요약화일 기법을 이용한 한글 문서 검색 시스템의 설계

°이정기, 김철완, 장재우
전북대학교 컴퓨터공학과

Design of Hangul Document Retrieval System Using Signature File Methods

Jeong-Ki Lee, Chul-Wan Kim, and Jae-Woo Chang
Chonbuk National Univ. Dept. of Computer Engineering

요 약

현재 국내에서 한국형 정보 검색 시스템의 중요성을 인식하고, 이를 국내 기술로 구축하려는 여러 시도들이 진행중에 있다. 이 가운데 정보 검색 시스템 구축에 기반이 되는 정보 저장 하부 구조로서, 한글 문서를 위한 효율적인 문서 검색 기법에 대한 연구는 필수적이다. 본 논문에서는 이를 위해 요약 화일 기법을 이용한 한글 문서 검색 기법을 설계한다. 아울러, 제안된 기법을 토대로 실제 시스템을 구현하여 성능을 분석하였다.

I. 서론

도서 정보 검색, 특허 또는 판례 검색 등과 같이 대규모의 문서를 다루는 응용을 위해 정보 검색 시스템(Information Retrieval System)이 사용되고 있다. 일반적으로 정보검색 시스템에서 사용하는 문서 검색 기법으로 우수한 검색 성능을 지닌 역화일 기법이 많이 사용되고 있지만, 역화일 기법은 데이타 화일의 50 ~ 300% 정도의 많은 부가 저장 공간을 필요로 하는 단점이 있다. 이러한 단점을 극복하기 위하여 검색 속도는 역화일 기법보다 뒤떨어지나, 부가 저장 공간의 크기가 데이타 화일의 10%내외인 요약 화일 기법이 제안 되었다[1,2]. 이 요약 화일 기법을 사용하는 시스템의 대표적인 예는 TITAN 시스템으로, 검색

속도나 부가 저장 공간 면에서 모두 우수한 것으로 알려져 있다[3].

한편, 국내에서도 한국형 정보 검색 시스템 구축의 중요성을 인식하고, 이를 구축하려는 여러 논의들이 진행중에 있다. 이를 위해 먼저 연구되어야 할 것은 정보 검색 시스템 구축의 밑바탕이 되는 정보 저장 하부 구조로서, 한글 문서를 효율적으로 검색할 수 있는 기법의 설계이다. 본 논문에서는 이를 위해 요약 화일 기법을 이용한 한글 문서 검색 기법을 설계하고, 이를 토대로 소규모 한글 문서 검색 시스템을 구현하여 성능을 평가한다.

본 논문의 구성은 다음과 같다. II장에서는 요약 화일 기법의 기본 개념과 기존의 영어 문서들을 위한 요약 화일 기법에 대해서 알아본다. III장에서는 요약 화일 기법을 이용한 한글 문서 검색 기법을 설계한다. 먼저 주어진 한글 문서에 대한 요약 화일의 크기를 정하기 위해, 일반 문서에서 색인어의 평균 개수와 색인어의 음절에 따른 비중복 음절 수를 통계적으로 계산한다. 둘째, 해싱(hashing) 함수를 통해 한글 문서에 존재하는 어절에 대해 요약을 추출하는 코딩 알고리즘을 비교한다. IV에서는 III장에서 제안한 기법들을 이용하여 실제 한글 문서 검색 시스템을 구현하여 성능을 분석한다. 마지막으로, V장에서 결론을 제시한다.

II. 영어 문서를 위한 요약 화일 기법

요약 화일 기법이란 원래 문서에 대한 요약(signature)을 추출하여 이를 요약 화일로 구성하고 요약 화일을 먼저 접근하여 질의에 만족할 가능성이 있는 문서들만을 선택적으로 접근함으로써 문서 화일의 검색시간을 줄이는 문서 검색 기법이다. 일반적으로 각 레코드의 요약은 각 단어에 대한 해싱을 하여 단어 요약을 만들고 이를 비트별로 중첩(bitwise ORing)하여 구성한다[4]. 다음은 한 레코드가 'information', 'retrieval', 'system'의 3개 단어로 구성된 경우 레코드 요약을 만드는 과정을 나타낸다. 여기서 m과 k는 각각 레코드 요약 크기와 한 단어에 할당된 '1'의 개수를 나타낸다.

해싱 H : m = 12, k = 2		
H(information)	:	0010 1000 0000
H(retrieval)	:	0000 0011 0000
H(systems)	:	1000 0000 0100
레코드 요약	:	1010 1011 0100

요약 화일에서 질의를 만족하는 레코드를 검색하는 방법은 다음과 같다. 먼저 사용자의 질의로 부터 위와 같은 방법으로 요약을 추출한다. 다음으로, 요약 화일을 검색하여 동일한 비트 패턴을 포함하고 있는 요약들을 추출한다. 이때 이렇게 추출된 요약들이 기리키는 레코드(문서)들은 사용자의 질의를 만족할 가능성이 있는 레코드로 간주된다. 마지막으로 데

이타 화일을 탐색(후위 탐색)하여 실제로 질의를 만족하는 레코드들만 추출한다. 이때 요약 화일에서는 질의를 만족하는 것으로 검색되었지만, 후위 탐색에서 제외되는 레코드를 탈락 착오(false drop) 레코드라 한다.

요약 화일 기법에서 효율적인 검색을 위한 연구로는 탈락 착오를 줄이려는 연구와 요약 화일 자체의 검색 시간을 줄이려는 연구로 분류된다. 탈락 착오를 줄이는 것은 결국, 후위 탐색시 불필요한 데이타 화일의 접근 수를 최소화 시킴으로써, 매우 큰 데이타 화일의 검색 시간을 줄인다[1,2]. 다음으로 요약 화일의 검색 시간을 줄이려는 연구는 요약 화일의 구조에 중점을 둔 연구로서 비트 슬라이스 요약 화일(bit-sliced signature file) 방법[6], 2단계 요약 화일(two-level signature file) 방법[7,8], 그리고 2경로 2단계 요약 화일(two-path two-level signature file) 방법[9,10]이 제안되었다.

III. 한글 문서를 위한 요약 화일 기법의 설계

지금까지의 요약 화일에 대한 연구는 대부분 영문 위주로 이루어 졌다. 따라서, 이를 한글에 적용하기 위해서는 한글에 대한 여러가지 특성에 관한 연구가 선행 되어야 한다. 한글은 첨가어로서 보통 한 단어는 체언, 용언 등과 같이 실질적인 의미를 나타내는 실질형태소(실사)와 조사, 어미 등과 같이 단지 문법적 관계만을 보여주는 형식형태소(허사)로 구성되며 조사를 제외한 다른 형식형태소는 활용을 한다는 특징을 가진다. 따라서 어절의 구성 형태가 매우 복잡하여 실제 색인에 중요한 실질형태소를 추출하기 힘들다. 또한 복합명사는 띄어쓰거나 붙여쓰기도 하기 때문에 빈칸이 무의미할 수도 있다. 따라서, 영문과 같이 어절 중심으로 한글 문서에 대한 요약을 추출할 경우 질의에 부합하는 문서를 찾지 못하는 경우가 발생하게 된다. 따라서 한글 문서에 대한 요약을 추출할 때는 음절 단위로 하여야 한다. 한편, 저장 구조 측면에서, 일단 일정한 코딩 방법에의해 요약이 구성되면, 이 요약의 구조는 영문 문서의 경우와 동일하므로 기존의 영문을 위한 효율적인 요약 화일 구조를 그대로 적용할 수 있다. 따라서 본 논문에서는 구조적인 측면보다 요약 생성의 측면에 대해서 논의 한다.

3.1 비중복 단어의 통계

주어진 문서에 대해 가장 효율적인 요약 크기를 정하기 위해서는 요약 구성에 사용되는 색인어 수를 알아야 한다. 그러나, 색인어 수는 각 문서마다 다르므로 많은 문서를 분석하여 이의 통계치를 구하는 것이 필요하다. 이를 위한 공식은 다음과 같다. 여기서 AW(X)는 문서 X에서의 전체 단어 개수를, CW(X)는 접속사, 부사와 같이 색인에 사용되지 않는 공통단어(common word)의 개수를, 마지막으로 NCDW(X)는 공통 단어를 제외한 색인어 가운데 중복을 제거한 색인어의 개수를 나타낸다.

(1) CW(X) = C__PER * AW(X)

(2) NCDW(X) = D__PER * (AW(X) - CW(X))

(1)에서 C__PER는 공통 단어의 비율로 영어나 이태리어의 경우 약 0.5의 값을 가진다[5]. 또한 (2)에서 D__PER는 색인어의 비율로 영어나 이태리어의 경우 0.02의 매우 작은 값을 가진다[5]. 따라서 영문의 경우 200만 단어로 구성된 문서에 대해 요약 구성을 할 경우 그것의 1%에 해당하는 2만 단어만으로 요약을 구성할 수 있다. 이는 매우 작은 크기의 요약으로도 요약 구성이 가능함을 나타낸다.

그러나, 현재 한글에 대한 위와 같은 자료는 없다. 일부 유사한 연구로서 단행본을 통한 색인어 추출의 연구에 따르면[11], 중복을 허락한 15,768 개의 색인어 가운데 중복을 제거한 단어는 4,267 개로 영어나 이태리어의 특성과 상당한 차이가 있음을 볼 수 있다. 그렇지만, 이 연구는 단행본 하나에 의한 연구이며, 음절에 대한 연구는 되어 있지 않다.

따라서 본 연구에서는 세가지 문서 부류에 대해 앞에서 설명한 C__PER와 D__PER의 값을 계산 하였다. 즉, 문자열형 필드(256Byte 이하)들을 포함하는 레코드 형태(문자열형)와 순수 단어 수가 약 5000자인 일반적이면서 비교적 짧은 문서 형태(단문형), 마지막으로 단어 수가 10만자가 넘는 긴 문서 형태(장문형)에 대한 것이다. 한편, 문자열형에 대한 실험을 수행할 때 사용한 실험 자료는 사진에 관련된 정보로서 사진 작가, 제목 사진 설명 등을 포함하는 2191건의 자료로 이들에 대해서는 각 레코드에 대한 평균을 취해 전체적인 결과를 계산하였다. 또한, 음절에 대한 분석도 실시 하여 음절의 단위로 1음절과 2음절을 단위로 사용 하였다. 그이유는, 음절 단위의 요약을 추출할 때 탈락 착오를 줄이기 위해 주로 1음절 방법, 2음절 방법, 혼합 방법이 사용되기 때문이다[12].

실험 방법은 공통 단어를 제외시키기 위해서 문자열형에 대해서는 모든 작업을 수작업으로 하였으며, 그외의 것들에 대해서는 한글 형태소 분석 알고리즘 KMA(Korean Morphological Analysis)를 사용하였고, 형태소 분석 결과 색인어를 추출하기 위한 방법으로, 형태소 분석 단계에서 모두 명사로 인식된 것만 색인어로 선정하는 엄격한 방법(Strict Method, SM)과, 분석 결과 명사의 비율이 높으면서 부사일 가능성이 없는 것을 색인어로 선정하는 느슨한 방법(Loose Method, LM)을 선택하였다. 단어 수에 대한 통계는 앞에서 제시된 인자를 가지고 하였으며, 1음절과 2음절에 대한 통계를 구하기 위해, 색인어로 추출된 단어를 각각 1음절과 2음절로 분리하여 중복을 제거한 다음, 1음절과 2음절에 대해 다음과 같이 단어 길이에 의한 비중복 음절수 통계를 구하였다.

1SP__PER = 비중복 색인어 내의 1SP수 / AWL

2SP__PER = 비중복 색인어 내의 2SP수 / AWL

* AWL : 문서 전체의 음절 수

여기서 한글의 1SP_PER와 2SP_PER 값은 영문에서 전체 문서 내의 비중복색인어 개수를 구하기 위한 인자와 동일한 의미를 지닌다. 즉,

D_PER(1 - C_PER) = NCDW / AW

	문자열형	단 문 형		장 문 형	
		SM	LM	SM	LM
AW	13.6595	5029		106282	
CW	6.5007	1942	1395	51675	41421
NCW	7.1588	3087	3634	54611	63861
NCDW	6.5354	515	619	7079	7686
C_PER	0.4759	0.3862	0.2774	0.4862	0.3991
D_PER	0.8876	0.1668	0.1703	0.1296	0.1204

〈 표1〉 단어 수에 대한 비중복 색인어 수

	문자열형	단 문 형		장 문 형	
		SM	LM	SM	LM
AWL	51.7147	13877		275420	
1SP수	15.3	288	307	839	870
2SP수	11.5	524	606	7296	7691
1SP_PER	0.2959	0.0208	0.0221	0.0030	0.0032
2SP_PER	0.2224	0.0378	0.0437	0.0265	0.0279

〈표2〉 음절수에 대한 색인어 내의 1음절과 2음절의 통계

〈표1〉을 보면 참고 문헌 [11]의 실험에서와 같이 비중복 단어 수의 통계치는 영문이나 이태리어 보다 상당히 높게 나타나고 있다. 그러나, 음절 단위의 실험 결과인 〈표2〉를 보면, 1음절과 2음절 모두 문서의 크기가 증가할 수록 1SP_PER, 2SP_PER가 감소함을 볼 수 있다. 또한 단문이나, 장문인 경우 영문의 단어 수에 대한 D_PER(1-C_PER)값인 0.01과 거의 비슷한 수준을 보이고 있다. 여기서 문자열형은 모든 색인어 통계치가 매우 높게 나타나고 있는데, 그것은 짧은 문장일 수록 대부분 명사 즉, 색인 후보어로만 이루어지기 때문이다.

3.2 한글 문서를 위한 요약 추출 기법

앞서 말한 것과 같이 한글 문서에 대한 요약을 추출할 경우에는 음절 단위로 하여야 한다. 여기서 고려해야 할 점은 해싱할 때의 음절 개수를 결정 하는 것이다. 지금까지 제시된 방법은 다음과 같은 3가지 경우이다.

Kopt = n*ln2 / m

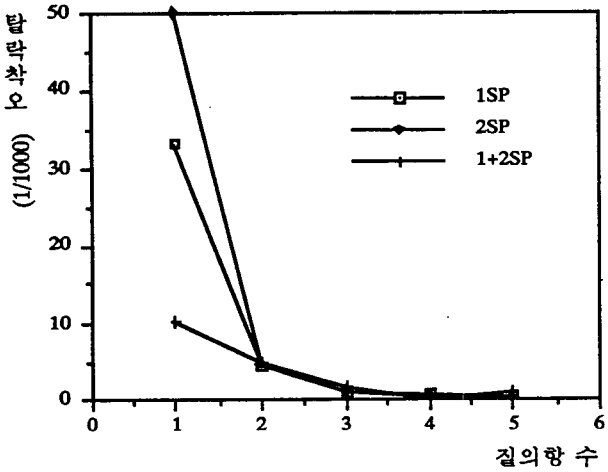
Kopt : 세트시켜야할 '1'의 수

n : 요약 길이

m : 레코드당 색인어 음절 수

이 경우 1SP방법에서는 6bit를, 2SP방법에서는 9bit를 세팅시켜야 함을 알 수 있다.

〈그림1〉은 질의항(query term)의 수에 따른 탈락 착오율을 나타낸다. 여기서, 1+2SP 방법이 질의항 수에 무관하게 거의 일정한 탈락 착오를 보임으로서 상당히 안정적임을 보이고 있다.



〈그림1〉 1SP, 2SP, 1+2SP의 질의항 수에 따른 탈락 착오

3.3 효율적인 해싱 함수

탈락 착오는 데이타 화일의 검색 시간을 최소화하는 가장 중요한 요소이다. 따라서, 요약 내에서 '1'이 주어진 범위 내에서 가능한 균등하게 분포되어 문서 요약의 밀도를 최적의 개수로 유지시키고 아울러 연산 시간도 짧은 해싱함수가 있어야 한다. 본 논문에서 선택한 해싱 함수[12]는 다음과 같고, 여기서 f1은 1SP방법에 f2는 2SP방법에 적용된다. 이 방법은 일반적으로 좋은 성능을 나타내는 것으로 판명되었다[13].

f1(c) = [p*unsigned_int(c)] mod m

f2(c1, c2) = [p1*unsigned_int(c1) + p2*unsigned_int(c2)] mod m

p, p1, p2 : 30보타 큰 서로 다른 솟수

m : 요약의 크기

(1) 1음절 단위(1 Syllable Pattern;1SP)의 코딩

(2) 2음절 단위(2 Syllable Pattern;2SP)의 코딩

(3) 1음절과 2음절의 혼합(1+2 Syllable Pattern;1+2SP)의 코딩

가장 간단한 1SP방법은 추출된 색인어를 1음절 단위로 해싱하여 요약을 만드는 방법으로 사용자의 모든 질의를 만족 시킬 수 있다는 장점이 있는 반면, 질의가 간단할 경우 높은 탈락 착오를 야기시키는 단점이 있다.

2SP방법은 한글의 단어가 보통 2음절 이상으로 구성된다는 특성을 이용한 방법이다. 이경우 복합명사의 띄어쓰기 처리가 문제가 된다. 따라서, 이경우 복합명사의 띄어쓰기에 무관하게 2SP가 추출되어야 한다. 그렇게 하기 위해서는 아래의 추출 예에서와 같이 복합명사의 띄어쓰기를 무시하고 1음절씩 중첩하여 요약을 추출한다. 그러나, 이방법은 2음절 이상의 검색조건을 처리할 때는 효과적이지만, '소', '말', '비' 등과 같이 한음절 단어는 처리할 수 없는 단점을 지닌다.

(추출 예)

데이타 베이스 시스템

데이타베이스 시스템

데이타베이스시스템

-----〉 2SP : 데이, 이타, 타베, 비이, 이스, 스시, 시스, 스템

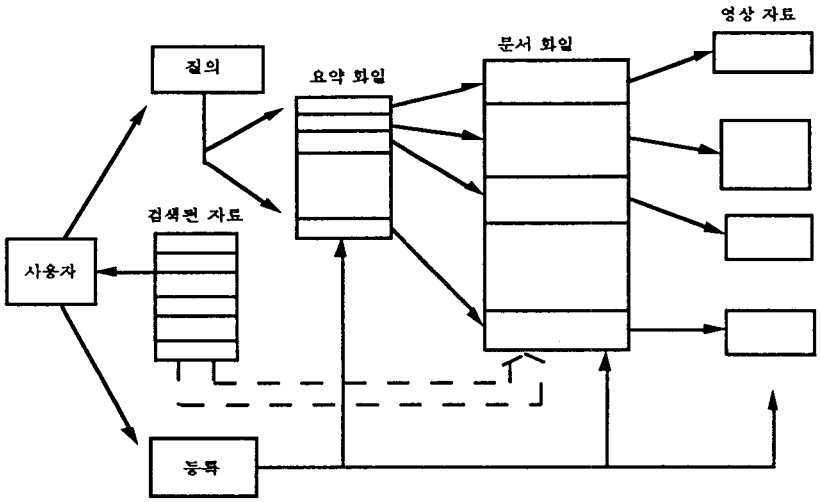
마지막으로 1+2SP방법은 1SP방법과 2SP방법을 결합한 형태이다. 즉, 하나의 요약 화일 안에 두 방법에 의해 추출된 요약을 같이 저장함으로써 각각의 단점을 보완하는 방법이다. 이방법은 요약 화일에 두방법의 요약을 저장하기 때문에 부가 저장 공간이 다소 증가하는 단점이 있지만 다음과 같은 장점이 있다.

첫째, 1음절을 포함하는 부분 매치(match) 질의를 처리할 수 있다. 둘째, 띄어쓰기 형태에 무관하게 질의에 해당되는 문서를 검색할 수 있다. 세째, 질의에 포함되는 색인어 수가 적을 경우에도 탈락 착오가 낮다.

본 논문에서는 앞의 문자열형 자료를 이용하여 각각의 방법에 따른 탈락 착오에 대한 실험을 수행 하였다. 요약의 크기는 각각의 레코드 크기가 평균적으로 90Byte이므로 전체 크기의 20%인 146을 요약의 크기로 잡을 수 있다. 하지만, 해싱값의 고른 분포를 위해 인접한 솟수 149로 요약 크기를 결정하였다. 주어진 요약 크기에서 탈락 착오를 최소화할 수 있는 비트 세팅 수는 다음 식에 의해 구할 수 있다.

IV. 시스템 구현

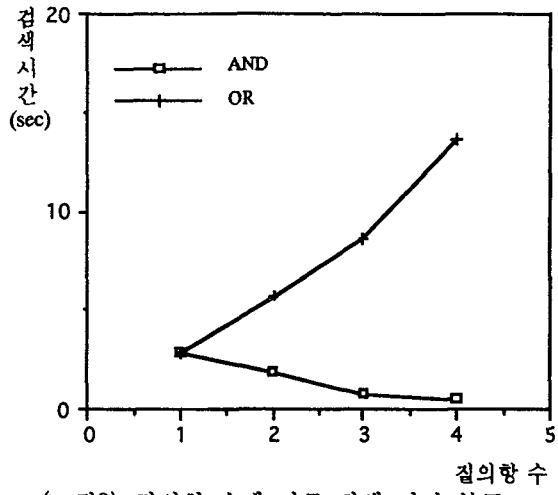
본 논문에서는 앞에서 제시한 2191건의 사진에 관련된 자료에 대해 실험용으로 정보 검색 시스템을 구현 하였다. 요약의 크기와 비트 세팅 수는 앞에서와 같고, 요약 화일은 1+2SP방법을 사용하였다. 시스템의 구성은 다음 〈그림2〉와 같다.



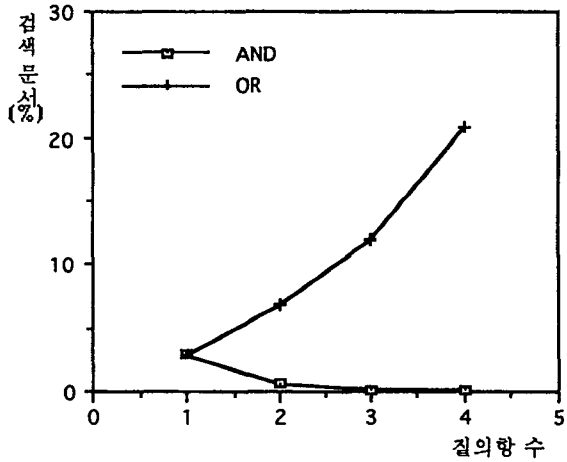
〈그림2〉 실험용 검색 시스템의 구성도

〈그림2〉와 같이 사용자가 질의를 하면, 먼저 요약 화일을 검색하여 후보 레코드를 얻은 다음 문서 화일을 검색하여 실제로 매치되는 레코드만을 "검색된 자료"로 보관 하고 사용자는 이를 순차 검색하게 된다. 한편, 질의 처리시 AND와 OR에 의한 검색을 지원 한다. 또한 등록의 경우 문서 화일에 문서를 저장하고, 문서가 저장된 위치와 함께, 색인어로 부터 추출된 요약을 요약 화일에 저장한다. 이 실험용 시스템에서, 그림은 각각 ID를 부여 하여 개개의 단일 화일로 관리 하였다.

시스템의 성능은 사용자가 AND, OR연산을 이용하여 다중항(multi-term)에 대한 질의를 하였을 때, 사용자가 질의를 한 후 최초 결과를 출력하기까지의 시간과 질의항에 따른 전체 문서중 검색된 문서 수에 대한 비율에 대해서 수행 하였으며, 그 결과는 〈그림3〉과 〈그림4〉와 같다. 〈그림3〉에서 보면 AND연산의 경우 질의항이 1개일 때 검색 시간은 평균 2.9초이나 질의항 수가 많아 질 수록 1초 이내의 검색 시간을 보이고 있다. 그러나, OR연산의 경우 질의항 수가 4개일 때 평균 13.67초가 걸린다. 그러나, 〈그림4〉를 보면 검색된 문서 수는 전체 문서의 21% 정도로 많은 문서가 검색 되었으며, 대부분의 검색 시간은 후위 탐색에 걸리는 시간이다.



〈그림3〉 질의항 수에 따른 검색 시간 분포



〈그림4〉 질의항 수에 따른 문서 검색 비율

V. 결론

본 논문에서는 일반적인 한글 문서에 요약 화일 기법을 적용하기 위해서 가장 필요한 문서의 비중복 통계를 구하였다. 여기서, 한글 단어의 특성상 단어 수에 따른 비중복 단어의 통계는 영문의 경우 보다 매우 높게 나타나고 있지만, 음절 단위로 하였을 경우 영문과 거의 비슷한 수준을 보임으로서 한글 문서도 아주 작은 요약으로 구성될 수 있다는 것을 보였다. 요약 추출 기법으로는 기존에 제시된 3가지 방법 중에서 1+2SP방법이 부가 저장 공간은 약간 증가하지만, 거의 균일한 탈락 착오를 가지며 사용자의 모든 질의 패턴을 만족시킬 수 있다.

앞으로의 연구 과제는, 비중복 색인어를 계산 하는데 있어 다양하고도 많은 문서에 대

해 연구하여 더 정확한 비중복 색인어 통계를 구해야 한다. 또한, 색인어를 효율적으로 추출할 수 있는 자동 색인이 되지 않으면, 완벽한 정보 검색 시스템이 될 수 없기 때문에, 앞으로는 효율적인 자동 색인 기법에 대한 연구가 수행되어야 한다.

참고 문헌

[1] S. Christodoulakis and C. Faloutsos, "Design Considerations for a Message File Server," IEEE Trans. on Software Engineering, Vol. 10, No. 2, pp.201-210, 1984.

[2] C. Faloutsos and S. Christodoulakis, "Signature Files : An Access Method for Documents and Its Analytical Performance Evaluation," ACM Trans. on Office Information Systems, Vol. 2, No. 4, pp.267-288, 1984.

[3] V.J. Calderbank, "TITAN : An Information Management System for Faster Retrieval from Massive Database Using Signatures, " Program, Vol. 24, No. 3, pp.253-268, 1990.

[4] D.E. Knuth, "The Art of Computer Programming, Volume 3 : Sorting and Searching," Addison-Wesley, 1973.

[5] F. Rabitti and J. Zizka, "Evaluation of Access Methods to Text Documents in Office Systems," Proc. of ACM SIGIR Conference, pp.21-40, 1984.

[6] C.S. Roberts, "Partial-match Retrieval via the Method of Superimposed Codes," Proc. of IEEE, Vol. 67, No. 12, pp.1624-1642, 1979.

[7] R. Sacks-Davis and K. Ramamohanarao, "A Two Level Superimposed coding Scheme for Partial Mathc Retrieval," Information Systems, Vol. 8, No. 4, pp.273-280, 1983.

[8] R. Sacks-Davis et al., "Multikey Access Methods Based on Superimposed Coding Techniques," ACM Trans. on Database Systems, Vol. 12, No. 4, pp.655-696, 1987.

[9] C. Faloutsos and S. Christodoulakis, "Design of a Signature File Method that Accounts for Non-Uniform Occurrence and Query Frequencies," Proc. of VLDB Conference, pp.165-170, 1985.

[10] J.W. Chang et al., "Multikey Access Methods Based on Term Discrimination and Signature Clustering," Proc of ACM SIGIR Conference, pp.176-185, 1989.

[11] 지능형 정보 검색에 관한 연구, 중간 연구 보고서, 한국 과학 기술원, 1991.

[12] 장재우, "한글 텍스트 검색을 위한 요약 화일 기법의 설계," 제3회 한글 및 한국어 정보처리 학술발표논문집, 한국인지과학회/한국정보과학회, pp.247-256, 1991.

[13] 송병호, 이석호, "한글 2음절 해싱함수의 비교," '92 봄 학술발표 논문집, 한국정보과학회, 19권, 1호, pp.643-646, 1992.