색인어 등 18개의 항목을 지니고 있으며, 각 질의에 대한 적합 문서들이 제시되어 있다.

4.2 검색 효과 비교

본 절에서는 제안하는 n-gram 기반의 색인 방법과 기존의 색인 방법들의 성능을 검색 효과의 측면에서 평가한다. 검색 효과의 측정 방법으로 11-포인트 평균정확률이 사용되었다. (그림1)은 제안하는 n-gram 기반의 색인 방법 중에서 2-gram 기반의 색인 방법이 가장 높은 검색 효과를 제공함을 보여준다. 1-gram 기반의 색인 방법을 사용할 경우, 2음절보다 1음절이 갖는 애매성이 크기 때문에 부적합한 문서들이 과다하게 검색

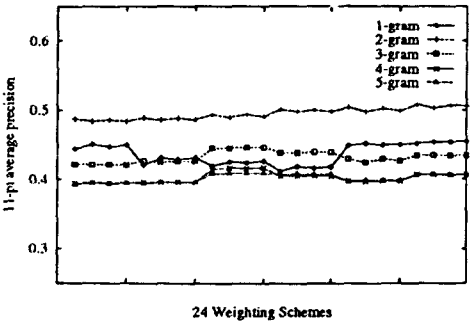

그림 1: n-gram 기반의 색인 방법 이용시의 검색 효과 비교 (n=1 ... 5)

될 수 있다. 예를 들면, 문서 d가 "기존 알고리즘들의 복잡도를 계산하고 이를 비교 분석하여"라는 문장을 포함하고 있고, 질의 q는 어절 "분산교환망"을 포함하고 있다고 가정하자. 이때 문서 d는 질의 q와 관련이 없을 지라도, 1-gram 기반의 색인 방법을 사용할 경우 '산'과 '분'이 색인어로 추출되어 질의 q의 검색 결과로 출력될 가능성이 있다.

3-gram, 4-gram, 5-gram 기반의 색인 방법을 이용할 경우에는 음절수로 인한 애매성이 적어 부적합한 문서들의 검색이 줄어들지만, 다음과 같은 경우로 인하여 검색 효과가 저하된다. 문서 d와 질의 q가 각각 '정보검색'과 '정보 검색'을 포함한다고 가정하자. 3-gram 기반의 색인 방법은 d와 q를 위한 색인어로 각각 {정보검, 보검색}과 {정보, 검색}을 추출한다. 따라서 동일한 어절들에 대하여 공통된 색인어가 추출되지 않는다.

다음으로 어절 단위 색인 방법, 형태소 단위 색인 방법과 2-gram 기반의 색인 방법을 비교한다. 일반적으로 어절 단위 색인과 형태소 단위 색인은 수작업으로 수행했을 때 가장 정확한 결과를 얻을 수 있다. 본 논문에서는 이러한 가정하에 수작업으로 비색인분절들을 제거함으로써 어절 단위 색인을 수행하고, 문서로부터 단순 명사들을 수작업으로 추출함으로써 형태소 단위의 색인을 수행하였다.

(그림2)는 2-gram 기반의 색인, 수작업에 의한 어절 단위 색인, 그리고 수작업에 의한 단순 명사 추출시의 검색 효과를 보여준다. (그림2)에서 2-gram 기반의 색인 방법은 수작업에 의한 어절 단위 색인에 비하여 우수한 검색 효과를 보인다. 이러한 결과는 제안하는 방법이 복합명사의 띄어 쓰기 문제를 어절 단위 색인 방법보다 잘 처리할 수 있기 때문이라고 해석될 수 있다. 한편, 수작업에 의한 단순 명사 추출과 비교하여 2-gram 기반의 색인 방법은 유사한 검색 효과를 보이고 있다.

표 1: 가중치 부여 기법의 구성 요소

출현 빈도		
b	1.0	색인어 출현 빈도를 무시하고 벡터를 구성하는 색인어에 1의 가중치 부여
n	tf	문서나 질의내에서 색인어의 출현 빈도
a	$0.5 + 0.5\frac{tf}{\max tf}$	보강된 정규화 출현 빈도 (tf를 $\max tf$로 나누고, 그 결과가 0.5에서 1.0 사이의 값을 갖도록 정규화)
l	$\ln tf + 1.0$	색인어 출현 빈도에 로그함수 적용
장서 빈도		
n	1.0	색인어 출현 빈도(b, n, a, l)만으로 가중치 생성
t	$\ln \frac{N}{n}$	색인어 출현 빈도와 역문헌 빈도를 곱한다 (N은 전체 문서들의 수이며, n은 그 색인어를 포함하고 있는 문서들의 수이다.)
정규화		
n	1.0	색인어 출현 빈도와 장서 빈도만으로 유도된 가중치를 사용
c	$\frac{1}{\sqrt{\sum_{vector} w_i^2}}$	유클리디안 벡터 길이를 이용한 코사인 정규화

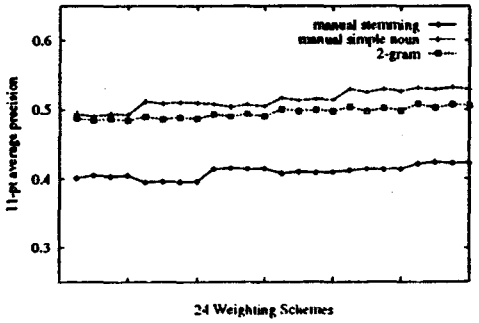

그림 2: 2-gram 기반의 색인, 수작업에 의한 어절 단위 색인, 수작업에 의한 단순 명사 추출시의 검색 효과 비교

5. 결론

기존의 한글 자동 색인을 위한 어절 단위 색인법은 구현이 간단한 반면, 복합 명사의 띄어 쓰기 문제 등을 적절히 처리할 수 없는 문제점을 지니고 있다. 한편 형태소 단위 색인법은 단순 명사를 추출함으로써 복합 명사의 띄어 쓰기 문제를 극복할 수 있고, 검색 효과가 좋은 것으로 알려지고 있다. 그러나 형태소 해석이나 구문 해석을 위한 언어 정보들의 개발을 요구한다.

본 논문에서는 n-gram 기반의 색인법을 제안하였다. 이 방법은 복합 명사의 띄어 쓰기 문제를 완화할 수 있으며, 형태소 단위 색인법에서와 같은 언어 정보의 개발도 거의 요구하지 않는다. 또한 색인어의 부분으로서 가치가 없는 음절들의 절단 오류나, '가공품'처럼 하나의 단일 형태소로 취급되는 복합명사로 인한 문제를 완화할 수 있고, 실제 문서에서 많이 발견되는 철자 오류나 일관성이 없는 외래어 표기 문제에도 대처할 수 있는 장점을 지닌다.

참고 문헌

[1] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Communications of the ACM, 18:11, pp.613-620, November 1975.

[2] W.B. Cavnar, "N-Gram-Based Text Filtering for TREC-2," Proceedings of the Second Text REtrieval Conference(TREC-2), NIST Special Publication 500-215, pp.171-179, 1994.

[3] M. Damashek, "Gauging Similarity with n-Grams: Language-Independent Categorization of Text," Science, Vol.267, pp.843-848, 1995.

[4] G. Salton, The SMART Retrieval System, Englewood Cliffs, N.J.: Prentice Hall, Inc., 1971.

[5] G. Salton, C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, Vol.24, No.5, pp.513-523. 1988.

[6] J.H. Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes," Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995. (to be published)

[7] 김영환, "한글 한자 혼용문의 자동 색인 시스템", 한국과학기술원 석사학위논문, 1982.

[8] 예용희, "국내 문헌 정보 검색을 위한 키워드 자동 추출 시스템 개발", 정보관리연구, 제23권 1호, p.39-62, 1992.

[9] 이현아, 홍남희, 이종혁, 이근배, "한국어 형태소 구조 규칙에 기반한 색인 시스템의 구현", 한국정보과학회 봄 학술발표논문집, pp.933-936, 1995.

[10] 정진성, "단일 문서내에서의 언어 및 통계 정보를 이용한 자동 색인", 한국과학기술원 석사학위논문, 1992.

[11] 한성현, "구문해석을 이용한 색인어 자동 추출 시스템의 설계와 구현", 한국과학기술원 석사학위논문, 1990.

[12] 김성혁 외 5인, "자동 색인기 성능 시험을 위한 Test Set 개발", 정보관리학회지 제11권 1호, 1994.