한글 문서의 효과적인 검색을 위한 *n*-Gram 기반의 색인 방법

An *n*-Gram-Based Indexing Method for Effective Retrieval of Hangul Texts

이준호(Joon-Ho Lee)*, 안정수(Jeong-Soo Ahn)**,
박현주(Hyun-Joo Park)*, 김명호(Myoung-Ho Kim)***

목 차

초 록

기존의 한글 자동 색인 방법들은 어절 단위 색인법과 형태소 단위 색인법으로 분류될 수 있다. 전자는 문서내의 어절에서 비색인 분절을 절단함으로써 색인어를 추출하는 방법으로, 문서들이 많은 복합 명사들을 포함할 경우 검색 효과가 저하된다. 후자는 형태소 해석이나 구문 해석을 이용하여 중요한 의미를 갖는 명사나 명사구를 추출하는 방법으로, 단일 명사를 추출함으로써 복합 명사의 띄어쓰기 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.

ABSTRACT

Conventional automatic indexing methods for Hangul texts can be classified into two groups as follows: One is to extract index terms by removing non-indexable segments from word-phrases, and the other is to generate index terms from the morphemes of word-phrases. The former suffers from the problem of word boundaries when documents contain many compound nouns. The latter can overcome the word boundary problem by extracting simple nouns, but has many overheads to develop a lot of linguistic knowledges needed in the indexing procedure. In this paper we propose a new indexing method based on n-grams. This method alleviates the problems of previous indexing methods related with word boundaries and linguistic knowledges. We also compare the effectiveness of the *n*-gram based indexing method with that of the previous ones.

*연구개발정보센터 연구개발부 선임연구원 **연구개발정보센터 연구개발부 연구원
***한국과학기술원 전산학과 부교수
■ 논문 접수일 : 1996년 4월 20일

1. 서 론

지난 30년 동안 과학과 기술 분야에서의 급속한 발전은 수많은 주제들에 대해 방대한 양의 정보가 생성되는 정보화 사회를 탄생시켰다. 원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정짓는 중요한 요소가 되었다. 정보 검색 시스템은 질의 요구에 적합한 문서들을 사용자에게 제공함으로써 대용량의 데이터로부터 주어진 시간 내에 원하는 정보를 발견할 수 있도록 도와 준다(Salton 1987).

정보 검색 시스템의 중요한 역할 중의 하나는 검색된 각각의 문서에 대하여 순위 결정 방법(ranking)을 적용하는 것이다. 문서 순위 결정 방법은 문서와 질의 사이의 관련 정도를 나타내는 유사도(similarity)를 계산하고, 계산된 유사도에 따라 문서에 순위를 부여한다. 높은 순위를 갖는 문서일수록 질의에 대한 만족도가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는 데 소모되는 시간을 최소화할 수 있다(Lee et al. 1994). 문서 순위 결정 방법을 제공하는 벡터 공간 모델은 문서와 질의를 가중치가 부여된 색인어들의 벡터로 표현하고, 표현된 벡터들의 내적으로써 문서와 질의 사이의 유사도를 계산한다(Salton et al. 1975 ; Salton 1989). 일반적으로 문서나 질의의 내용을 표현하는 색인어들은 이를 추출하는 색인 방법에 따라 달라지므로, 벡터 공간 모델에서의 문서 순위 결정은 사용하는 색인 방법에 의해 영향을 받는다. 본 논문에서는 한글 문서를 위해 고안된 다양한 색인 방법들이 벡터 공간 모델하에서 검색 효과에 미치는 영향을 고찰한다.

기존의 한글 자동 색인 방법들은 추출되는 색인어의 단위에 따라 어절 단위 색인법과 형태소 단위 색인법으로 분류될 수 있다. 어절 단위 색인법은 문서와 질의의 각 어절들에 대해 색인어의 일부분으로서 가치가 없는 비색인 분절(non-indexable segment) 즉, 조사, 어미, 접미사 등의 음절들을 절단하여 원문에 가까운 형태로 색인어를 추출하는 것으로(김영환 1982 ; 안현수 1986 ; 예용희 1992), 색인 과정이 비교적 간단하다. 그러나 이 방법은 비색인 분절을 절단할 때 오류가 발생할 수 있으며, 특히 의미가 동일한 복합 명사의 띄어쓰기 문제를 적절히 처리하지 못하기 때문에 문서들 내에 많은 복합 명사들이 포함되어 있을 경우 검색 효과가 저하되는 문제점을 지니고 있다.

형태소 단위의 색인법은 일반적으로 형태소 해석을 수행함으로써 문장 중의 각 어절을 명사, 조사, 부사 등의 형태소 단위로 분리한 후, 문서나 질의의 내용 표현에 적절한 명사 또는 명사구들을 추출한다(강승식 외 2인 1995 ; 이현아 외 3인 1995 ; 정진성 1992 ; 최기선 1991 ; 한성현 1991). 이 방법은 복합 명사를 단일 명사들로 분리할 수 있어 앞에서 언급한 어절 단위 색인법에서의 복합 명사 띄어쓰기 문제를 극복할 수 있다. 그러나 형태소 해석에 의존적인 형태소 단위 색인법은 형태소 해석을 위한 규칙이 복잡하고, 형태소 해석 결과의 애매성, 미등록어, 비문법적인 어절 등의 이유로 부정확한 색인어가 추출될 수 있다. 또한 형태소 사전과 같은 언어 정보

들을 개발하고 유지 관리해야 하는 부담을 안고 있다.

본 논문에서는 어절 단위 색인법과 *n*-gram 방법(Cavnar 1994 ; Damashek 1995)을 결합한 *n*-gram 기반의 색인 방법을 제안한다. *n*-gram이란 인접한 *n*개의 음절을 의미한다. 제안하는 방법은 문장 내의 각 어절에 대하여 어절 단위의 색인법을 적용하고, 그 결과로 생성된 분절에 *n*-gram 방법을 적용함으로써 색인어들을 추출한다. 예를 들어 "정보검색을"이란 어절에 대하여 2-gram 기반의 색인 방법은 '정보', '보검', '검색'의 색인어들을 추출한다. 비록 제안하는 색인 방법은 '보검'과 같은 의미 없는 색인어를 추출할지라도, 어절 단위 색인법에서의 복합 명사 띄어쓰기 문제를 완화할 수 있다. 또한 형태소 해석을 수행하지 않기 때문에 형태소 단위 색인에서와 같은 복잡한 문장 해석 규칙이나 언어 정보의 개발을 요구하지 않으며, 단일 명사를 추출할 수 있는 형태소 단위 색인법과 유사한 검색 효과를 제공한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 벡터 공간 모델과 기존의 한글 자동 색인 방법에 대해 기술하고, 3장에서는 제안하는 *n*-gram 기반의 색인 방법에 대하여 설명한다. 4장에서는 벡터 공간 모델을 기반으로 하는 SMART 시스템과 KT Test Set을 사용하여 제안하는 색인 방법과 기존 한글 자동 색인 방법들의 성능을 평가한다. 그리고 마지막으로 5장에서 결론을 제시한다.

2. 관련 연구

2.1 벡터 공간 모델

벡터 공간 모델은 정보 검색 시스템의 검색 과정을 표현하기 위한 모델의 하나로서, 사용자의 질의를 완전히 만족하는 문서들만을 검색하는 부울 모델(Boolean model)과는 달리 질의를 부분적으로 만족하는 문서들도 검색할 수 있으며, 또한 질의와의 관련 정도에 따라 검색되는 문서들의 순위를 결정할 수 있다. 이 모델에서는 문서와 질의로부터 색인어를 추출하고 추출된 색인어에 가중치를 부여함으로써, 문서와 질의를 다음과 같은 *n*개의 가중치가 부여된 색인어들의 벡터로서 표현한다 (Salton 1975 ; Salton 1989).

$$d = \{(t_1, w_{d1}),\ (t_2, w_{d2}),\ \cdots,\ (t_n, w_{dn})\}$$
$$q = \{(t_1, w_{q1}),\ (t_2, w_{q2}),\ \cdots,\ (t_n, w_{qn})\}$$

여기서 t_i는 색인 방법에 의해 추출된 색인어를 표시하며, w_{di}와 w_{qi}는 색인어 t_i가 문서 d와 질의 q에서 갖는 중요도를 반영하는 가중치이다.

문서나 질의에 대한 색인어 벡터들이 형성되면, 이 벡터들을 기반으로 문서와 질의 사이의 유사도를 계산하고 계산된 유사도에 따라 문서들의 순위를 결정한다. d와 q를 각각 문서와 질의를 표현하는 벡터라 할 때, 문서 d와 질의 q 사이의 유사도는 다음 식과 같이 두 벡터들의 내적으로 계산된다.

$$\mathrm{Sim}(d, q) = \sum_{i=1}^{n} w_{di} \times w_{qi}$$

따라서 이러한 벡터 공간 모델에서 문서 순위 결정의 정확성은 색인어에 가중치를 부여하는 기법뿐만 아니라 문서와 질의를 표현하기 위한 색인어들을 추출하는 색인 방법에 의해서도 영향을 받는다.

2.2 기존의 한글 자동 색인 방법

과거의 색인 작업은 훈련된 사서나 주제 전문가에 의해 수작업으로 수행되었다. 그러나 수작업에 의한 색인은 과다한 지적 노력을 필요로 하며, 비용대 효과면에서도 그다지 효율적이지 못하다는 사실이 여러 실험 결과 밝혀졌다. 또한 수작업의 색인은 색인자의 주관에 따라 색인의 양이나 질이 달라질 수 있어 색인의 일관성이 결여되는 문제점을 안고 있다. 따라서 컴퓨터를 사용하여 문서와 질의를 자동적으로 분석함으로써 색인어를 추출하는 자동 색인 기법들이 출현하게 되었다(Cleverdon 1984 ; Salton 1986).

한글 문서에서 문서의 내용을 나타내는 것은 주로 명사이므로, 기존의 한글 자동 색인 방법들은 명사나 명사구 추출에 중점을 두어 왔다. 즉, 명사 뒤에 조사가 붙는다는 특성, 조사나 어미로써 명사의 문장 내에서의 성분이 결정된다는 한글의 언어 구조 특징 등이 이용되었다. 대체로 이들 기존의 한글 자동 색인 방법들은 추출되는 색인어의 단위에 따라 크게 어절 단위 색인법과 형태소 단위 색인법으로 〈표 1〉과 같이 분류될 수 있다.

2.2.1 어절 단위 색인법

어절 단위 색인법은 문서나 질의에서 불용어를 제외한 모든 어절들을 색인어 후보로 간주하고, 각 어절로부터 색인어의 부분으로서 무의미한 비색인 분절(non-indexable segment)을 제거한 나머지 색인 분절(indexable segment)을 색인어로 선택하는 방법이다(김영환 1982 ; 안현수 1986 ; 예용희 1992). 특히 한글에서 문서나 질의를 표현할 수 있는 체언이나 용언의 명사형 뒤에 조사나 접미사

〈표 1〉 한글 자동 색인 방법

	어절 단위 색인법	형태소 단위 색인법	
		형태소 해석	구문 해석
색인 과정	1. 문장내의 어절 인식	1. 문장내의 어절들의 형태소 해석	1. 문장내의 어절들의 형태소 해석
	2. 비색인 분절의 제거	2. 형태소 해석의 애매성 제거	2. 형태소 해석의 애매성 제거
			3. 형태소 해석 결과를 이용한 구문 해석
		3. 명사 추출	4. 명사 추출
	3. 불용어 제거	4. 불용어 제거	5. 불용어 제거

등이 붙는다는 특성에 근거하여 문서 내에서 명사형을 포함하고 있는 어절들을 인식하고, 인식된 어절에서 조사나 접미사 등을 제거하는 데 중점을 두고 있다. 비색인 분절이란 아래와 같이 체언의 뒤에 붙여 쓰이지만 색인어에 포함시키기에는 무의미한 조사, 어미, 접미사 등의 음절들을 말한다.

는, 은, 가, 이	을, 를, 에, 에게
와, 과	부터, 로부터
들, 들도, 들의	마다, 만큼, 보다
로서, 로써	와의, 과의, 처럼
하다, 하는, 하도록	되다, 되는, 되도록
당하다, 시키다	…

어절 단위 색인법의 색인 과정은 대체로 세 단계로 이루어진다. 먼저 빈칸과 같은 문자를 구분자로 하여 문서 내에서의 모든 어절들을 인식한다. 그리고 인식된 각 어절에 대해 비색인 분절을 제거한다. 일반적으로 비색인 분절의 검출을 위하여 최장 일치의 원칙(principle of the longest match)이 이용된다. 최장 일치의 원칙이란 주어진 어절 내에서 검출될 수 있는 비색인 분절들 중에서 가장 긴 분절을 선택하는 방법이다. 예를 들면, "시스템으로부터"라는 어절이 있을 때 '으로부터', '로부터', '부터'의 비색인 분절 중에서 가장 긴 '으로부터'를 선택하여 이를 제거한다. 그리고 마지막 단계에서 어절 단위 색인법은 불용어를 제외한 나머지 색인 분절들을 색인어로 선정한다.

2.2.2 형태소 단위 색인법

형태소 단위 색인법은 〈표 1〉에서 보는 바와 같이 문장 분석의 정도에 따라 형태소 해석만을 이용하는 방법과 구문 해석을 이용하는 방법으로 구분된다. 형태소 해석만을 이용하는 방법은 문장의 모든 어절들에 대해 형태소 해석을 수행하여 최소 의미의 명사들을 문서와 질의의 표현을 위한 색인어로 선정한다(이현아 외 3인 1995). 이 방법은 형태소 해석, 애매성 제거, 단일 명사 추출, 불용어 제거의 네 단계를 거쳐 색인을 수행한다. 형태소 해석 단계에서는 문장을 최소의 의미 단위로 구분하여 문장 내의 각 어절을 구성하고 있는 형태소들을 파악한다(강승식 외 2인 1995). 그리고 단어나 어절 자체의 모호함 때문에 하나의 단어나 어절에 대해 여러 개의 해석 결과가 산출되는 형태소 해석의 애매함을 제거한다. 즉, 여러 개의 해석 결과로부터 하나의 해석 결과를 선정한다. 그리고 선택된 해석 결과로부터 단일 명사와 같은 최소 의미의 형태소들을 추출하고, 불용어를 제외한 나머지를 색인어로 선정한다.

구문 해석을 통한 방법은 형태소 해석에서 한 단계 더 나아가 문장 단위의 구문 해석을 수행하여 문장에서 중요한 의미를 갖는 특정한 명사나 명사구를 색인어로 선정한다. 예를 들면, 색인어 추출을 위해 문장의 서술어와 그것과 연관된 명사구들의 의미 역할을 설정하는 격문법을 이용한 구문 해석이 시도되었다(한성현 1991 ; 최기선 1991 ; 정진성 1992). 여기에서는 서술어가 문장 중에 반드시 가져야 되는 격을 필수격이라 정의하고,

한국어 용언이 가질 수 있는 격틀(case frame)을 이용하여 각 문장의 필수격을 찾는 방식으로 구문해석을 수행한다. 그리고 필수격에 해당하는 명사나 명사구들을 색인어 후보로 채택하고 불용어를 제거한다.

어절 단위 색인법과 형태소 단위 색인법은 최종적으로 추출되는 색인어의 단위에서 구별된다. 어절 단위 색인법은 여러 개의 명사가 결합된 복합 명사 어절의 경우에도 단순히 조사나 접미사 등의 절단만을 수행하고 나머지 분절을 색인어로 취한다. 반면에 형태소 단위 색인법은 복합 명사 어절을 최소 의미의 형태소로 분리하여 단일 명사를 색인어로 선정할 수 있다. 예를 들면, '정보검색서비스가' 라는 어절에 대해, 전자의 방법은 '정보검색서비스' 를 색인어로 선정하지만, 후자의 방법은 '정보', '검색', '서비스' 의 색인어들을 선정할 수 있다.

3. 한글 문서를 위한 새로운 색인 방법

3.1 기존 색인 방법들의 분석

최장 일치의 원칙을 사용하는 어절 단위 색인법은 구현이 간단한 반면, 비색인 분절의 절단 과정에서의 오류로 인해 추출되는 색인어의 일관성이 떨어질 수 있다. 예를 들면, '벨기에로서는' 과 '벨기에' 의 두 어절에 대해 어절 단위 색인법은 각각 '로서는' 과 '에' 를 비색인 분절로 판정하고, '벨기에' 와 '벨기' 의 서로 다른 색인어들을 추출한다.

한글은 복합 명사를 구성하는 단일 명사들 사이의 띄어쓰기를 자유롭게 규정하고 있다. 이러한 한글 체계에서 문서들이 복합 명사를 많이 포함하고 있을 경우, 어절 단위 색인법은 검색 효과를 저하시키는 다음과 같은 문제점을 지닌다. 예를 들면, 문서 d_1과 d_2가 각각 "정보검색"과 "정보 전송"을 포함하고, 사용자의 질의 q_1은 "정보 검색"이라고 가정하자. 문서 d_1은 질의 q_1과 의미가 동일한 복합 명사를 다른 띄어쓰기 형태로 포함하고, 문서 d_2는 문서 d_1보다 질의와 관련성이 적은 명사를 포함한다. 이들 문서 d_1, d_2와 질의 q_1에 대하여 어절 단위 색인법을 적용할 때, 문서와 질의의 벡터 표현과 질의에 대한 각 문서들의 유사도는 다음과 같다.

d_1 : { (정보검색, w_1) }
d_2 : { (정보, w_2), (전송, w_3) }
q_1 : { (정보, w_4), (검색, w_5) }
$Sim(d_1, q_1) = 0$
$Sim(d_2, q_1) = w_2 w_4$

따라서 문서 d_1은 질의 q_1과 다른 형태의 띄어쓰기로 인해 일치하지 않는 색인어를 갖기 때문에, 질의 q_1에 관련성이 적은 문서 d_2가 관련성이 많은 d_1보다 높은 유사도를 갖는다.

복합 명사의 띄어쓰기 문제는 형태소 단위 색인법을 사용하여 단일 명사들을 색인어로 추출함으로써 극복될 수 있다. 앞에서의 문서 d_1, d_2와 질의 q_1에 대해 형태소 단위 색인법을 이용하여 단일 명사를 색인어로 추출할 때, 문서와 질의의 벡터 표현과 질의에 대한 각 문서들의 유사도는 다음과 같다.

d_1 : { (정보, w_6), (검색, w_7) }
d_2 : { (정보, w_8), (전송, w_9) }
q_1 : { (정보, w_{10}), (검색, w_{11}) }
$Sim(d_1, q_1) = w_6 w_{10} + w_7 w_{11}$
$Sim(d_2, q_1) = w_8 w_{10}$

따라서 문서 d_2의 색인어 '정보'의 가중치 w_8이 아주 큰 값을 갖지 않는 한 유사도는 $Sim(d_1, q_1) > Sim(d_2, q_1)$로 계산되어 정확한 순위 결정이 이루어진다.

형태소 단위 색인법은 이와 같이 복합 명사의 띄어쓰기 문제를 잘 처리할 수 있으며, 검색 효과도 좋은 것으로 보고되고 있다. 그러나 형태소 해석이나 구문 해석과 관련하여 다음과 같은 몇 가지의 문제점을 지니고 있다. 형태소 단위 색인법은 형태소 해석이나 구문 해석 과정에서 형태소 사전, 격틀 사전 등의 언어 정보들을 필요로 한다. 특히 형태소 사전은 많은 개발 시간과 비용을 요구하며, 형태소 해석의 대상이 되는 문서들의 성질에 크게 의존하는 경향이 있어 문서의 종류마다 서로 다르게 개발되어야 하는 부담을 안고 있다.

형태소 단위 색인법은 형태소 해석에 의존하므로 형태소 해석 과정에서의 오류는 부정확한 색인어들의 생성을 야기한다. 형태소 해석의 가장 큰 오류는 사전 내에 등록되지 않은 미등록어로 인해 발생하며, 특히 과학 기술 분야에서는 많은 전문 용어들이 미등록어로 처리되어 검색 효과가 떨어질 수 있다. 그 외에도 단어나 어절 자체의 모호함에서 오는 형태소 해석의 애매성이나, 실제 문서에서 많이 발견되는 철자 오류와 띄어쓰기 오류 등의 비문법적인 어절들도 형태소 해석 오류의 원인이 되고 있다. 〈표 2〉는 이러한 오류의 예를 보여준다.

형태소 단위 색인법은 형태소 해석이나 구문 해석 과정에서 복잡한 규칙을 요구한다. 앞에서 언급된 오류들에 대처하기 위한 형태소 해석의 규칙은 자연 복잡해질 수 밖에 없으며, 문서 내의 어절이나 문장들은 다양한 형태의 구조를 갖고 있고 예외적인 상황도 많이 발생할 수 있어 복잡한 구문 해석 규칙이 요구된다. 예를 들어, 문장 구성 성분들의 순서가 도치되거나 구성 성분의 역할을 결정하는 조사나 어미 등의 생략은 구문 해석을 어

〈표 2〉 형태소 해석 오류의 예

원 인	입력 어절	잘못된 형태소 해석 결과
애매한 형태소 해 석	10년 형을 선고함	선/보통명사 + 고함/보통명사 선고/보통명사 + 함/보통명사
미등록어	가내공장으로	가/보통명사 + 내/보통명사+ 공장/보통명사 + 으로
비문법적 어절	추출하였다고하자	추출하였다고/동작성보통명사+ 하/파생접미사 + 자/연결 어미
띄어쓰기오류	공유할수있는	공유할수있/보통명사 + 는/주격조사

렵게 한다.

3.2 *n*-Gram 기반의 색인 방법

본 절에서는 어절 단위 색인법에서의 복합 명사 띄어쓰기 문제를 완화할 수 있으며, 형태소 단위 해석에서와 같은 복잡한 문장 해석 규칙이나 언어 정보의 개발을 요구하지 않는 색인 방법을 제안한다. 제안하는 방법은 기존의 어절 단위 색인법과 ***n*-gram** 방법의 결합에 의해 구성된다. 〈표 3〉은 제안하는 방법의 색인 과정을 간략히 보여주며, 각 단계에 대한 자세한 설명은 다음과 같다.

(1) 문서나 질의를 색인하기 위해 먼저 빈칸, 마침표, 쉼표, 따옴표 등을 구분자로 하여 모든 어절들을 추출한다.

(2) 불용어 리스트를 이용하여 색인어로서 무의미한 어절들을 제거한다. 한글에서는 단어에 다양한 종류의 조사나 어미 등이 붙을 수 있고 복합어와 동사의 활용이 다양하므로 불용어 선정에 신중을 기해야 한다.

(3) 나머지 어절들에 대해 최장 일치법을 이용하여 비색인 분절을 절단한다. 비색인 분절은 단일 조사(-가, -이, -를, -으로, -부터), 복합 조사(-으로부터, -에서부터), 조사, 어미, 접미사 등이 결합된 다양한 형태의 음절들을 포함한다. 예를 들면, 다음과 같은 어절들에서 '색인' 뒤에 오는 모든 문자열이 여기에 포함된다.

*색인**을***	*색인**하여***	*색인**하였는데***
*색인**되어***	*색인**되었으니***	*색인**임을***
*색인**이기에***	*색인**이라고***	*색인**이지만***

(4) 생성된 각각의 색인 분절에 대해 ***n*-gram** 방법을 적용한다. ***n*-gram**이란 인접한 n개의 음절을 말한다(Cavnar 1994 ; Damashek 1995). 예를 들면, '프로그래밍'이란 어절에 대해 2-gram은 '프로', '로그', '그래', '래밍'이며, 3-gram은 '프로그', '로그래', '그래밍'이다. 색인 분절의 음절 수가 n보다 큰 경우에는 색인 분절을 여러 개의 ***n*-gram**들로 분리하고, 작은 경우에는 색인 분절 전체를 하나의 ***n*-gram**으로 취한다.

〈표 4〉는 제안하는 방법을 이용한 색인 과정의 예를 보여준다.

3.3 *n*-Gram 기반 색인 방법의 장단점

제안하는 ***n*-gram** 기반의 색인 방법은 검색 효과의 측면에서 다음과 같은 장점을 갖는다.

〈표 3〉 *n*-Gram 기반의 색인 과정

단계 1 : 문서나 질의 내의 모든 어절들을 인식한다.
단계 2 : 불용어를 제거한다.
단계 3 : 각 어절에서 비색인 분절들을 절단한다.
단계 4 : 나머지 색인 분절을 ***n*-gram**들로 분할하여 색인어로 선정한다.

〈표 4〉 *n*-Gram 기반 색인 방법의 예 (2-Gram)

내년 중반부터 정보검색서비스가 실시된다.
단계 1 : 문장 내의 어절 인식
내년, 중반부터, 정보검색서비스가, 실시된다
단계 2 : 불용어 제거
정보검색서비스가, 실시된다
단계 3 : 비색인 분절의 절단
정보검색서비스, 실시
단계 4 : 2-Gram의 적용
정보, 보검, 검색, 색서, 서비, 비스, 실시

(1) ***n***-gram 기반의 색인법은 어절 단위 색인법을 이용할 때의 절단 오류로 인한 파급 효과를 완화한다. 예를 들면, 어절 '벨기에로서는'과 '벨기에'는 어절 단위 색인 과정에서 '벨기에'와 '벨기'로 색인된다. 여기에 2-gram 방법을 적용하면 모두 '벨기'의 공통된 색인어가 생성된다.

(2) 제안하는 방법은 복합 명사의 띄어쓰기 문제를 완화한다. 예를 들면, 아래와 같은 문서 d_1, … , d_5와 질의 q_1, q_2가 있다고 가정하자.

d_1 : 과학기술정보 유통의
d_2 : 과학기술 정보유통의
d_3 : 과학 기술 정보 유통의
d_4 : 과학기술 분야의 정보를 유통하기 위한
d_5 : 과학과 기술의 정보를 유통하기 위한
q_1 : 과학기술정보유통에 관한
q_2 : 과학 기술 정보 유통에 관한

2-gram 기반의 색인 방법은 이들 문서와 질의에 대해 다음과 같은 색인어들을 생성한다.

d_1 : {과학, 학기, 기술, 술정, 정보, 유통}
d_2 : {과학, 학기, 기술, 정보, 보유, 유통}
d_3 : {과학, 기술, 정보, 유통}
d_4 : {과학, 학기, 기술, 분야, 정보, 유통}
d_5 : {과학, 기술, 정보, 유통}
q_1 : {과학, 학기, 기술, 술정, 정보, 보유, 유통}
q_2 : {과학, 기술, 정보, 유통}

이와 같은 경우 질의 q_1과 q_2의 복합 명사 띄어쓰기가 서로 다르지만, 유사도를 계산하는 벡터 공간 모델에서 검색을 수행할 때, 모든 문서들은 두 질의에 대하여 높은 유사도를 갖는 문서로서 검색될 가능성이 크다.

(3) 한글 문서들을 살펴보면 아래의 예와 같이 단일 명사의 뒤에 한 글자의 명사가 붙거나 또는 파생 접사가 붙어서 형성된 명사들을 많이 발견할 수 있다. 형태소 단위 색인법에서는 이러한 명사를 보통 단일 형태소로 취급

하여 색인어로 추출한다.

가공기 가공력 가공도 가공량 가공면
가공물 가공법 가공부 가공비 가공사
가공성 가공상 가공수 가공압 가공업
가공열 가공용 가공률 가공재 가공칩
가공품 가공형 가공자 가공학 …

이러한 경우 '가공'의 질의가 입력되면, 관련된 많은 문서들이 검색되지 않을 수 있다. 제안하는 색인 방법은 이와 같은 경우에 관련 문서의 검색을 도와 준다.

(4) 제안하는 색인 방법은 철자 오류나 일관성이 없는 외래어 표기 문제를 적절히 극복할 수 있다. 예를 들면, 문서 d_1이 '정보검섹'으로 잘못 표기된 어절을 포함하고, 사용자는 '정보검색'으로 질의 q_1을 입력한다고 가정하자. 2-gram 기반의 색인법은 문서 d_1과 질의 q_1에 대해 각각 다음과 같은 벡터 표현을 형성한다.

d_1 : {(정보, w_1), (보검, w_2), (검섹, w_3)}
q_1 : {(정보, w_4), (보검, w_5), (검색, w_6)}

따라서 문서에 '검섹'의 철자 오류가 있더라도 문서는 질의의 결과로 검색될 가능성이 크다. 서로 다른 외래어 표기의 문제도 이와 유사하다. 사용자마다 'database'를 '데이터베이스'로 표기하기도 하고 '데이터베이스'로 표기하기도 한다. 어떤 식으로 문서에 표기되어 있든 n-gram 기반의 색인법을 이용하는 시스템에서는 서로 다른 표기법이 사용된 문서가 비슷한 수준의 유사도를 갖고 검색될 가능성이 크다.

제안하는 n-gram 기반의 색인 방법의 단점은 다음과 같다.

(1) 의미 없는 n-gram의 생성으로 인해 질의에 부적합한 문서들이 검색될(false match) 가능성이 있으며, 특히 가중치 기법과 관련하여 이들 부적합 문서들이 상위의 순위를 부여받을 수 있다. 예를 들어, 다음과 같은 문서 d_1과 질의 q_1이 있다고 가정하자.

d_1 : 자방친 및 화분친에 따라 감자 반수체 유기효율이 컸으며
q_1 : 배기관 형상에 따른 2 행정 기관의 소기효율 및 성능 예측

이때 문서 d_1의 '유기효율'과 질의 q_1의 '소기효율'에 대해 제안하는 색인 방법은 각각 {유기, 기효, 효율}과 {소기, 기효, 효율}의 색인어를 형성한다. 여기에서 '기효'가 일치하므로 문서 d_1은 질의 q_1에 관련이 없는데도 검색 결과로서 출력될 수 있으며, 만일 '기효'가 높은 가중치 값을 부여받는다면 문서 d_1과 질의 q_1 사이의 유사도가 커져 문서 d_1이 상위의 순위를 부여받을 수 있다. 따라서 이러한 문제를 해결할 수 있는 처리 방안이 고려되어야 한다.

(2) 제안하는 색인 방법에서는 추출되는 색인어의 수가 많아지며, 이를 위해 필요한 부가적인 저장 공간이 늘어날 수 있다.

4. 성능 평가

4.1 성능 평가 환경 및 실험 자료

본 논문에서는 *n*-gram 기반의 색인 방법과 기존의 색인 방법들의 성능을 비교하기 위해 SMART 시스템을 이용하였다(Salton 1971). SMART 시스템은 하버드와 코넬 대학에서 35년간에 걸쳐 개발된 시스템으로, 문서와 질의를 색인어들의 벡터로서 표현하는 벡터 공간 모델을 기반으로 한다. 즉, 색인 과정을 통해 추출된 색인어들에 가중치를 부여함으로써 문서와 질의 표현을 위한 벡터를 형성하고, 형성된 벡터를 이용하여 문서와 질의 사이의 유사도를 계산한다. 그리고 계산된 유사도에 따라 질의에 적합한 문서들을 검색하고, 검색된 문서들 사이의 순위를 결정한다.

정보 검색에 관한 많은 연구들은 가중치 부여 기법을 구성하는 데 있어서 색인어의 출현 빈도(term frequency), 장서 빈도(collection frequency), 정규화(mormalization)의 세 가지 요소를 고려한다(Lee 1995 ; Salton 1988). 출현 빈도는 문서 내에서 자주 출현하는 색인어에 보다 높은 가중치를 부여한다. 장서 빈도는 전체 문서들 중에서 적은 수의 문서에 출현하는 색인어에 보다 높은 가중치를 부여한다. 그리고 정규화 요소는 데이터베이스내의 모든 문서 벡터들의 길이를 일치시키는 요소로서, 작은 크기의 문서들이 문서값 계산에 있어 불공정하게 취급되는 것을 피하도록 한다. 〈표 5〉는 각각의 구성 요소에 대해 잘 알려진 공식들을 보여준다.

문서와 질의를 표현하는 벡터의 색인어들에 가중치를 부여하는 방법은 이미 언급된 세 가지 요소의 조합으로 구성된다. 예를 들어, lnc · ltc는 문서와 질의 벡터의 색인어들에 대해서 각각 lnc와 ltc의 가중치 기법을 적용함을 의미한다. 즉, 출현 빈도의 로그 값을 코사인 정규화함으로써 문서 벡터의 색인어들에 가중치를 부여하고, 출현 빈도와 역 문헌 빈도(inverse document frequency)를 곱한 값을 코사인 정규화함으로써 질의 벡터의 색인어들에 가중치를 부여한다. 본 논문에서는 문서와 질의 벡터를 위해 얻을 수 있는 128가지 조합의 가중치 부여 기법들을 성능 비교에 사용한다.

성능 평가에 사용된 실험 자료는 자동 색인기의 성능 시험을 위해 구축된 KT Test Set이다(김성혁 외 5인 1994). 여기에는 정보과학회논문지, 1993 한국정보과학회 학술발표대회 논문집, 정보관리학회지에 수록된 논문들로 구성된 1,053개의 문서들과 30개의 질의가 포함되어 있다. 모든 문서는 국문 및 영문 저자, 서명, 서지 사항, 초록, 분류 번호, 색인어 등 18개의 항목을 지니고 있으며, 각 질의에 대한 적합 문서들이 제시되어 있다.

4.2 성능 비교 분석

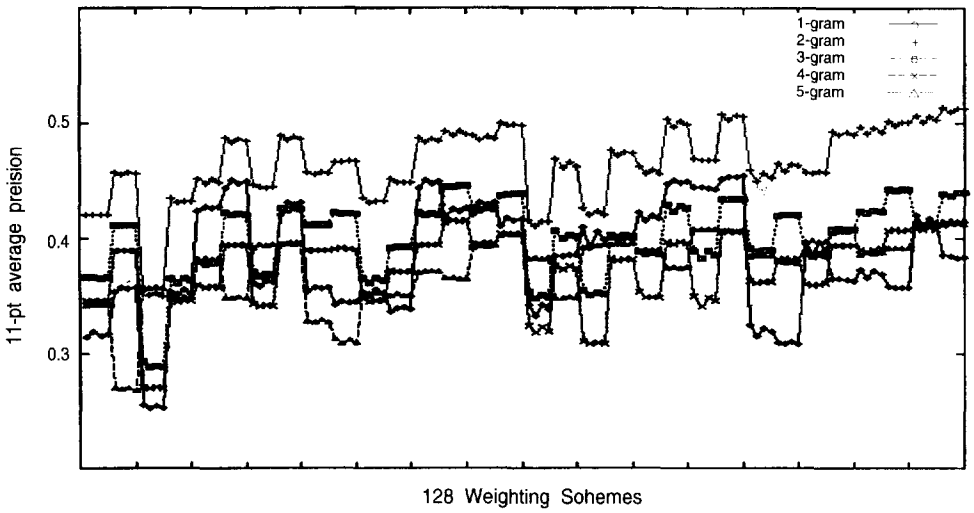
본 절에서는 *n*-gram 기반의 색인 방법과 기존의 한글 자동 색인 방법들의 성능을 검색 효과와 저장 공간의 측면에서 평가한다. 각 색인 방법을 적용할 때의 검색 효과를 측정하기 위해, KT Test Set의 모든 문서들의 서명과 초록의 항목에 대해 색인을 수행하고, 128개의 가중치 부여 기법을 바꾸어 적용하여 30

〈표 5〉 색인어 가중치 부여 기법의 구성 요소

출현빈도(term frequency)	
b 1.0	색인어의 출현 빈도를 무시하고 벡터를 구성하는 색인어에 1의 가중치 부여
n tf	문서나 질의 내에서 색인어의 출현 빈도
$a\ \ 0.5 + 0.5\frac{tf}{\max tf}$	보강된 정규화 출현 빈도(*tf*를 max *tf*로 나누고, 그 결과가 0.5 ~1.0의 값을 갖도록 정규화)
$l\ \ \ln tf + 1.0$	색인어의 출현 빈도에 로그 함수 적용
장서 빈도(collection frequency)	
n 1.0	색인어의 출현 빈도(*b, n, a, l*)만으로 가중치 생성
$t\ \ \ln\frac{N}{n}$	색인어 출현 빈도와 역문헌 빈도를 곱한다(N은 전체 문서들의 수이며, *n*은 그 색인어를 포함하고 있는 문서들의 수이다.)
정규화(normalization)	
n 1.0	출현 빈도와 장서 빈도만으로 유도된 가중치를 사용
$c\ \ \frac{1}{\sqrt{\sum_{\text{vector}} w_i^2}}$	유클리디안 벡터 길이를 이용한 코사인 정규화

개의 질의를 테스트하였다.

정보 검색 시스템의 검색 효과는 일반적으로 재현율과 정확률로써 평가된다(Salton et al. 1983). 재현율(recall)은 문서 집합에서 사용자가 원하는 문서를 어느 정도 검색하였는가를 나타내고, 정확률(precision)은 검색된 문서들 중에서 사용자가 원하는 문서가 얼마나 포함되어 있는가를 나타낸다. 예를 들어, 전체 문서 집합에 200개의 문서가 저장되어 있고, 이 문서 집합 속에 사용자의 질의에 관련된 문서가 5개 있다고 가정하자. 이때 사용자가 검색 시스템을 사용하여 6개의 문서를 검색하였고 검색된 문서 중에서 4개의 문서가 질의에 관련된 문서라고 하면, 재현율과 정확률은 각각 0.8과 0.67이 된다. 문서 순위 결정 방법을 제공하는 검색 시스템은 보간 기법을 사용하여 고정된 재현율에 대한 정확률을 계산할 수 있다. 본 논문에서는 고정된 11개의 재현율에 대한 모든 질의의 정확률을 평균한 값을 나타내는 11-포인트 평균 정확률을 이용하여 검색 효과를 측정하였다.



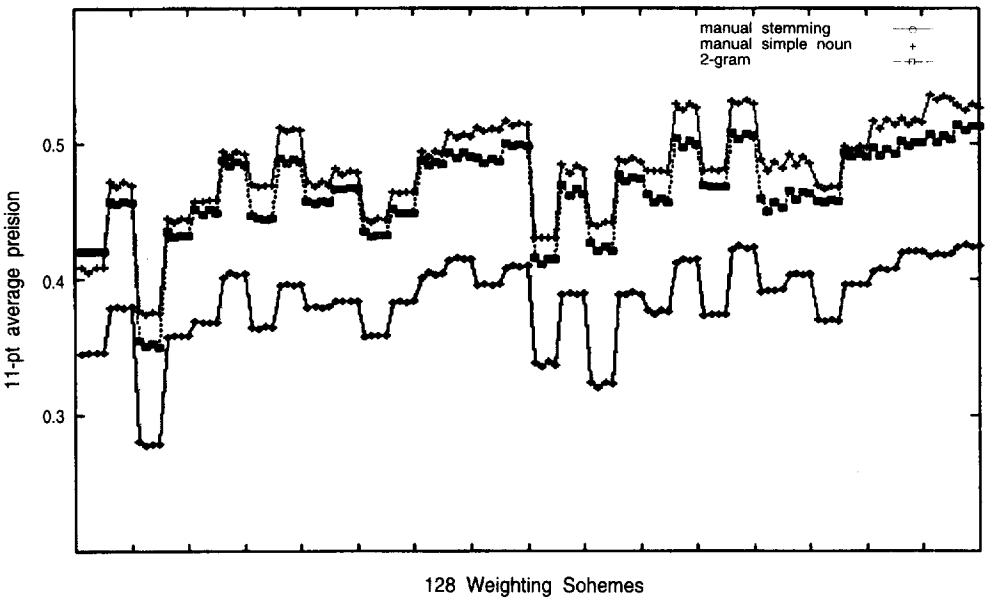
〈그림 1〉 *n*-Gram 기반 색인법의 검색 효과 비교(*n*은 1~5)

4.2.1 n-Gram기반 색인법의 검색효과

〈그림 1〉은 제안하는 ***n*-gram** 기반의 색인 방법으로 색인을 수행했을 때의 11-포인트 평균 정확률을 보여준다. 이 그림은 2-gram 기반의 색인 방법을 적용할 때 가장 높은 검색 효과를 제공함을 보여준다. 1-gram 기반의 색인법을 적용할 경우 일반적으로 2음절보다 1음절이 갖는 애매성이 크기 때문에 부적합한 문서들이 과다하게 검색될 수 있다. 예를 들면, 문서 ***d***가 "기존 알고리즘들의 복잡도를 계산하고 이를 비교 분석하여"라는 문장을 포함하고 있고, 질의 ***q***는 어절 "분산교환망"을 포함한다고 가정하자. 이때 문서 ***d***는 분명히 질의 ***q***와 관련이 없으나, 1-gram 기반의 색인법을 사용할 경우 '산'과 '분'이 색인어로 추출되어 질의 ***q***의 검색 결과로 출력될 수 있다. 따라서 이러한 부적합 문서들의 과다한 검색으로 인해 1-gram 기반의 색인법을 적용할 때 검색 효과가 떨어짐을 추측할 수 있다. 3-gram, 4-gram, 5-gram 기반의 색인법을 적용할 경우에는 음절수로 인한 애매성이 줄지만, 다음과 같은 이유로 2-gram 기반의 색인법에 비해 검색 효과가 저하됨을 추측할 수 있다. 예를 들면, 문서 ***d***와 질의 ***q***가 각각 '정보검색'과 '정보 검색'을 포함한다고 가정하자. 3-gram 기반의 색인법은 문서 ***d***와 질의 ***q***를 위한 색인어로 각각 {정보검, 보검색}과 {정보, 검색}을 추출한다. 따라서 동일한 어절에 대해 공통된 색인어가 생성되지 않는다.

4.2.2 한글 자동 색인법의 검색효과

일반적으로 어절 단위 색인과 형태소 단위



〈그림 2〉 한글 자동 색인 방법들의 검색 효과 비교

색인은 수작업으로 수행할 때 가장 정확한 결과를 얻을 수 있다. 본 논문에서는 이러한 가정하에 수작업으로 비색인 분절들을 제거함으로써 어절 단위 색인을 수행하고, 문서로부터 단일 명사들을 수작업으로 추출함으로써 형태소 단위 색인을 수행하였다. 〈그림 2〉는 2-gram 기반 색인 방법의 검색 효과와 수작업에 의한 어절 단위 색인, 그리고 수작업에 의한 단일 명사 추출시의 검색 효과를 보여준다. 그림에서 2-gram 기반의 색인법은 수작업에 의한 어절 단위 색인법보다 높은 평균 정확률을 보여주며, 형태소 단위 색인법과 거의 유사한 평균 정확률을 보여주고 있다.

본 논문에서는 이러한 검색 효과의 차이가 통계적으로 유의한지를 검증하기 위해 paired t-test를 실시하였다. 색인어 가중치 기법으로서 대부분의 색인법에서 높은 검색 효과를 보인 atc · atc 가중치 기법이 사용되었으며, 그 결과 2-gram 기반의 색인법과 수작업에 의한 어절 단위 색인법간의 검색 효과의 차이에 대해 유의 수준 0.05에서 확률값이 1.68E-09로 계산됨으로써 두 방법간에 차이가 있는 것으로 나타났다. 반면에 2-gram 기반의 색인 방법과 수작업에 의한 형태소 단위 색인법간의 차이에 대해서는 확률값이 0.1114로 계산되어 두 방법간에 검색 효과의 차이가 없는 것으로 나타났다. 결국 이러한 사실은 2-gram 기반의 색인법이 어절 단위 색인법보다 복합 명사의 띄어쓰기 문제를 더욱 잘 해결한다는 것을 보여주며, 무의미한 2-gram들의 생성으로 인해 발생하는 검색 효과의 저하가 크지 않다는 것을 의미한다.

〈표 6〉 명사 어절들의 출현 비율

명사 어절	출현 비율
1 음 절	4.93%
2 음 절	62.67%
3 음 절	17.02%
4 음 절	11.01%
5 음 절	2.81%
기 타	1.56%

4.2.3 저장 공간의 비교

색인어를 위한 저장공간은 추출된 색인어의 수와 관련이 있다. 본 논문에서는 각각의 색인 방법에서 요구하는 저장 공간의 비교를 위해 KT Test Set 내의 문서들 중에서 임의의 문서들에 대한 색인어의 수를 조사하였는데, 높은 검색 효과를 보인 2-gram 기반의 색인 방법은 1-gram을 제외한 다른 색인 방법들보다 최대 약 1.5배의 색인어들을 생성하였다. 이러한 완만한 증가는 한글 문서에 나타나는 명사들 중에 대부분이 2음절이나 3음절이기 때문인 것으로 조사되었다. 〈표 6〉은 KT Test Set 문서들에 포함된 명사들의 출현 비율을 보여준다.

5. 결 론

정보 검색 시스템의 목적은 단순히 사용자 질의를 만족하는 문서들의 검색뿐만 아니라, 문서와 질의 사이의 유사도 계산을 통해 검색되는 문서들에 순위를 부여함으로써 사용자들이 필요로 하는 정보를 얻는 데 소모되는 시간을 최소화하는 역할도 포함한다. 문서 순위를 결정할 수 있는 벡터 공간 모델은 문서와 질의를 색인어의 벡터로서 표현하고, 두 벡터들의 내적을 통해 유사도를 계산한다. 따라서 색인 방법은 검색되는 문서들의 순위 결정에 영향을 주는 중요한 요소 중의 하나이다.

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

본 논문에서는 어절 단위 색인법과 *n*-gram 방법을 결합한 *n*-gram 기반의 색인법을 제시하였다. 제안하는 색인 방법은 복합 명사의 띄어쓰기 문제를 완화하며, 형태소 단위 색인법에서와 같은 언어 정보의 개발도 거의 요구하지 않는다. 또한 비색인 분절의 절단 오류나 하나의 단일 형태소로 취급되는 복합 명사들로 인한 문제를 완화할 수 있고, 실제 문서에서 많이 발견되는 철자 오류나 일관성이 없는 외래어 표기 문제에도 대처할 수 있다.

본 논문에서는 벡터 공간 모델을 지원하는 SMART 시스템과 KT Test Set를 이용하여 제안하는 색인 방법과 기존의 한글 자동 색인 방법들의 검색효과를 평가하였다. 그 결과, 제안하는 *n*-gram 기반의 색인법 중에서 2-gram 기반의 색인법이 가장 좋은 검색 효과를 나타냈다. 그리고 2-gram 기반의 색인법은 수

작업에 의한 어절 단위 색인법보다 높은 검색 효과를 보였으며, 수작업으로 단일 명사를 추출한 형태소 단위 색인법과 유사한 검색 효과를 보였다.

참 고 문 헌

Cleverdon, C.W. 1984. Optimizing Convevient On-line Access to Bibliographic Databases, Information Service and Use, 4:1, 37-47.

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

Damashek, M. 1995. Gaushing Similarity with *n*-Grams: Language -Independent Categorization of Text, Science, Vol. 267, 843-848.

Lee, J.H., Kim, M.H., Lee, Y.J. 1994. Ranking Documents in Thesaurus Based Boolean Retrieval Systems, Information Processing & Management, Vol.30, No.1, 79-91.

Lee, J.H. 1995. Combining Multiple Evidence from Different Properties of Weighting Schemes, Proceedings of the 18th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval.

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

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

Salton, G. and McGil, M.J. 1983. Introduction to Modern Information Retrieval, McGraw-Hill, Inc.

Salton, G. 1986. Another Look at Automatic Text Retrieval, Communications of ACM, 29:7, 648-656.

Salton, G. 1987. Historical Note: The Past Thirty Years in Information Retrieval, Journal of the American Society for Information Science, Vol.38, No.5.

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

Salton, G. 1989. Automatic Text Pro-

cessing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison Wesley.

강승식, 권혁일, 김동렬. 1995. 한국어 자동 색인을 위한 형태소 분석 기능, 한국정보과학회 봄 학술발표논문집, 제22권 1호, 930-932.

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

김영환. 1982. 한글 한자 혼용문의 자동 색인 시스템, 한국과학기술원 석사학위논문.

안현수. 1986. 한글 문헌의 자동 색인에 관한 실험적 연구, 정보관리학회지, 제3권 2호, 108-306.

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

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

정진성. 1992. 단일 문서내에서의 언어 및 통계 정보를 이용한 자동 색인, 한국과학기술원 석사학위논문.

최기선. 1991. 구문 및 의미 분석을 통한 한국어 자동 색인, 정보관리학회지 제8권 2호.

한성현. 1991. 구문해석을 이용한 색인어 자동 추출 시스템의 설계와 구현, 한국과학기술원 석사학위논문.