工學碩士 學位論文

한글 정보 검색을 위한 혼합 *n*-그램 기반의 색인 방법

A*n* I*n*dex*ing* Method Based o*n* the Mixed *n*-Gram for Korea*n* I*n*formatio*n* Retrieval

指導敎授 金 載 熏

2004年 8月

韓國海洋大學校 大學院

컴 퓨 터 工 學 科

鄭 昌 溶

本 論文을 鄭昌溶의 工學碩士 學位論文으로 認准함

委員長　工學博士　朴　烋　讚　印

委　員　工學博士　柳　吉　洙　印

委　員　工學博士　金　載　熏　印

2004年　7月

韓國海洋大學校 大學院

컴퓨터工學科　鄭　昌　溶

목 차

표 목차

그림 목차

An Indexing Method Based on the Mixed n-Gram for Korean Information Retrieval

Chang-yong Jung

Department of Computer Engineering, Graduate School,
Korea Maritime University, Busan, Korea

Abstract

In Korean information retrieval systems, several indexing methods are proposed such as morpheme-based, word-phrase-based, and n-gram-based. An n-gram-based indexing method is widely used among these methods where n is 2 or 3. The method is very simple, but outperforms others in precision and recall, which are basic measures for evaluating information retrieval systems. On the other hand, the method generates too many index terms that contain meaningless terms, and then the size of index files is huge. To relieve this problem, this paper proposes a new indexing method, which chooses between 2 and 3-grams according to probabilistic criteria for removing the meaningless terms. It is called a mixed n-gram indexing method. The t-score is used for the criteria for choosing between 2 and 3-grams. Also this paper describes a new stemming method for speed-up of Korean indexing systems by using a greedy algorithm.

For experiments, KT-SET and KEMONG-SET are used for reference test collections in Korean and storage and retrieval components of Lemur information retrieval toolkit 2.2 are used. Experiments have shown that the proposed method is not inferior to others in recall and precision, but is superior to others in the number of index terms.

제 1 장 서 론

정보 검색 시스템(information retrieval system)은 정보 이용자가 필요로 하는 정보를 수집하여 검색하기 쉬운 형태로 저장하여 두었다가 정보에 대한 요구가 발생했을 때, 적합한 정보를 검색하여 제공하는 시스템이다. 일반적으로 정보 검색 시스템은 색인과 저장과 검색의 세 부분으로 이루어진다. 특히 정보 검색 시스템의 성능에 가장 큰 영향을 미치는 것은 문서에 대한 색인어 선택 방법과 색인어 가중치를 결정하는 검색 모델이다(Baeza-Yates and Ribeiro-Neto, 1999).

색인(indexing)이란 문서에서 정보 검색의 대상이 되는 색인어(index term) 집합을 표현하는 과정이다. 정보 검색 시스템에서 사용자 요구에 적합한 문서(relevant documents)를 검색할 수 있다는 것은 적합 문서에 대해 색인되어 있다는 것이다. 즉, 적합 문서에 대해 얼마나 정확히 색인되어 있느냐가 정보 검색 시스템의 성능에 영향을 미칠 수 있다.

저장(storing)이란 추출된 색인어와 색인어를 포함한 문서의 위치, 색인어의 가중치를 기록하는 과정이다. 색인어를 저장하는 파일구조는 역파일(inverted file), 요약 파일(signature file) 등이 있다(Blumer *et al.* 1987). 이들 중에 어떤 구조로 색인어를 저장하느냐에 따라 저장 공간의 효율과 검색 속도에 영향을 줄 수 있다.

검색(retrieving)이란 질의어와 저장된 색인어를 비교하여 원하는 문서를 찾는 과정이다. 또한 사용자의 요구에 더 적합한 문서를 상위에 배치하는 방법들이 사용되는데 이런 방법을 순위화(ranking)라고 한다. 순위화하기 위한 방법으로 색인어 가중치 계산을 이용한다. 색인어 가중치 계산 방법에 따라 사용자에게 더 적합한 문서가 상위에 배치될 수도 있고 그렇지 않을 수도 있다.

앞서 살펴본 바와 같이 정보 검색 시스템의 성능에 영향을 미치는 요인으로는 적절한 색인어 선택과 저장 구조의 선택, 색인어 가중치 부여 방법 등이 있다(Baeza-Yates and Ribeiro-Neto, 1999). 이들 중에서 본 논문은 색인 방법에 해당하므로 색인 방법에 대해 좀 더 구체적으로 살펴보고자 한다.

색인 방법은 크게 언어적 성질을 이용한 방법과 비언어적 성질을 이용한 방법이 있다. 언어적 성질을 이용한 방법에는 형태소 단위 색인 방법과 어절 단위 색인 방법이 있으며, 비언어적 성질을 이용한 방법으로는 *n*-그램 색인 방법이 있다(이준호 외, 1996; 강승식, 2002).

형태소 단위 색인 방법과 어절 단위 색인 방법은 색인할 어절에서 동일 어근(stem)을 찾아 색인어로 선택하는 방법이다. 두 방법의 차이는 형태소 분석을 이용할 경우에는 문법적 관계를 이용하여 동일 어근을 찾아 불규칙 활용된 용언에 대해 그 원형을 복구하는 반면, 어절 단위 색인 방법은 단순히 조사, 어미, 접미사만 제거하여 색인한다. 형태소 단위 색인 방법은 어절들 사이의 문법적 관계를 밝혀야 하기 때문에 사전(dictionary) 등의 추가적인 정보가 필요하고, 사전에 따라 성능에 영향을 받기도 한다. 위의 두 방법은 다른 색인 방법에 비해 색인어 개수가 적다는 장점이 있지만 복합 명사를 띄어 쓰는 차이에 따라 검색 성능이 떨어지는 문제를 가지고 있다(이준호 외, 1996; 강승식, 2004).

n-그램 색인 방법은 언어적 특성을 고려하지 않고 문장을 단순히 *n*개의 음절로 분리하는 방법이다. *n*이 2인 경우는 문장을 두 음절 단위로 분리하여 색인어를 추출하는데 이 방법을 2-그램 색인 방법이라고 한다. 마찬가지로 *n*이 3인 경우는 문장을 세 음절 단위로 분리하여 색인어를 추출하는데 이 방법을 3-그램 색인 방법이라고 한다. *n*-그램 색인 방법은 복합 명사의 띄어쓰기 문제가 자연스럽게 해결된다. 또한, 형태소 단위 색인에서와 같은 복잡한 문장 해석 규칙 등을 필요로 하지 않는다(이준호 외, 1996). 그러나 추출되는 색인어의 수가

증가하게 되어 검색 효율과 저장 공간의 효율이 저하된다는 단점이 있다(강승식, 2004).

이러한 문제를 보완하기 위해 본 논문에서는 혼합 *n*-그램 색인 방법을 제안한다. 이 방법은 한국어 단어의 약 80%가 2~3 음절로 구성되어 있다(이준호 외, 1996)는 점에 착안하여 주어진 어절에 따라 2-그램 색인 방법과 3-그램 색인 방법을 선택적으로 사용한다. 실험 집합을 사용하여 제안된 방법으로 추출된 색인어의 수는 2-그램 보다 최대 40% 이상 줄었고, 검색 효과도 기존의 *n*-그램과 유사하거나 좋은 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 정보 검색 시스템의 개요와 한글 문서를 위한 색인 방법에 대해 기술하고, 3장에서는 제안된 혼합 *n*-그램 색인 방법과 새로운 동일 어근 추출(stemming) 방법에 대해 설명한다. 4장에서는 KT-SET과 KEMONG-SET을 사용하여 제안된 혼합 *n*-그램 색인 방법과 기존 한글 색인 방법들의 성능을 평가하고 비교하고 분석한다. 마지막으로 5장에서는 결론에 대해 제시한다.

제 2 장 정보 검색 시스템과 한글 문서 색인 방법

본 장에서는 정보 검색 시스템의 구성과 한글 문서 색인 방법에 대해 살펴본다.

2.1 정보 검색 시스템

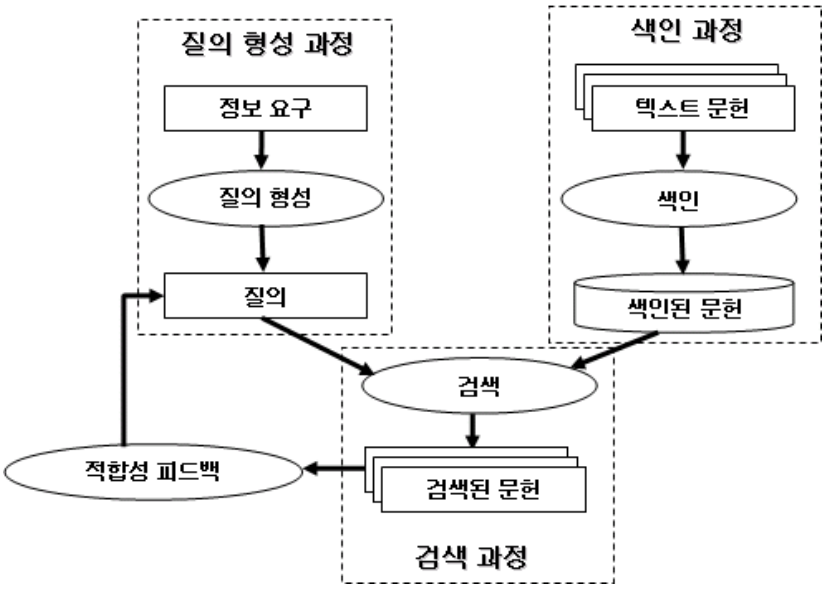

그림 2.1 정보 검색 시스템의 개관
Fig. 2.1 Overview of information retrieval system

정보 검색 시스템은 사용자가 원하는 정보에 적합한 문서를 저장된 데이터베이스에서 검색하는 시스템이며, 색인, 저장, 검색 기능으로 구성된다. 정보 검색 시스템의 개관은 그림 2.1과 같다. 색인 과정은 문서를 색인어 집합으로 표현하는 과정을 말한다. 이렇게 색인된 문서는 저장 과정에서 색인된 문헌 형태로 저장 공간에 저장된다. 질의 형성 과정은 색인과 마찬가지로 색인 과정에서 사

용한 색인 방법을 이용해 사용자 질의를 질의어 집합으로 표현하는 과정이다. 검색 과정은 색인어와 질의어를 비교하여 원하는 문서를 찾는 과정이다. 적합성 피드백(relevance feedback)은 검색된 문서에 대해 사용자가 추가적인 정보를 입력하여 재검색하는 과정을 말한다.

가. 색인

색인은 문서를 정보 검색의 대상이 되는 색인어(index term) 집합으로 표현하는 과정이다. 그림 2.2는 가상의 문서에 대한 색인 과정을 보이고 있다. 색인 과정은 색인어 추출과 색인어 저장 단계로 구분된다. 색인어 추출 단계는 주어진 문서에서 색인어를 선택하고, 각 색인어에 대해 가중치를 구한다. 예제에서는 색인어의 출현 빈도를 가중치로 사용한 예이다. 색인어 저장 단계는 색인어를 포함하고 있는 문서와 가중치로 구성된 역파일을 생성한다. 이렇게 생성된 역파일을 이용하여 검색하게 된다.

일반적으로 한글 문서를 색인하는 방법에는 언어적 성질을 이용한 방법과 비언어적 성질을 이용한 방법이 있다. 각 방법에 대한 설명은 2.2절에서 다룬다.

문서 번호	문서 내용
001	한글 정보 검색 시스템이란 한글 문서를 위한 검색 시스템이다.
002	한글은 세종대왕이 만들었다.
003	학생 정보 시스템이 학생들에 의해서 완성되었다.
004	아버지는 학생 주임 선생님이시다.

색인어 추출 단계

문서 번호	색인어 및 가중치
001	(한글,2), (정보,1), (검색,2), (시스템,2) (문서,1)
002	(한글,1), (세종대왕,1), (만들,1)
003	(학생,2), (정보,1), (시스템,1), (완성,1)
004	(아버지,1), (학생,1), (주임,1), (선생님,1)

색인어 저장 단계

색인어	문서번호 및 가중치
검색	(001, 2)
만들	(002, 1)
문서	(001, 1)
선생님	(004, 1)
세종대왕	(002, 1)
시스템	(001, 2), (003, 1)
아버지	(004, 1)
완성	(003, 1)
정보	(001, 1), (003, 1)
주임	(004, 1)
학생	(003, 2), (004, 1)
한글	(001, 2), (002, 1)

그림 2.2 색인 과정의 예

Fig 2.2 An example of indexing steps

나. 검색

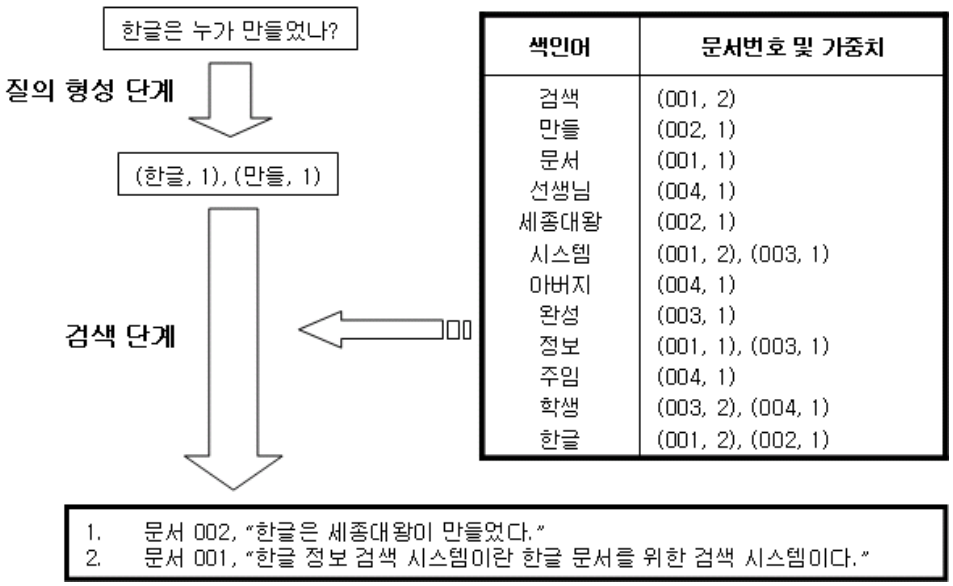

그림 2.3 검색 과정의 예

Fig. 2.3 An example of retrieval steps

검색은 데이터베이스에 저장되어 있는 색인어와 사용자의 질의를 비교하여 일치하는 문서를 찾아내는 과정이며, 크게 질의 형성 단계와 검색 단계로 이루어진다. 질의 형성 단계는 주어진 질의에 대해 문서에서 색인어를 추출하는 방법과 같은 방법으로 질의 문장에서 질의어를 선택하는 과정이다. 즉, 문서가 색인되는 과정과 동일한 방법으로 질의에 대해서도 중심 질의어와 가중치를 결정한다. 위의 예와 같이 "한글을 누가 만들었나?"라는 문장에 대해 질의 형성 단계를 거치면 '한글'과 '만들'이라는 질의어를 추출한다. 질의어의 가중치는 색인할 때와 마찬가지로 질의어 빈도수로 계산한다. 검색 단계에서는 이렇게 선택된 질의어와 가중치를 색인할 때 생성된 역파일과 비교하여 검색을 끝낸다. 검

색 결과 중에 사용자의 요구에 더 적합한 문서를 상위에 표시하기 위해서 다양한 방법을 사용하는데 대표적인 방법으로 벡터 모델(vector-space model)과 확률 모델(probabilistic model) 등이 있다.

2.2 한글 문서 색인 방법

색인 방법은 크게 언어적 성질을 이용한 방법과 비언어적 성질을 이용한 방법이 있다. 언어적 성질을 이용한 방법에는 형태소 단위 방법과 어절 단위 색인 방법이 있고, 비언어적 성질을 이용한 방법에는 *n*-그램 색인 방법이 있다.

영어와 같은 언어에서는 띄어쓰기 단위로 어절이 분리되어 있기 때문에 어절 단위 색인 방법으로 동일 어근을 추출(stemming)하여 색인할 수 있다. 반면에 일본어나 중국어의 경우에는 문장이 연속된 음절들의 나열로 이루어져 있기 때문에 색인어 추출 과정에서 띄어쓰기 문제를 해결하는 것이 큰 문제가 된다 (Nie and Ren, 1999). 이런 언어의 경우에는 형태소 분리를 이용하는 것이 일반적이다.

한글 문서의 경우에는 교착어라는 특성 때문에 일본어나 중국어처럼 형태소 단위 색인 방법이 적합하다. 또한, 영어의 경우처럼 어절이 띄어쓰기를 기준으로 나뉘어져 있고, 조사, 접미사, 어미를 분리하는 과정이 필요한데 이는 어절 단위 색인 방법과 유사하다. 이러한 특성으로 인해 언어적 성질을 이용한 색인 방법에서 한글 문서의 색인 방법은 형태소 단위 색인 방법과 어절 단위 색인 방법을 사용한다.

중국어나 일본어의 경우에는 띄어쓰기 문제로 인해 형태소 단위 색인 방법보다 *n*-그램 색인 방법을 이용하는 것이 더 나은 성능을 보인다고 보고되고 있다 (Nie and Ren, 1999). 한국어에서도 역시 *n*-그램 색인 방법이 좋은 성능을 나

타낸다고 보고되고 있다(이준호 외, 1996).

가. 언어적 성질을 이용한 방법

언어적 성질을 이용한 색인 방법은 사람이 단어를 인지하는 과정에 기반을 둔 색인 방법이다. 이러한 방법에는 다시 형태소 단위 색인 방법과 어절 단위 색인 방법이 있다.

형태소 단위 색인 방법은 문서의 내용을 대표하는 용어로부터 동일 어근(stem)을 추출하여 추출된 동일 어근들을 색인하는 방법이다. 색인 방법은 불용어(stopword) 제거, 조사와 접미사와 어미의 제거, 동일 어근 추출의 과정으로 이루어진다. '이', '그', '저' 등과 같은 대부분의 단음절과 문서간의 발생빈도가 아주 높은 단어들을 불용어라고 하고, 이들은 문서들 간의 분별력을 떨어뜨리기 때문에 제거하는 것이 효과적이다. 조사, 어미, 접미사 제거 과정은 체언 뒤의 조사, 접미사와 용언의 어미를 제거한다. 예를 들어, '용어로부터'라는 어절에서 조사, '로부터'를 제거하거나 '먹으니'라는 어절에서 어미, '으니'를 제거하는 과정이다. 형태소 단위 색인 방법에서 동일 어근 추출은 불규칙 활용 등으로 변형된 어간을 원형으로 복구하는 과정이다. 예를 들어, '아름답다', '아름다워', '아름다운'의 경우는 각 어절의 원형인 '아름답'이 색인된다.

형태소 단위 색인 방법에서는 추출된 색인어 수가 다른 색인 방법에 비해 최대 30%~50% 정도 감소되는 효과가 있지만(강승식, 2004), 형태소를 분석하는데 사전(dictionary)과 같은 언어적 정보가 필요하기 때문에 알려지지 않은 단어를 포함할 경우 분석 성능을 떨어뜨린다.

형태소 분석의 단점을 보완하기 위해 어절 단위 색인 방법이 등장하였다. 이는 색인할 문장을 어절 단위로 분리하고 분리된 어절에서 조사와 접미사와 어

미만을 제거하여 색인어로 선택한다. 이 과정에서는 불규칙 용언의 어간을 복원하지 않는다. 영어의 경우에는 같은 명사에 대해 단수와 복수의 표현이 서로 다른 색인어로 간주되기도 한다. 어절 단위 색인 방법은 사전 정보와 문법적 정보가 필요 없기 때문에 동일 어근을 추출하는데 많은 비용을 줄일 수 있다. 하지만 두 방법 모두 복합 명사의 띄어쓰기에 따른 검색 오류를 포함하고 있다. 예를 들어, '정보검색시스템'이라고 색인된 문서가 있을 때, 질의어 '정보 검색시스템'을 요구한다고 가정하자. 질의 형성과정에서 질의어는 '정보'와 '검색시스템'으로 나누어지고, 이들 질의어로 색인어 역파일에서 검색한다. 이때, '정보검색시스템'이라고 색인된 문서는 의미적으로 '정보 검색시스템'과 동일하지만 색인어와 질의어가 다르기 때문에 검색되지 않는다.

나. 비언어적 성질을 이용한 방법

비언어적 성질을 이용한 방법은 색인할 문장에 대해 문법적 관계를 고려하지 않고 음절 단위로 색인어를 추출하는 방법이다. 이러한 방법을 n-그램 색인 방법이라고 부르며, 문장을 연속된 n개의 음절로 나누어 색인어를 생성한다. n-그램 색인 방법에서 n의 수는 사용자가 임의로 지정할 수 있지만 한글에서는 명사 어절 중 60% 이상이 2음절로 이루어졌다는 특성 때문에 n=2일 때 가장 좋은 성능을 나타낸다고 보고되었다(이준호 외, 1996).

표 2.1 2-그램 색인 방법을 이용한 색인어 추출의 예
Table 2.1 Indexing examples using 2-gram indexing method

(a) '정보 검색 시스템'	'정보', '검색', '시스', '스템'
(b) '정보검색 시스템'	'정보', '보검', '검색', '시스', '스템'
(c) '정보검색시스템'	'정보', '보검', '검색', '색시', '시스', '스템'

예를 들어, n=2일 때(2-그램)의 색인어 추출은 표 2.1과 같다. 이 경우, '정보'라는 질의어가 주어지면 (a), (b), (c) 모두 찾아준다. 하지만 형태소 단위 방법이나 어절 단위 색인 방법에서는 (b)와 (c)의 예는 복합명사의 띄어쓰기 문제 때문에 검색하지 못한다.

이 방법은 색인 방법이 단순하고 복합명사의 띄어쓰기 문제를 해결할 수 있는 장점이 있는 반면, '보검', '색시' 같은 의미 없는 색인어가 다수 생성되어 부적합한 문서가 검색될 가능성이 있고, 색인어가 지나치게 증가한다는 문제점이 있다. 이상의 단점을 보완하기 위해 본 논문에서는 비언어적 성질을 이용한 방법의 하나인 2-그램과 3-그램을 이용한 혼합 n-그램 색인 방법을 제안한다.

제 3 장 한글 문서를 위한 혼합 *n*-그램 색인 방법

본 장에서는 기존의 한글 문서 색인 방법에서 나타난 복합 명사의 띄어쓰기로 인한 검색 오류와 의미 없는 색인어의 과다 생성 문제점을 보완하기 위한 혼합 *n*-그램 색인 방법과 본 논문에서 사용한 동일 어근 추출 방법을 소개한다.

3.1 동일 어근 추출 방법

한국어에서 동일 어근 추출(stemming)이란 문법적으로 체언 뒤의 조사를 제거하거나 용언의 어간을 복원하는 과정을 말한다. 한국어의 경우 표 3.1과 같이 크게 세 가지 유형으로 분류될 수 있다. 첫째 유형은 표 3.1의 a)와 같이 '체언+조사'로 이루어진 어절에서 체언만 추출하는 유형이고, 둘째 유형은 표 3.1의 b)와 같이 '용언+어미'로 이루어진 어절에서 형태소 분석을 통해 용언의 어간을 추출하고 불규칙 활용된 어간에 대해 원형을 복원하는 유형이다. 셋째 유형은 표 3.1의 c)와 같이 '용언+어미'로 이루어진 어절에서 용언의 어간을 추출하고 불규칙 활용된 용언에 대해서 원형을 복구하지 않는 유형이다.

표 3.1 한국어에서 동일 어근 추출의 예

Table 3.1 Stemming examples in Korean

a) 체언 + 조사
(a1) '정보검색시스템은' ⇒ '정보검색시스템'
(a2) '정보검색시스템을' ⇒ '정보검색시스템'
(a3) '정보검색시스템이란' ⇒ '정보검색시스템'

b) 용언 + 어미(용언의 원형을 복구한 경우)
(b1) '아름답다' ⇒ '아름답'
(b2) '아름다워' ⇒ '아름답'
(b3) '아름다운' ⇒ '아름답'

c) 용언 + 어미(용언의 원형을 복구하지 않은 경우)
(c1) '아름답다' ⇒ '아름답'
(c2) '아름다워' ⇒ '아름다'
(c3) '아름다운' ⇒ '아름다'

원형을 복원할 경우에는 형태소 분석 과정이 필요하다. 형태소 분석은 계산 과정이 복잡해서 분석 속도가 느려지며, 사전의 정보에 따라 오류를 포함하게 된다. 이를 보완하기 위해 본 논문에서는 기능어 사전과 한 어절의 길이와 그 어절에 포함된 기능어의 길이에 대한 통계를 이용하는 방법을 제안한다. 기능어(fu*nctio*nal word)란 조사와 같이 문장에서 단어의 역할을 결정하는 단어를 말하며, 명사, 동사, 형용사, 부사같이 문장에서 실질적 내용을 담고 있는 내용어(co*nten*t word)에 결합되어 사용된다. 본 논문에서는 특히 조사, 접미사 및 어미를 기능어라고 정의하였다. 형태소 분석에서 사용되는 사전은 신조어에 대해 유지되고 관리되어야 하지만, 기능어는 국어에서 신조어가 거의 생기지 않는 특징 때문에 유지하고 관리하는 추가 비용이 발생하지 않는다.

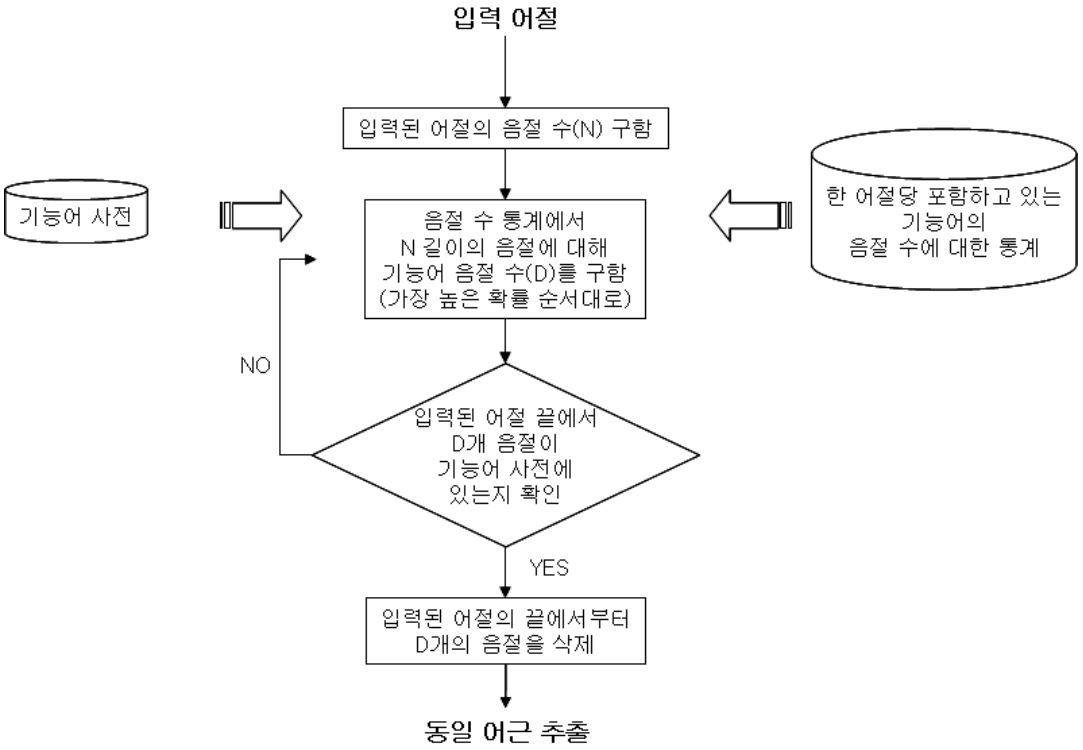

그림 3.1 제안된 동일 어근 추출 방법

Fig. 3.1 The proposed stemming method

본 논문에서 제안하는 동일 어근 추출 방법은 그림 3.1과 같다. 기능어 사전은 태깅된 코퍼스(문화관광부, 2003)를 이용하여 가능한 모든 경우의 기능어 목록을 추출하여 생성하였다. 표 3.2는 추출된 기능어 목록의 일부분을 나타낸 것이다.

표 3.2 기능어 사전의 일부분

Table 3.2 A part of functional word dictionaries

가	나	다
간	납니다	단
갈	나가	단다
가가	나게	단다는
가게	나고	단데데
가고까지	나그려	단들
가나	나나	단이
가는	나나이에요	달
가다	나나지요	답니까
가다는	나는	답니더
가당	나니	다가
가도	나다	다간
가라고	나다고	다가는
가라우	나던	다가도
가로	나던데	다가들
가를	나드라	다가들의
…	…	…

한 어절의 길이와 어절에 포함된 기능어의 길이에 대한 통계는 태깅된 코퍼스에서 한 어절의 길이와 그 어절에서 기능어로 태깅된 음절의 수를 구하고 각 경우에 대해 발생할 확률을 계산하였다. 예를 들어, 표 3.3에서 보면 어절의 길이가 2인 경우에는 그 어절에 기능어의 길이가 0개인 음절을 포함할 확률이 51% 정도이고, 어절의 길이가 3인 경우에는 그 어절에 기능어의 길이가 1개인 음절을 포함할 확률이 약 66%임을 나타낸다.

표 3.3 어절의 길이가 2, 3일 때 기능어의 길이에 대한 통계

Table 3.3 Statistics on lengths of functional words where the length of word-phrase is 2 or 3

어절 길이(L)	기능어의 길이(D)	확률(P)
2	0	0.5113
2	1	0.4698
3	0	0.1923
3	1	0.6658
3	2	0.1318

표 3.4는 제안된 동일 어근 추출 방법을 기술한 알고리즘이다. 먼저, 입력 받은 어절의 길이(L)를 구한다. 통계 정보에서 어절 길이(L)에 해당하는 기능어의 길이 중 확률이 가장 높은 길이(D)를 구한다. 다음으로, 입력 받은 어절의 오른쪽 D개의 음절이 기능어 사전에 존재하는지 확인하여 존재한다면 D개의 음절을 삭제한다. 만약 존재하지 않는다면 어절 길이(L)에 해당하는 기능어의 길이 중 확률이 다음으로 높은 것의 길이(D)를 구하여 위의 과정을 반복한다.

표 3.4 제안된 동일 어근 추출 알고리즘

Table 3.4 The proposed stemming algorithm

```
string  s = <입력 문장>;
table   t = <표 3.3의 통계표>;
dict    dt = <기능어 사전>;
eojol   e[N] = split(s, 공백문자);  // 어절 단위로 분리
count   N = <어절 개수>;
index   i = 0;

while( i < N ) {
  l = length(e[i]);              // 분리된 어절의 길이 구함
  table st = t에서 어절 길이 l에 해당하는 부분만 선택;
  sort(st, desc, by 확률);       // 내림차순 정렬
  while( st의 마지막 원소까지 ) {
    p = st에서 순서대로 선택한 확률;
    d = st에서 p에 해당하는 기능어 길이;
    if( 어절 e[i]의 오른쪽 d개의 음절이 dt에 있는가? ) {
      e[i]의 오른쪽 d개의 음절 삭제;
      break;
    }
  }
  e[i]를 색인어로 추출;
  i++;
}
```

예를 들어, 표 3.5의 “철수가 학교에 간다.”라는 문장에서 ‘철수가’에 대해 동일 어근의 추출 과정을 살펴보자. ‘철수가’는 3음절로 이루어져 있기 때문에 길이가 3이 된다. 이 때, 기능어의 길이에 대한 통계 사전(표 3.3)을 보면 기능어의 길이가 1일 때 확률이 가장 높다. 마지막으로 ‘철수가’에서 오른쪽 한 음절 ‘가’가 조사 사전에 포함되었기 때문에 ‘가’를 제거한 ‘철수’라는 음절이 동일 어근으로 추출된다.

표 3.5 제안된 방법의 동일 어근 추출의 예

Table 3.5 An example of the proposed stemming method

“**철수가** 학교에 간다.” 1) ‘철수가’에서 어절의 길이 L : 3 2) 어절 길이(L)가 3일 때, 조사/접미사/어미 길이(D)가 1인 확률이 가장 높음 3) ‘철수가’의 오른쪽 1개의 음절 ‘가’가 조사 사전에 포함되었는지 조사 4) ‘가’가 조사 사전에 있으므로 입력 어절에서 ‘가’를 제거 5) 동일 어근 ‘**철수**’를 추출

3.2 혼합 *n*-그램을 이용한 색인 방법

기존의 *n*-그램 색인 방법은 문장을 연속된 음절의 나열로 보고 *n*개의 음절을 순서대로 분리하여 색인하는 방법이다. 이 때 *n*은 하나의 숫자로 고정되어 있다. 예를 들어, *n*이 2인 경우에는 입력된 문장을 2개의 음절로 분리하여 색인어를 추출하고, *n*이 3인 경우에는 입력된 문장을 3개의 음절로 일정하게 분리하여 색인어를 추출한다. ‘정보검색시스템은’이라는 어절에 대해 2-그램은 ‘정보’, ‘보검’, ‘검색’, ‘색시’, ‘시스’, ‘스템’, ‘템은’이라는 색인어를 추출하고, 3-그램은 ‘정보검’, ‘보검색’, ‘검색시’, ‘색시스’, ‘시스템’, ‘스템은’이라는 색인어가 추출된다. 이 방법은 앞 장에서 살펴본 바와 같이 계산 과정이 단순하며 복합명사의 띄어쓰기에 따른 검색 오류 문제를 해결할 수 있지만, ‘시스’, ‘보검색’ 등과 같은 의미 없는 색인어가 많이 추출되어 부적합 문서를 검색할 가능성이 커지며, 저장 공간을 많이 차지 한다는 단점이 있다.

본 논문에서 제안된 혼합 *n*-그램 색인 방법은 2-그램과 3그램 사이의 공기

정보를 이용해 n을 가변적으로 선택할 수 있도록 한 방법이다. 이는 한글의 경우, 명사 어절들의 출현 비율이 2, 3음절 단어가 전체의 약 80%를 차지한다(이준호, 1996)는 특징에서 착안하였다.

기존의 2-그램에서는 '시스템'이라는 어절에 대해 항상 '시스'와 '스템'이라는 색인어를 추출하였다. 그러나 만약 문서 전체에서 '시스'와 '스템'이라는 색인어가 항상 '시스템'이라는 단어가 나올 때만 발생한다고 가정하면 이 어절에 대해 '시스템'이라는 한 단어만 색인하면 더 효과적일 것이다. 이처럼 혼합 n-그램 방법은 세 음절 단어와 그것을 구성하는 두 쌍의 두 음절 단어 사이의 발생 확률을 비교하여 2-그램을 선택할 것인지 3-그램을 선택할 것인지 결정하게 된다.

세 음절의 단어와 두 쌍의 두 음절 단어 사이의 발생 확률은 식 (3.1)을 이용하여 계산한다. 이 때, 확률은 t-분포(Manning and Schütze, 1999)를 따르고, 계산된 값이 t-분포에 의하여 기준치를 넘어가면 3-그램을 선택하고 그렇지 않으면 2-그램을 선택하게 된다.

$$t = \frac{P(xyz) - P(xy)P(yz)}{\sqrt{\frac{P(xyz)}{N}}} \quad \cdots\cdots (3.1)$$

$P(xyz)$: 3-그램 xyz가 출현할 확률

$P(xy)$: 2-그램 xy가 출현할 확률

$P(yz)$: 2-그램 yz가 출현할 확률

N : 전체 문서 집합에서 3-그램의 개수

제안된 방법을 이용한 색인 과정은 그림 3.2와 같다. 어절 분리 단계에서는 입력된 문장을 공백 문자를 기준으로 나누고, 동일 어근 추출 단계에서는 앞 절에서 제안된 방법을 이용하여 조사/접미사/어미를 제거한다. 2-그램과 3-그램을 선택하는 기준은 (식 3-1)의 t-점수를 이용하여 t-분포에 따라 결정한다.

또한 유의수준(level of significance)에 따라 기준 값을 가변적으로 선택할 수 있다.

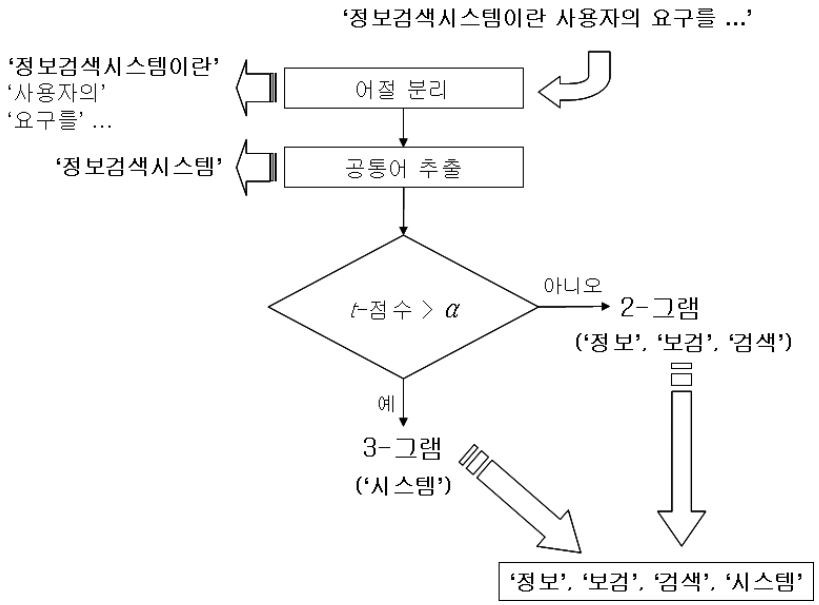

그림 3.2 혼합 n-그램 색인 방법

Fig. 3.2 The indexing method based on the mixed n-gram

t-분포는 $X_1, X_2, \cdots, X_n$가 정규모집단 $N(\mu, \sigma)$에서 추출된 확률표본이고, 표본평균(sample mean)과 분산(variance)이 각각 $\overline{X} = \frac{1}{n}\sum X_i$와 $s^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$ 이면, $t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$의 분포를 말하고, 이를 자유도 $n-1$인 t-분포라고 부른다.

표 3.6 t-분포표

Table 3.6 t-distribution table

d.f.	α = .1	α = .05	α = .025	α = .01	α = .005	α = .001
1	3.078	6.314	12.709	31.821	63.657	318.309
2	1.886	2.92	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.341	5.841	10.215
…	…	…	…	…	…	…
120	1.289	1.658	1.98	2.358	2.617	3.16
240	1.285	1.651	1.97	2.342	2.596	3.125
…	…	…	…	…	…	…
∞	1.282	1.645	1.96	2.326	2.576	3.09

표 3.6에서 d.f.는 자유도(degree of freedom)를 나타내며, 본 논문에서는 전체 문서 집합에서 3-그램의 총 개수가 자유도에 해당한다. t-분포는 자유도가 클 때, 근사적으로 표준정규분포를 따르는 특징이 있다. 따라서 본 논문에서는 자유도를 ∞로 가정하고 확률 구간에 따른 기준 값을 설정하였다.

제 4 장 실험 및 평가

본 장에서는 기존의 색인 방법과 본 논문에서 제안된 혼합 n-그램 색인 방법의 성능을 비교하고 분석한다.

4.1 실험 환경

색인 방법의 객관적인 성능 비교를 위해 국내에서 문서검색시스템의 성능을 평가하기 위해 개발된 KT-SET(박영찬 외, 1995), KEMONG-SET(강현규, 1997)을 이용하였다. 이들 평가 실험 집합의 특성을 살펴보면 아래와 같다.

KT-SET : KT-SET 1.0은 한국과학기술원에서 1994년에 구축하였고, 정보과학회 논문으로 이루어진 1,053개의 문서와 30개의 단순 질의로 구성되었다. 1996년에는 KT-SET 1.0을 확장하여 KT-SET 2.0을 구축하였다. 전기/전자, 컴퓨터 분야의 논문, 신문기사 등으로 된 4,414개의 문서와 50개의 자연언어 질의 및 불리언 질의로 구성되어 있다. 문서의 일부는 한국어-영어 정렬문서로 이루어져있다.

KEMONG-SET : 한국전자통신연구원(ETRI)에서 1997년에 구축하였으며, 계몽사 백과사전(계몽사, 1992)으로 하나의 표제어에 해당하는 것을 하나의 문서로 만든 것이다. 23,113개의 문서와 46개의 질의, 각 질의에 대한 적합 문서 집합으로 구성되어 있다. 각 문서에 대해 12개의 대분류와 76개의 소분류의 분류 정보를 포함하고 있어서 문서분류시스템의 성능 평가를 위한 실험 집합으로 이용할 수 있다.

표 4.1 각 실험 집합의 문서 수와 질의 수

Table 4.1 The numbers of documents and queries on reference test collections

Test collection	문서 수	질의 수
KT-SET	4,414	50
KEMONG-SET	23,113	46

자료 저장과 검색을 위해서는 Lemur Toolkit 2.2[1]를 사용하였다. Lemur Toolkit 2.2는 Carnegie Mellon University와 University of Massachusetts, Amherst에서 언어 모델링과 정보 검색 분야 연구를 위해서 설계되고 개발되었다.

4.2 평가 방법

형태소 단위 색인 방법과 어절 단위 색인 방법과 n-그램 색인 방법과 본 논문에서 제안된 혼합 n-그램 색인 방법의 성능 평가는 검색 효과(retrival effectiveness)와 공간 효율(space efficiency)로 비교한다(Baeza-Yates and Ribeiro-Neto, 1999).

일반적으로 정보 검색 시스템의 검색 효과는 재현율(recall)과 정확률(precision)로 평가되며, 각각 식 (4.1)과 식 (4.2)와 같이 정의된다. 재현율은 사용자가 원하는 문서 중에 시스템이 얼마나 많은 적합 문서를 검색하였는가를 나타내고, 정확률은 검색된 적합 문서 중에 실제로 사용자가 원하는 문서가 얼마나 포함되었는가를 나타낸다.

1) http://www-2.cs.cmu.edu/~lemur

$$\text{재현율}(R) = \frac{(\text{검색된 적합 문헌 수})}{(\text{적합 문헌 총수})} \times 100 \quad \cdots\cdots (4.1)$$

$$\text{정확률}(P) = \frac{(\text{검색된 적합 문헌 수})}{(\text{검색된 문헌 총수})} \times 100 \quad \cdots\cdots (4.2)$$

예를 들어, 전체 문서가 100개, 그 문서들 중에 사용자 질의에 적합한 문서가 5개라고 가정할 때, 시스템이 검색한 문서의 수가 6개이고, 그 중 적합 문서는 4개일 때, 재현율은 80%이고 정확률은 67%이다.

일반적으로 재현율이 높으면 정확률이 낮고, 정확률이 높으면 재현율이 낮게 나타난다. 따라서 재현율과 정확률을 포함한 성능 평가 방법으로 재현율과 정확률의 조화 평균(harmonic mean) F를 이용하는 방법과 재현율과 정확률의 교차점을 점수로 하는 교차 점수(BEP: Break Even Point)를 사용하기도 한다. 조화 평균 F는 식 (4.3)과 같이 정의된다. 본 논문에서는 교차 점수를 이용하여 전체 성능을 평가한다.

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}} = \frac{2RP}{R+P} \quad \cdots\cdots (4.3)$$

검색 시스템은 보간 기법을 사용하여 고정된 재현율에 대한 정확률을 계산할 수 있다. 일반적으로 11개의 재현율에 대해 정확률을 평균한 값을 사용하는데, 이를 11-포인트 평균 정확률(11-point average precision)이라고 한다. 본 실험에서도 11-포인트 평균 정확률을 이용해서 성능을 비교한다. 또한 색인 방법에 따른 색인어 개수를 비교하여 저장 공간에 대한 성능을 평가한다.

4.3 성능 평가

가. 검색 모델에 따른 성능 비교

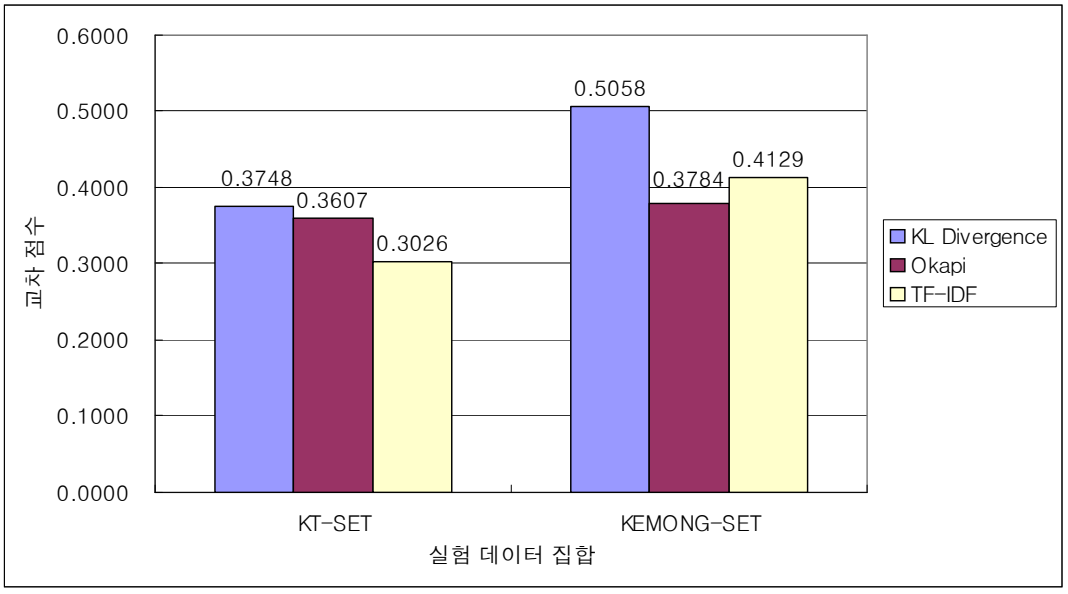

그림 4.1 검색 모델에 따른 교차점수
Fig. 4.1 Comparison of BEPs on retrieval models

Lemur Toolkit 2.2에서 제공하는 검색 모델은 크게 TF-IDF, Okapi(Robertson *et al.*, 1996), KL-Divergence(Zhai and Lafferty, 2001)로 나누어지고, 각 모델은 입력 매개변수에 따라 세분화되어있다. 그림 4.1은 실험 집합에 대해 수행한 검색 모델에 따른 교차 점수를 비교하기 위한 막대그래프이다. 실험에서 사용한 매개변수는 Lemur Toolkit 2.2에서 제공하는 기본 값을 사용하였고, 기존의 연구(이준호, 1996)에서 좋은 성능을 보인 2-그램 색인 방법을 이용하였다. 본 논문에서는 실험 집합에 대해 우수한 성능을 보인

KL-Divergence 모델을 채택하였다. 따라서 이하의 실험에서는 KL-Divergence 모델을 사용한다.

나. 동일 어근 추출 방법에 따른 성능 비교

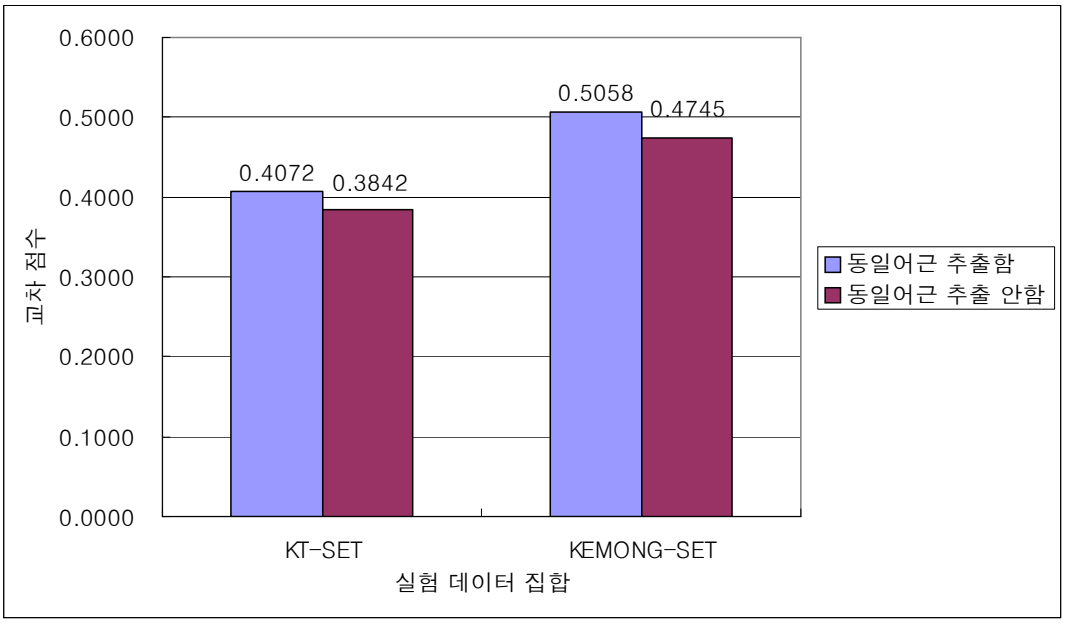

그림 4.2 동일 어근 추출에 따른 교차 점수 비교
Fig. 4.2 Comparison of BEPs depending on stemming

본 절에서는 제안된 동일 어근 추출 방법에 따른 성능을 비교하였다. 실험에 사용한 색인은 2-그램 색인 방법을 이용하였다. 그림 4.2에 나타난 것과 같이 제안된 방법으로 동일 어근을 추출한 경우가 더 좋은 성능을 보였다. 또한, 색인어 개수도 표 4.2와 같이 동일 어근을 추출하지 않은 경우보다 최대 37% 정도 적게 추출되므로 저장 공간의 효율에서도 높은 성능을 보였다. 따라서 이하의 실험에서는 제안된 동일 어근 추출 방법을 사용한다.

표 4.2 동일 어근 추출에 따른 색인어 개수

Table 4.2 The numbers of index terms depending on stemming

	동일 어근을 추출한 경우	동일 어근을 추출하지 않은 경우	축소율(%)
KT-SET	816,860	1,280,629	36
KEMONG-SET	1,548,963	2,474,631	37

다. 형태소 단위 색인 방법과 어절 단위 색인 방법의 성능 비교

본 절에서는 형태소 단위 색인 방법과 어절 단위 색인 방법에 대해 성능을 평가하였다. 어절 단위 색인 방법은 본 논문에서 제안된 동일 어근 추출 방법을 사용하였다. 그림 4.3과 같이 KT-SET에 대해서는 두 가지 방법에 대해 0.006 정도의 차이로 형태소 분석 색인 방법이 나은 성능을 보였지만 미미한 수준이었다. KEMONG-SET에 대해서는 어절 단위 색인 방법이 나은 성능을 보였다. 이는 앞서 설명했던 것처럼 형태소 분석은 사전과 언어적 지식이 필요한데 이 과정에서 사전에 등록되지 않은 단어나 애매한(ambiguous) 분석 결과로 인한 오류가 있었기 때문에 낮은 성능을 보인 것으로 판단된다.

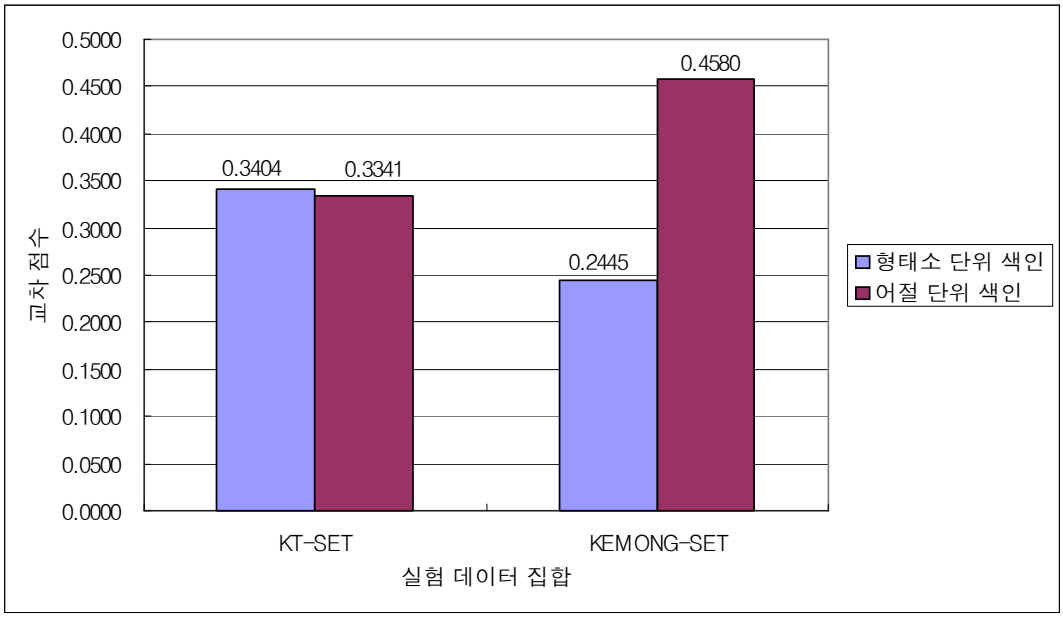

그림 4.3 형태소 단위 색인 방법과 어절 단위 색인 방법의 교차 점수
Fig 4.3 Comparison of BEPs between a morpheme-based indexing method and a word-phrases-based indexing method

라. 유의수준에 따른 혼합 *n*-그램 색인 방법의 성능 비교

혼합 *n*-그램 색인 방법은 *t*-점수를 계산하여 *t*-분포에 따라 2-그램과 3-그램을 선택적으로 결정한다. 2-그램과 3-그램을 선택할 때 사용하는 기준 값은 유의수준에 따라 달라진다. 본 절에서는 유의수준에 따라 나타나는 성능을 비교하였다. 유의수준에 따라 나타난 혼합 *n*-그램 색인 방법은 그림 4.4와 같이 유의수준에 따른 큰 차이는 없으나 유의수준 0.5%일 때 전체적으로 좋은 성능을 보였다. 따라서 이하의 실험에서는 유의수준 0.5%에 해당하는 혼합 *n*-그램 색인 방법을 이용하였다.

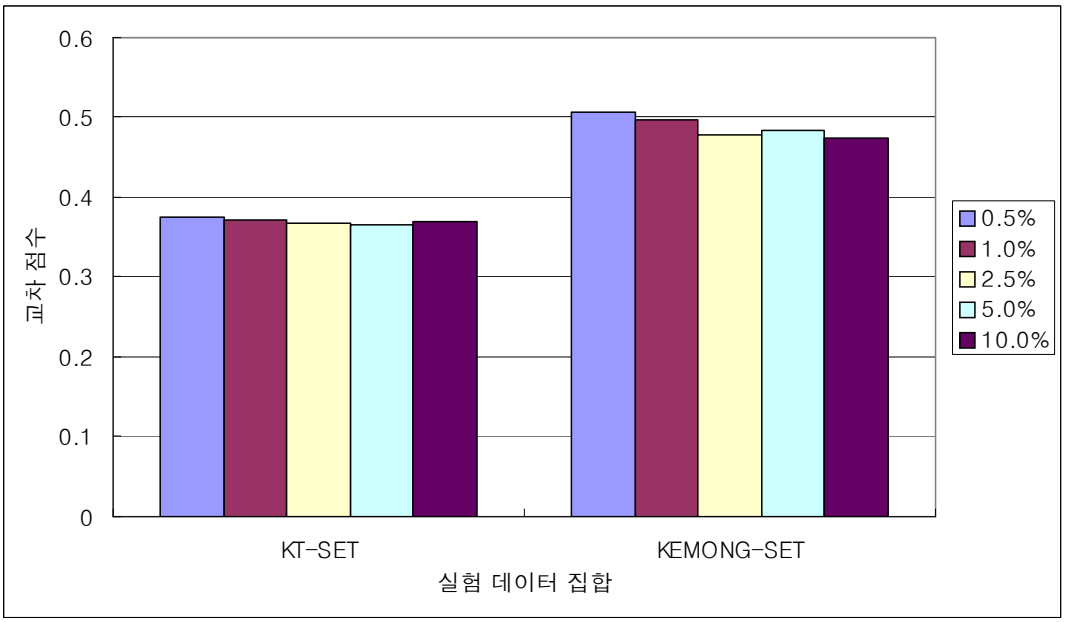

그림 4.4 유의수준에 따른 혼합 n-그램 색인 방법의 교차 점수 비교
Fig 4.4 Comparision of BEPs on levels of significance in mixed n-gram indexing method

마. 색인 방법에 따른 성능 비교

그림 4.5와 그림 4.6의 그래프는 각 실험 집합에 따른 11-포인트 평균 정확률을 나타낸 것이다.

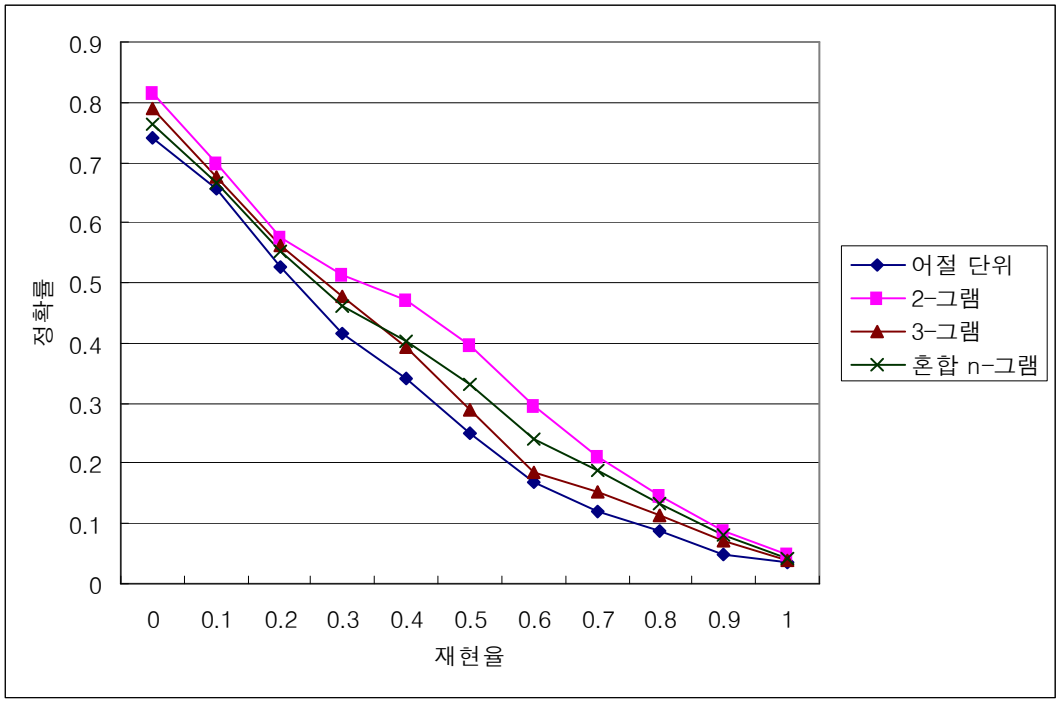

그림 4.5 KT-SET의 11-포인트 평균 정확률

Fig. 4.5 11-point average precision on KT-SET reference test collection

그림 4.5의 KT-SET에 대한 11-포인트 평균 정확률은 2-그램 색인 방법이 전체적으로 높게 나왔다. 이는 혼합 *n*-그램 색인 방법은 계산된 확률이 *t*-분포를 따르게 되는데 *t*-분포는 전체 데이터 수가 큰 경우 정확률이 높아진다는 특징에 기인한 것으로 추측할 수 있다. KT-SET의 문서 수는 4,414개이다(표 4.1). 하지만 기존의 연구에서 *n*-그램 색인 방법이 형태소 분석이나 어절 단위 색인 방법보다 좋은 성능을 보인 것(이준호, 1998)처럼 혼합 *n*-그램 색인 방법에 대해서도 형태소 분석이나 어절 단위 색인 방법보다 높은 성능을 보였다.

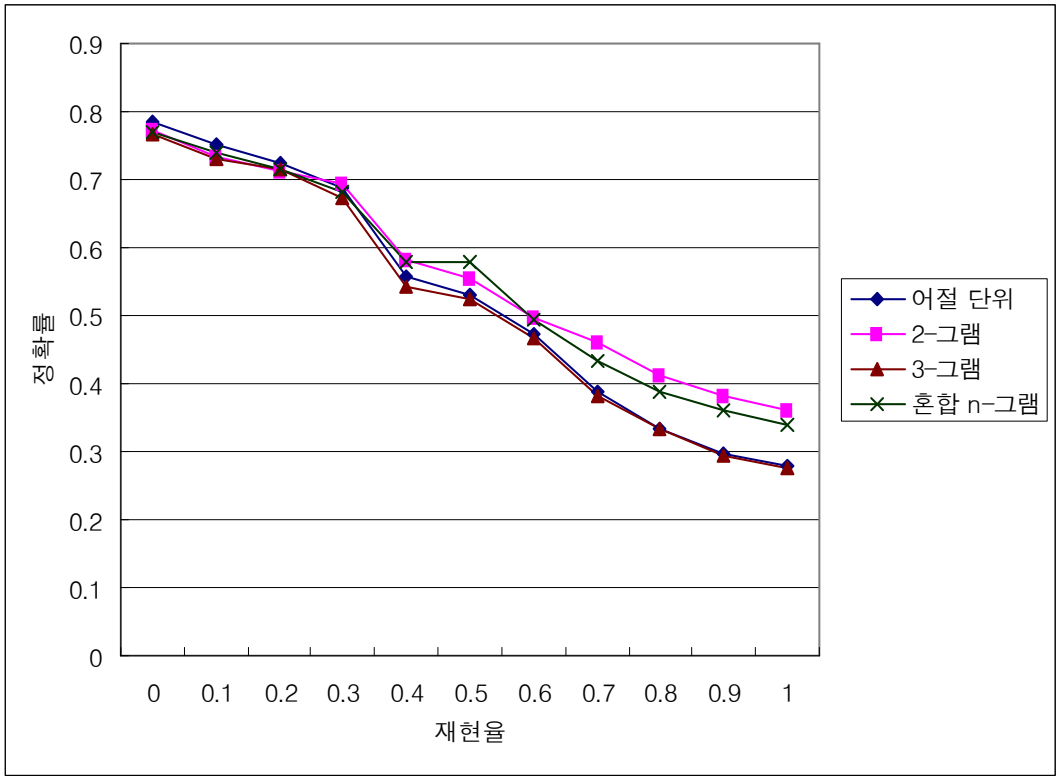

그림 4.6 KEMONG-SET의 11-포인트 평균 정확률

Fig. 4.6 11-point average precision on KEMONG-SET reference test collection

그림 4.6의 KEMONG-SET에서 11-포인트 평균 정확률은 2-그램 색인 방법과 혼합 *n*-그램 색인 방법이 유사한 수준이었다. 하지만 표 4.3과 그림 4.7에 나타난 교차 점수는 혼합 *n*-그램 색인 방법이 가장 높은 성능을 보였다. 이는 앞서 언급했던 것처럼 *t*-분포는 자유도, 즉 전체 데이터 수가 많은 경우에 정확도가 높아지기 때문이라고 추측할 수 있다. KEMONG-SET의 전체 문서 수는 23,113개이다(표 4.1).

표 4.3 색인 방법에 따른 교차 점수 비교

Table 4.3 Comparison of BEPs on indexing methods

	KT-SET	KEMONG-SET
형태소 분석	0.34040	0.44429
동일 어근 추출만	0.33412	0.45804
2-그램(동일 어근 추출 안함)	0.38423	0.47452
2-그램(동일 어근 추출)	0.40720	0.50580
3-그램(동일 어근 추출)	0.36606	0.43895
혼합 *n*-그램	**0.37484**	**0.50664**

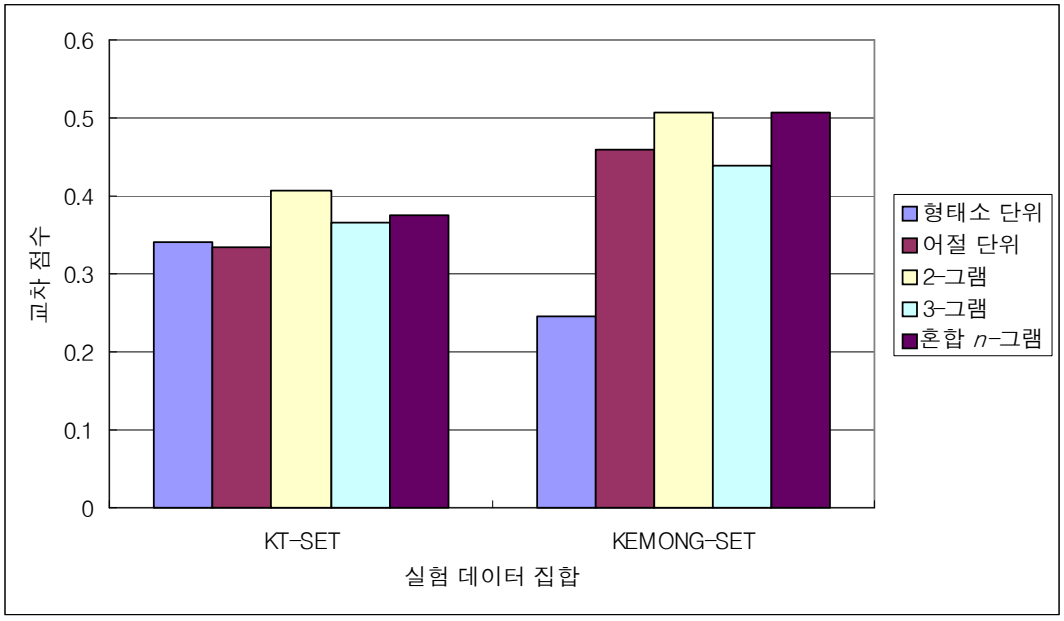

그림 4.7 색인 방법에 따른 교차 점수 비교

Fig. 4.7 Comparison of BEPs on indexing methods

표 4.4 색인 방법에 따른 색인어 개수 비교

Table 4.4 Comparison of the numbers of index terms on indexing methods

	KT-SET	축소율(%)	KEMONG-SET	축소율(%)
형태소 단위	750,598	41.4	1,503,357	39.2
어절 단위	631,484	50.7	1,332,861	46.1
2-그램(동일 어근 추출 안함)	1,280,629	0.0	2,474,631	0.0
2-그램(동일 어근 추출)	816,860	36.2	1,548,963	37.4
3-그램(동일 어근 추출)	708,135	44.7	1,389,052	43.9
혼합 *n*-그램	746,361	41.7	1,456,100	41.2

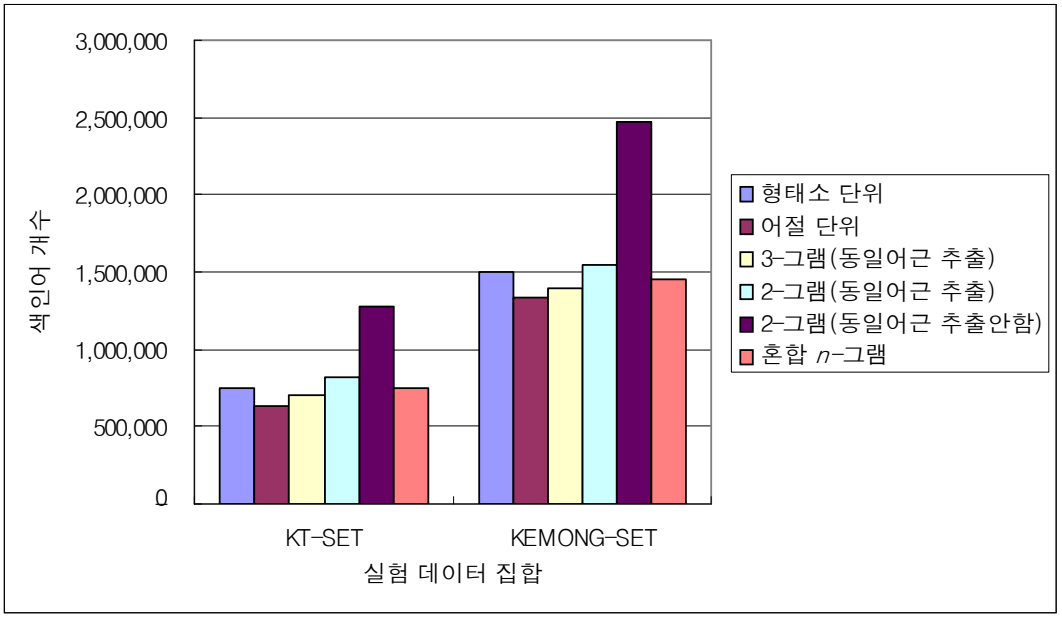

그림 4.8 색인 방법에 따른 색인어 개수 비교

Fig. 4.8 Comparison of the numbers of index terms on indexing methods

정보 검색 시스템의 저장 공간의 효율에 대한 성능을 평가하기 위해 색인 방법에 따른 색인어 개수를 표 4.4와 그림 4.8에 나타내었다. 표에서 축소율은 동일 어근을 추출하지 않은 2-그램 색인 방법을 기준으로 색인어 개수가 줄어든 비율을 나타내었다. 표에 나타난 것처럼 가장 많은 색인어를 생성한 방법은 동일 어근 추출 없이 2-그램을 이용한 방법이었고, 가장 적은 색인어를 생성한 방법은 동일 어근만 추출해서 색인어를 선택한 어절 단위 색인 방법이었다.

4.4 토의

실험에서 살펴보았듯이 검색 시스템을 평가하기 위해서는 다양한 측도를 사용한다. 각각의 측도에 따른 성능을 살펴보면 첫째, 11-포인트 평균 정확률에서는 전체적으로 기존의 *n*-그램 색인 방법이 우수한 성능을 보였다. 반면에 본 논문에서 제안된 혼합 *n*-그램 색인 방법은 적은 양의 문서에서는 기존의 *n*-그램과 유사하거나 약간 밑도는 성능을 보였다. 하지만 상대적으로 많은 양의 실험 집합에서는 가장 우수한 성능을 보였다. 둘째, 교차 점수로 살펴본 성능도 전체적으로 기존의 *n*-그램 색인 방법이 우수하였으나 실험 데이터 량이 많은 KEMONG-SET에서는 제안된 혼합 *n*-그램 색인 방법이 가장 우수하게 나타났다. 마지막으로 저장 효율의 성능 평가를 위한 색인어 개수 비교에서는 동일 어근을 추출하지 않은 2-그램이 가장 많은 개수의 색인어를 생성하였고, 동일 어근을 추출하는 어절 단위 색인 방법이 가장 적은 양의 색인어를 생성하였다.

전체 실험에서 나타난 결과를 보면 재현율과 정확률의 측면에서는 기존의 *n*-그램 방식과 본 논문에서 제안된 혼합 *n*-그램 색인 방법이 우수하게 나타났다. 특히 기존의 *n*-그램 색인 방법에서 동일 어근을 추출하는 방법을 형태소 분석이 아닌 제안된 방법을 이용하므로 형태소 분석에 필요한 추가 비용이 발생하

지 않는다. 또한, 동일 어근을 추출하지 않는 2-그램 색인 방법보다 색인어 개수가 약 36% 정도 적게 나타났다.

KT-SET에서는 혼합 *n*-그램 색인 방법이 기존의 *n*-그램 색인 방법의 성능에 약간 밑도는 수준이었다. 이는 혼합 *n*-그램 색인 방법에서 2-그램과 3-그램을 선택하는 기준 값으로 사용하는 *t*-점수가 많은 데이터에 대해서 더 정확하기 때문에 상대적으로 문서의 수가 적은 KT-SET에 대해서는 성능이 낮게 나타난 것으로 추측된다. 하지만 이 경우에 있어서도 형태소 분석이나 어절 단위 색인 방법보다 좋은 성능을 보였다.

전체적으로 본 논문에서 제안된 혼합 *n*-그램 색인 방법이 기존의 색인 방법과 성능이 유사하거나 대량의 데이터에 대해서는 우수하였고, 색인어의 개수는 색인어가 가장 많이 생성된 동일 어근을 추출 하지 않은 2-그램에 비해 최대 43%정도 줄어들었다. 또한, 동일 어근 추출을 한 2-그램에 대비해서도 약 5% 정도 줄어들었다.

제 5 장 결 론

본 논문에서는 한글 문서를 색인하기 위한 동일 어근 추출 방법과 혼합 *n*-그램 색인 방법을 제안하였다. 기존의 *n*-그램 색인 방법이 형태소 단위 색인 방법이나 어절 단위 색인 방법보다 나은 성능을 보이지만, 의미 없는 색인어가 다수 생성되어 부적합 문서를 검색하기 때문에 검색 효율을 떨어뜨리는 단점이 있다. 이를 보완하기 위한 방법으로 혼합 *n*-그램 색인 방법을 제안하였다.

기존의 연구에서는 *n*-그램을 이용할 때, 동일 어근을 추출하지 않거나, 형태소 분석을 이용하여 동일 어근을 추출하였다. 동일 어근을 추출하지 않고 *n*-그램을 이용하면 실험에서 나타난 것처럼 색인어의 개수가 가장 적은 색인 방법에 비해 약 36~55%가 증가한다. 또한, 형태소 분석을 이용하여 동일 어근을 추출하면 형태소 분석을 위한 사전 정보와 문법 정보가 필요하며, 상대적으로 연산 량이 많은 형태소 분석 때문에 속도가 느려지는 단점이 있다.

본 논문에서는 크게 형태소 단위 색인 방법, 어절 단위 색인 방법, *n*-그램 색인 방법, 혼합 *n*-그램 색인 방법에 대해 KT-SET과 KEMONG-SET을 가지고 성능을 비교하였다. 실험 집합에 대해 데이터 량에 상관없이 기존의 *n*-그램을 이용한 방법이 대체적으로 높은 성능을 나타내었다. 하지만 상대적으로 실험 집합이 큰 KEMONG-SET에 대해서는 혼합 *n*-그램 색인 방법이 좋은 성능을 보였다. 이는 혼합 *n*-그램 색인 방법에서는 2-그램과 3-그램의 선택을 *t*-점수를 이용하는데, 사용하는 *t*-분포는 대량의 문서에 대해 정확도가 높게 나타나는 특성 때문인 것으로 판단된다. 따라서 상대적으로 큰 KEMONG-SET에 대해 더 좋은 성능을 보였다.

실험 결과에 의하면 대체적으로 기존의 *n*-그램 색인 방법의 성능이 좋았다. 하지만 대량의 데이터에 대해서는 본 논문에서 제안된 혼합 *n*-그램 색인 방법

이 더 나은 성능을 보였다. 인터넷 상의 웹 페이지를 대상으로 검색하는 시스템의 경우에는 엄청나게 많은 양의 문서를 다룬다. 실제로 현재 상용 서비스를 하고 있는 구글[2]에 색인되어 있는 사이트 수는 대략 10억 개 이상이다(구글, 2004). 이렇게 대량의 문서를 다루는 정보 검색 시스템에서 제안된 혼합 n-그램 색인 방법을 채택한다면 재현율과 정확률 뿐 아니라 저장 공간의 효율에 있어서도 많은 이득을 얻을 수 있을 것이다.

혼합 n-그램 색인 방법은 2-그램과 3-그램을 선택하는 방법으로 t-분포를 따르는 t-점수를 이용한다. t-분포의 특성상 문서의 양이 적은 실험 집합에 대해서는 기존의 2-그램과 비슷한 성능을 나타내었다. 따라서 앞으로 본 연구를 확장하여 2-그램과 3-그램을 선택하는 측도를 다양한 방법으로 시도하여 문서의 양에 큰 영향을 받지 않는 측도를 찾아야 할 것이다. 또한 동일 어근 추출 방법에 사용되는 한 어절에 포함된 기능어의 개수에 대한 통계를 더 큰 코퍼스를 통해 계산하여 좀 더 나은 성능을 얻도록 실험해야 할 것이다.

2) http://www.google.co.kr

참 고 문 헌

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*, Addison-Wesley.

Blumer, A., Blumer, J., Haussler, D., McConnell, R. and Ehrenfeucht. A. (1987) "Complete inverted files for efficient text retrieval and analysis". *Journal of the ACM*, vol. 34, no. 3, pp. 578-595.

Lafferty, John and Zhai, Chengxiang (2002) "Document language models, query models, and risk minimization for information retrieval", *Proceedings of Special Interest Group on Information Retrieval*, pp. 111-119.

Lee, Joon Ho, Cho, Hyun Yang and Park, Hyouk Ro (1999) "*n*-Gram-based indexing for Korean text retrieval", *Information Processing and Management*, vol. 35, pp. 427-441.

Manning, C. D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press.

Nie, Jian-Yun and Ren, Fuji (1999) "Chinese information retrieval: using characters or words?", *Information Processing and Management*, vol. 35, pp. 443-462.

Robertson, S. E., Walker, S., Beaulieu, M. M, Gatford, M. and Payne, A.

(1996) “Okapi at TREC-4”, *Proceedings of The Fourth Text REtrieval Conference (TREC-4)*, pp. 73-96.

Zhai, Chengxiang and Lafferty John (2001) “Model-based feedback in the language modeling approach to information retrieval”, *Proceedings of Conference on Information and Knowledge Management*, pp. 403-410.

Zhai, Chengxiang and Lafferty, John (2001) “Model-based feedback in the language modeling approach to information retrieval”, *Proceedings of Conference on Information and Knowledge Management*, pp. 403-410.

강승식 (2002) *한국어 형태소 분석과 정보 검색*, 홍릉과학출판사.

강승식 (2004) “한글 문서의 색인어와 색인 기법”, *정보과학회지*, 제22권, pp. 72-77.

계몽사 (1992), *계몽사 백과 사전*, (주)계몽사, 서울.

구글 (2004) http://www.google.com/intl/ko/why_use.html

김재훈 (1998) “가중치망 모델을 이용한 한국어 품사 태깅”, *정보과학회논문지(B)*, 제25권, 제6호, pp. 951-969.

김판구 (1994) *한국어 정보 검색을 위한 상호 정보량에 기반한 복합어 자동색인*, 서울대 박사학위 논문.

박영찬, 최기선, 김영환, 김재군 (1996) “한국어 정보검색 연구를 위한 시험용 데이터 모음 2.0 (KT-SET 2.0) 개발”, *한국어정보과학회 인공지능연구*

회 춘계학술대회, pp. 59-65.

이준호, 안정수, 박현주, 김명호 (1996) “한글 문서의 효과적인 검색을 위한 *n*-Gram 기반의 색인 방법”, *정보관리학회지*, 제13권, 제1호, pp. 47-63.

임해창, 윤보현, 강승식 (1995) “한국학 서지정보와 전자텍스트를 위한 자동색인 및 검색시스템 개발 연구”, *한국어전산학*, 제2집, pp. 279-292.

최기선 (1991), “구문 및 의미분석을 통한 한국어 자동색인”, *정보관리학회지*, 제8권, 제2호, pp. 96-107.

감사의 글

본 논문이 완성되기까지 인내와 열정으로 지도해 주신 김재훈 교수님께 감사드립니다. 학부를 졸업하던 해에 교수님께서 다시 한국에 들어오신다는 소식을 듣고 특별한 준비도 없이 대학원 진학 결정을 내렸던 저를 지금까지 지켜봐주시고 가르쳐 주셔서 너무 감사드립니다. 교수님의 세심한 배려와 격려 덕분에 이 논문을 마칠 수 있었습니다. 또한 바쁘신 와중에도 부족한 논문을 심사해 주시고 일일이 지적해 주시고 고쳐 주신 류길수 교수님과 박휴찬 교수님께도 감사의 말씀을 드립니다. 지금은 외국에 계시지만 학부 때부터 깊은 가르침을 주신 신옥근 교수님과 손주영 교수님께도 감사드립니다.

매일 아침 도시락을 싸주고 늦은 밤까지 간식을 챙겨주며 잠을 아꼈던, 이제는 저의 신부가 된 아내에게도 고마움을 전합니다. 혹시라도 먼저 잠이 들면 남편의 논문이 늦어질까 무거운 눈을 비비며 옆에서 지켜봐주던 아내가 아니었으면 논문을 마치지 못했을 것입니다. 결혼을 얼마 앞두지 않고 사고로 병석에 누워계신 아버지와 늘 곁에서 격려를 아끼지 않으시던 어머니께도 감사의 마음을 전합니다. 어릴 때부터 당신들은 늘 부족하게 지내시면서 저에게는 항상 좋은 것으로 먹이고 입히셨던 것을 잊지 못합니다. 항상 묵묵히 근엄하게 지켜봐주신 장인어른과 맛있는 반찬을 챙겨주시며 기도해 주신 장모님께도 감사드립니다. 부족한 사위에게 예쁜 딸을 주셔서 더욱 감사드립니다.

공부에만 전념하라고 격려해주시고 어려울 때마다 기도와 여러 가지 후원으로 도와주신 부산 비전교회의 박길서 담임 목사님과 안영숙 사모님께도 감사드립니다. 지칠 때마다 힘주시고 기도해주신 이재국 부목사님과 오현옥 사모님, 고교 시절부터 지금까지 기도로 후원해 주신 장상근 강도사님과 권시자 사모님, 아침마다 학교까지 차를 태워주신 김영대 전도사님, 저를 볼 때마다 미소를

아끼지 않으시던 이동숙 전도사님, 홍영 전도사님, 그리고 비전교회 모든 식구들에게도 감사드립니다. 지금은 캐나다에 계신 정은주 전도사님께도 감사의 마음을 전합니다. 고교시절부터 영적 어머니가 되어 지금까지도 기도의 끈을 놓지 않으셔서 너무 감사드립니다.

실험실에서 나보다 더 선배같은 후배 강민이와 최선이라는 단어가 어울리는 병걸이, 영어에는 그 누구도 부럽지 않은 형우, 항상 나의 궁금증과 어려움을 풀어주는 태욱이, 작은 체구의 악발이 희영이, 그리고 늦게 실험실에 합류한 정철이와 동욱이, 긴 시간은 아니었지만 있는 동안 날 즐겁게 해주고 내가 살아있다는 것을 알게 해준 많은 후배들에게도 감사를 전합니다. 점심때마다 같이 식사하고, 나와 같이 졸업을 하려고 무던히 노력하던 성미에게도 감사의 마음을 전합니다. 예쁜 후배 성미 때문에 논문의 압박으로부터 많이 벗어날 수 있었던 것 같습니다. 그리고 학부 때부터 같이 했던 선배님들과 많은 동기들, 후배들에게도 대학의 낭만과 열정을 알게 해주셔서 감사하다는 말을 전합니다. 특히, 학부 때부터 대학원을 졸업하는 오늘까지도 모든 학사 일정과 대학 생활을 꼼꼼히 챙겨주시며 고등학생에서 대학생으로 대학생에서 대학원생으로 거듭나게 해주셨던 강군호 조교 선생님과 김경언 조교 선생님께도 깊이 감사드립니다. 그리고 조교님께서 직접 만들어 주셨던 커피도 잊지 못할 것입니다.

시간이 지나면 모든 것이 저절로 이루어질 것 같고 혼자서 다 할 수 있을 것 같은 어리석음으로 대학과 대학원 생활을 하였는데 그 모든 시간이 지난 지금에서야 중요한 진리를 깨닫게 되었습니다. 이렇게 많은 사람들이 옆에서 지켜봐 주시고 도와주시고 힘을 주시기 때문에 오늘의 기쁨을 누릴 수 있다는 것입니다. 이제는 제가 누군가에게 그런 힘과 용기, 그리고 기쁨을 줄 수 있는 사람으로 사회와 주변에 공헌하길 원합니다.

무엇보다 본 논문을 마칠 수 있도록 하시고 이 모든 감사의 말을 할 수 있도

록 허락하신 하나님 아버지께 진심으로 감사드립니다. 주님을 기쁘시게 하기보다 나의 필요만을 기도하던 어리석은 모습이지만 그래도 저를 사랑해 주시고 인도해 주셔서 정말로 감사합니다.