문자열 부분검색을 위한 색인기법의 설계 및 성능평가

강 승 헌† · 유 재 수††

요 약

신장해싱이나 B⁺-트리와 같은 기존의 색인구조들은 문자열의 부분검색을 지원하지 못하거나 부분검색에 제약점을 가지고 있다. 최근 웹 환경에서 동작하는 정보검색 엔진들이 사용하는 역 화일 기법과 요약 화일 기법, 또한 부분검색을 지원하지 못하거나 검색효율이 떨어지는 문제점을 갖는다. 본 논문에서는 역 화일의 빠른 검색성능을 가지면서 문자열 부분검색을 효율적으로 지원하는 색인기법을 제안한다. 제안된 색인기법은 기본적으로 역 화일 구조이며, 2음절 단위의 패턴으로 색인을 구성함으로써 문자열 부분검색을 지원한다. 제안된 색인기법의 특성을 분석하기 위해 제안된 방법의 성능을 다양한 환경에서 실험을 통하여 비교하고 분석한다. 또한 성능평가를 위해 기존의 역 화일 기법, 요약 화일 기법들과 제안하는 색인기법의 분석적 모델을 검색시간과 저장공간 측면에서 제시하고, 그 모델을 기반으로 그들의 성능을 비교한다. 분석적 비교모델을 통한 성능비교 결과, 제안된 부분검색을 위한 색인기법은 저장공간의 오버헤드는 크지만 기존 요약 화일 기법에 비해 검색성능을 상당히 향상시킨다.

Design and Performance Evaluation of an Indexing Method for Partial String Searches

Seung-Heon Kang† · Jae-Soo Yoo††

ABSTRACT

Existing index structures such as extendible hashing and B⁺-tree do not support partial string searches perfectly. The inverted file method and the signature file method that are used in the web retrieval engine also have problems that they do not provide partial string searches and suffer from serious retrieval performance degradation respectively. In this paper, we propose an efficient index method that supports partial string searches and achieves good retrieval performance. The proposed index method is based on the Inverted file structure. It constructs the index file with patterns that result from dividing terms by two syllables to support partial string searches. We analyze the characteristics of our proposed method through simulation experiments using wide range of parameter values. We also derive analytic performance evaluation models of the existing inverted file method, signature file method and the proposed index method in terms of retrieval time and storage overhead. We show through performance comparison based on analytic models that the proposed method significantly improves retrieval performance over the existing methods.

※ 본 논문은 정보통신부의 정보통신 우수 시범학교 지원사업에 의하여 수행된 것입니다.
† 준 회 원 : 충북대학교 대학원 정보통신공학과
†† 정 회 원 : 충북대학교 전기전자공학부 교수
논문접수 : 1998년 11월 21일, 심사완료 : 1999년 5월 3일

1. 서 론

현재까지 연구되고 개발된 색인구조들은 원하는 레코드에 대한 빠른 접근이 그 주된 목적이었다. 일반적으로 색인을 구성하는데 가장 많이 이용되는 것은 B⁺-트리라 할 수 있다. B⁺-트리는 색인을 구성할 때 기준 값보다 작거나 같은 값들은 왼쪽 서브트리에, 큰 값들은 오른쪽 서브트리에 위치시킨다[1]. 이러한 색인구조는 비교 대상을 대폭적으로 줄여 검색 속도를 향상시켜 준다.

그러나 문자열에 대해 B⁺-트리로 색인을 구성하였을 경우, 문자열 부분검색에 제약점을 갖게된다. 그 이유는 문자열에 대해서 B⁺-트리는 구조상 전방향절단검색(forward truncate search)만이 가능하기 때문이다. 즉, 질의어와 일치하거나 질의어를 문자열의 처음 부분에 포함하고 있는 값들만을 검색할 수 있다. 예를 들면, 질의어를 '데이타'라고 주었을 때 '데이타'로 시작하는 문자열들은 검색할 수 있지만 '삼성데이타시스템', '멀티미디어데이타'와 같이 '데이타'가 중간이나 끝에 오는 경우의 문자열들은 검색할 수 없게 된다.

최근 정보검색에서 역 화일과 요약 화일을 이용하여 검색서비스를 지원하고 있다[2]. 역 화일은 형태소 분석된 낱어로 색인을 구성하고 이 단어를 포함하고 있는 문서의 정보를 포스팅 화일에 유지한다. 역 화일은 검색성능은 우수하지만 문자열 부분검색이 불가능하고 형태소 분석이라는 추가적인 부담을 안고있다. 요약 화일에서는 단어요약을 중첩시켜 만든 문서요약을 이용하여 검색하기 때문에 문자열의 부분검색이 가능하다. 그러나 문서요약의 가중치(요약에서 '1'로 설정된 비트의 수)가 낮을 경우 성능 저하가 심해지는 단점을 가지고 있다. 또한 질의를 만족하지 않는 것을 결과로 가져오는 false match라는 문제점을 가지고 있다[3,4,5].

본 논문에서는 기존 색인구조들의 이러한 문제점을 해결하기 위해 문자열 부분검색을 효율적으로 지원하는 색인기법을 제안한다. 제안된 색인기법은 역 화일 구조를 따르며, 음절단위 패턴으로 색인을 구성하고 검색하므로써 문자열 부분검색을 가능토록 한다.

본 논문은 다음과 같은 순서로 구성된다. 2장에서는 정보검색에서 널리 이용되는 색인기법인 역 화일과 요약 화일에 대해 알아본다. 3장에서는 문자열 부분검색을 지원하는 색인기법을 제안하고, 4장에서는 제안된 색인기법에 대하여 성능평가를 한다. 마지막으로 5장에서는 본 논문의 결론을 내린다.

2. 정보검색에 이용되는 색인기법

2.1 역 화일

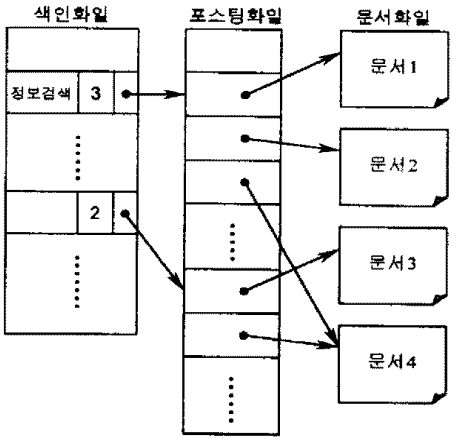
문서로부터 단어들을 추출해 내면 그 문서는 낱어들의 열로 표현할 수가 있다. 예를 들어, 문서 D가 단어(키워드) W1, W2, W3, ... Wm으로 구성되어 있다면 다음과 같이 표현할 수 있을 것이다.

$$D = (W1, W2, W3, \ldots Wm)$$

이것을 모든 문서에 대해 적용시킨 후 단어를 기준으로 바꾸어 표현하면 다음과 같다.

$$W1 = (\ldots, Dn)$$
$$W2 = (D1, D3, \ldots Dn)$$
$$W3 = (D2, D3, \ldots)$$
$$Wm = (D1, \ldots)$$

이와 같은 원리를 이용한 역 화일 기법은 색인 화일과 포스팅 화일로 구성된다. 색인 화일은 사용자의 질의 단어를 효율적으로 탐색하기 위하여 전체 데이타화일에서 추출한 색인어를 B-Tree, Hash 등의 인덱스 기법을 이용하여 구성한다. 포스팅 화일은 색인어와 색인어가 출현한 문서간의 관계를 저장하는 화일로서 색인어와 문서간의 밀접도(relatedness), 문서 내에서 색인어가 출현한 위치 정보 등으로 이루어진다. 다음 (그림 1)은 간단한 역 화일의 구조를 보여준다. (그림 1)에서 색일 화일의 두 번째 필드는 그 색인어가 가지고 있는 링크의 수를 나타낸다.



(그림 1) 역 화일의 구조

의 화일에서 색인 화일을 구성하는 단어들은 형태소 분석에 의해 얻어진다. 형태소 분석은 문서로부터 각 어절(띄어쓰기 단위)들을 분석하여 명사, 조사, 동사, 어미 등으로 분해하는 작업으로 단어 사전을 필요로 한다. 형태소 분석은 자연어처리의 가장 기본적인 작업이면서도 잘 풀리지 않는 어려운 문제이다. 형태소 분석에서 더욱 문제가 되는 것은 복합명사 및 고유명사 혹은 신조어의 존재이다. 이들은 생성 규칙이나 단어사전에 존재하지 않기 때문이다.

한편 정보검색에서 불용어(stopword)라고 불리는 검색에서 제외되는 단어들이 있다. 예를 들어, 영어로 된 문서의 경우 'a', 'the', 'of'를 가지지 않은 문서는 거의 없을 것이다. 이들 단어들은 거의 모든 문서에 존재하므로 검색의 의미가 없기 때문에 검색에서는 제외된다. 정보검색에서는 이런 성격의 단어들을 모아서 불용어라고 하여 특별히 검색에서 제거대상으로 삼는다[6]. 불용어를 사용하면, 색인의 크기가 급격하게 줄어들기 때문에 역 화일 기법에서는 불용어의 사용이 거의 필수적이다.

2.2 요약 화일

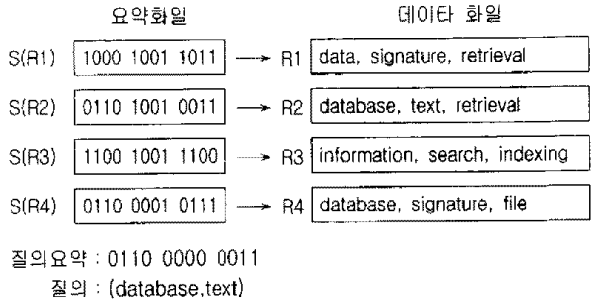
요약 화일 색인기법은 각 레코드(또는 텍스트)에 대한 요약(signature)을 요약 화일에 저장하고, 데이타 화일을 검색하기에 앞서 요약 화일을 검색하여 질의를 만족할 가능성이 있는 레코드만을 선택 접근함으로써, 데이타화일 검색 시간을 감소시키는 텍스트 검색 기법이다[7,9]. 일반적으로 각 레코드의 요약은 단어 요약을 비트별로 중첩하여 구성한다[5,10]. 각 단어에 대한 요약은 해싱을 사용하여 구성하며, 기존의 해싱 방법과는 달리 요약 화일에서 사용하는 해싱은, 해싱에 따른 결과 비트 스트링의 '1'의 개수가 일정하게 유지되는 방법이다[3,5]. (그림 2)는 레코드가 'database', 'text', 'retrieval'의 3개의 단어로 구성된 경우 레코드 요약을 만드는 과정을 나타낸다. 여기서는 m과 k는 각각 레코드 요약의 크기와 한 단어에 할당된 '1'의 개수를 나타낸다.

해싱 H : m=12, k=2

H(database)	: 0110 0000 0000
H(text)	: 0000 0000 0011
H(retrieval)	: 0000 1001 0010
레코드 요약	: 0110 1001 0011

(그림 2) 레코드 요약 생성

한편 요약 화일 기법에서 질의의 선택 조건을 만족하는 레코드를 검색하는 방법은 다음과 같다. 먼저 질의로부터 레코드 요약을 구성하는 방법과 같은 질의 요약을 만들고, 데이타 화일을 접근하기 전 단계로서 요약 화일을 접근하여, 질의 요약의 비트 패턴(bit pattern)을 포함하는 요약들을 추출한다. 이때 추출된 요약에 해당하는 레코드는 질의를 만족할 가능성이 있는 레코드로 간주하고, 이러한 레코드만 데이타 화일에서 최종적으로 검색하여 질의를 만족하는 레코드로 추출한다. (그림 3)은 4개의 레코드로 구성된 데이타 화일과 4개의 요약으로 구성된 요약 화일에서 질의에 대한 처리 과정을 보여준다.



(그림 3) 요약 화일의 질의 처리 과정

질의 단어 'database', 'text'로부터 질의 요약을 구성하고, 요약 화일의 각 요약 중에서 S(R2)와 S(R4)가 질의 요약을 포함하고 있으므로 질의를 만족할 가능성이 있다고 판정한다. 아울러 S(R2)와 S(R4)에 해당하는 데이타 화일의 R2와 R4를 실제로 검색하여 R2만이 두 개의 질의 단어를 포함하고 있으므로 질의를 만족한다고 한다. 이때 질의를 만족할 가능성이 있지만 실제로 질의를 만족하지 않는 R4는 false match 레코드라 한다[6,8].

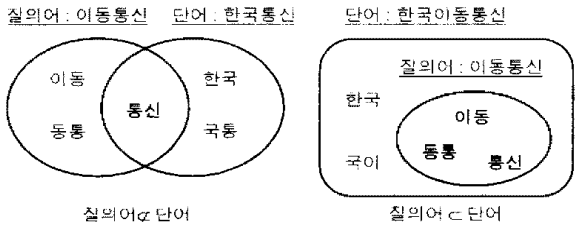
3. 제안된 문자열 부분검색을 위한 색인기법

3.1 문자열 부분검색 원리

본 논문에서 제안하는 문자열 부분검색을 위한 색인기법은 기본적으로 패턴 단위의 검색을 수행한다. 여기에서 패턴이란 문자열을 일정한 길이의 음절단위로 분할한 것이다. 본 논문에서는 한글정보에서 평균 어절의 길이가 2.75인 점을 감안하여 2음절 패턴을 사용한다[15]. 문자열을 패턴단위로 분할하는 것은 이미 설

명된 요약 화일 기법에서 분석요약을 만든 때 문서를 일정한 크기의 바이트로 나누는 것과 동일한 작업이다. 제안된 색인기법에서 패턴 분할의 주된 목적은 문자열 부분검색의 지원이며, 또한 정보검색의 역 화일에서 문제가 되었던 형태소 분석이라는 오버헤드를 줄일 수 있다. 문자열 부분검색 원리는 질의어와 비교되는 단어의 패턴들간의 집합관계를 이용한다. 즉, 질의어 패턴 집합이 비교되는 단어 패턴 집합의 부분집합일 경우 질의어는 비교되는 단어의 부분문자열이 된다. 다음은 패턴들을 이용하여 문자열 부분검색이 이루어지는 간단한 예이다.

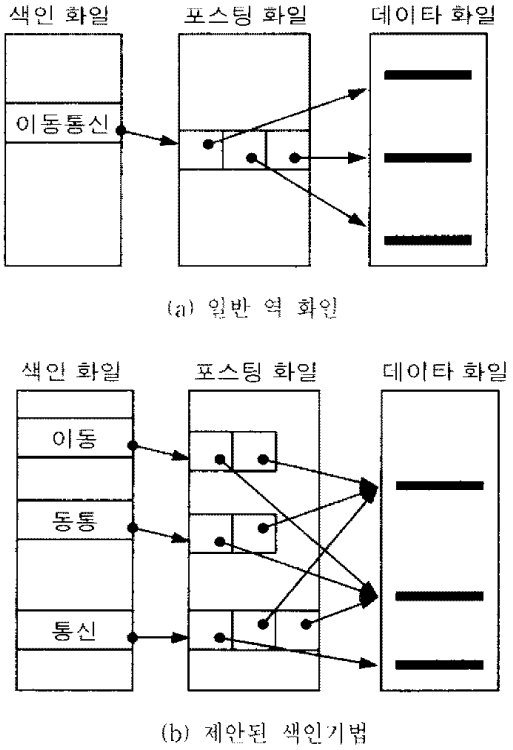
예를 들어, 패턴의 길이가 2음절이고 사용자의 질의어는 '이동통신'이며, 검색 대상이 되는 단어는 '한국통신'과 '한국이동통신'이라 하자. 질의어는 각각 '이동', '동통', '통신'의 질의 패턴으로 분할된다. '한국통신'을 단어 패턴으로 분할하고 질의 패턴들과 비교했을 때, 공통되는 것은 '통신' 하나이다. 반면에 단어 '한국이동통신'에 대해 같은 과정을 적용시키면, 질의어의 모든 패턴이 단어의 패턴 집합에 포함되는 것을 알 수 있다. 결과적으로, 질의어 '이동통신'의 패턴집합이 '한국이동통신'의 패턴집합의 부분집합이 되므로 '한국이동통신'은 질의어를 부분문자열로 포함하고 있다고 할 수 있다. 위의 예를 도식적으로 표현하면 (그림 4)와 같다.



(그림 4) 문자열 부분검색의 원리

3.2 문자열 부분검색을 위한 색인구조 및 검색 방법

본 논문에서 제안한 부분검색을 위한 색인구조는 정보검색의 역 화일과 동일한 구조를 갖는다. 정보검색에서 역 화일의 기본 단위는 단어(키워드)인 반면에 본 논문에서 제안하는 색인기법은 2음절 패턴을 기본 단위로 한다. (그림 5)는 단어 '이동통신'에 대한 색인을 일반 역 화일과 문자열 부분검색을 위한 색인기법으로 구성하는 것을 보여주고 있다. 일반 역 화일의 경우 데이타 화일에서 단어 '이동통신'이 존재하는 모든 위치 정보를 포스팅 화일에 유지하고, 이것을 색인 화일에서 포인트한다. 검색할 때에는 색인화일을 이용하여 포스팅 화일로부터 단어 '이동통신'이 위치한 데이타 화일의 주소들을 구하고, 이 주소들로부터 검색결과를 얻는다. 반면에 본 논문에서 제안한 문자열 부분검색을 위한 색인기법에서는 단어 '이동통신'을 2음절 패턴으로 나누고, 각각의 패턴에 대해 일반 역 화일에서와 같은 방법으로 색인을 구성한다. 검색할 때에는 역 화일을 만들 때와 같은 방법으로 질의어를 2음절 패턴들로 나누어 각각의 질의 패턴에 대하여 검색을 수행하고, 각 질의 패턴의 검색 결과들 중 공통되는 것만을 최종 검색결과로 취한다.



(그림 5) 일반 역 화일과 제안된 방법의 색인구조

3.3 불용패턴

앞에서 언급한 바와 같이 불용어는 역 화일의 크기를 줄이고, 무의미한 검색을 제거하여 검색 성능을 향상시키는데 그 목적이 있다. 본 논문에서 제안하는 문자열 부분검색을 위한 색인기법에서는 한 패턴이 임계치 이상의 출현 빈도수를 가질때 이 패턴을 불용패턴으로 처리하여 색인 및 검색에서 제외한다. 그러므로 검색기의 성능을 좌우하는 중요한 요소가 불용패턴의 임계치라 할 수 있다. 즉, 임계치가 높으면 더욱 많은

검색결과를 얻을 수 있지만 검색속도가 느려지고 색인의 크기가 커진다는 단점을 갖게된다. 반면에 임계치가 낮으면 많은 질의패턴이 걸러져 검색이 제대로 수행되지 않을 수도 있다. 그러므로 임계치는 시스템의 성능, 사용자의 질의, 검색시스템의 사용 환경 등 여러 가지 요소를 고려하여 결정되어야 한다.

3.4 문자열 부분검색 오류

본 논문에서 제안한 부분검색을 위한 색인기법에서는 질의어를 부분문자열로 가지지 않는 단어가 검색결과에 포함될 수 있다. 이러한 오류에는 두 가지 원인이 있다.

첫째, 불용패턴 처리로 인한 검색 오류이다. 불용패턴 처리가 되는 질의 패턴은 무시되기 때문에 불용패턴 처리되는 패턴의 위치에 다른 패턴이 있더라도 이를 처리하지 못해 잘못된 결과가 발생한다. 다음은 불용패턴 처리에 의한 검색 오류의 예이다.

질의어 : 한국통신
질의 패턴 : 한국(불용패턴 처리), 국통, 통신
검색 오류 : 동국통신, (주)홍국통신 등1

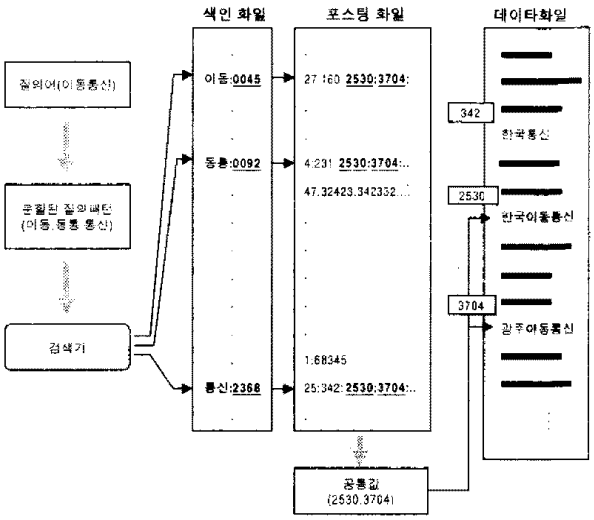
둘째, 검색 알고리즘으로 인한 검색 오류이다. 본 논문에서 제시한 알고리즘은 질의어에서 질의 패턴의 위치나 순서를 고려하지 않기 때문에 잘못된 결과가 발생한다. 다음은 검색 알고리즘에 의한 검색 오류의 예이다.

질의어 : 한국전관
질의 패턴 : 한국(불용패턴 처리), 국전, 전관
검색 오류 : 한국전력안전관리공사

두 가지 검색오류 중에서 후자의 발생 빈도는 극히 작다. 전자의 경우도 질의 패턴 수가 많아지면 검색오류는 거의 발생하지 않는다. 그러므로 불용패턴이 포함된 3~4음절의 질의어에 대해서는 부분일치 알고리즘[11,12,13]을 사용하여 검색오류를 제거해야 하지만 나머지 경우에 대해서는 부분일치 알고리즘을 적용할 필요성은 거의 없다.

3.5 부분검색 예

아래 (그림 6)은 질의어 '이동통신'의 부분검색 예이다. 질의어 '이동통신'은 질의 패턴 '이동', '동통', '통신'으로 분할된다. 각 질의 패턴을 키값으로 색인화일을 검색한 결과는 45, 92, 2368이다. 이 검색 결과가 가리키는 포스팅 화일의 위치에는 각 패턴을 포함하고 있는 단어의 주소들이 기록되어 있고, 주소들 중 공통된 값을 구하면 2530, 3704이다. 데이타 화일의 2530, 3704 위치에 있는 '한국이동통신'과 '광주이동통신'이 검색결과가 된다.



(그림 6) 부분검색의 예

3.6 제안된 색인기법의 특징

본 논문에서 제안한 색인기법은 3음절 이상의 패턴으로 색인을 구성하였을 경우 문자열 부분검색을 완벽히 지원하지 못한다. 예를 들어, 4음절 패턴을 기반의 색인에서 '한국이동통신'은 '한국이동', '국이동통', '이동통신'의 3개의 패턴으로 분할되어 색인된다. 이때, 질의어가 패턴의 길이보다 작은 '한국'이나 '통신'에 대하여 검색을 수행할 경우, 색인 화일은 4음절의 패턴으로 구성되므로 질의어와 일치하는 패턴은 존재하지 않게 되고 '한국이동통신'이 '한국'과 '통신'의 부분문자열을 가지고 있음에도 불구하고 검색은 실패하게 된다.

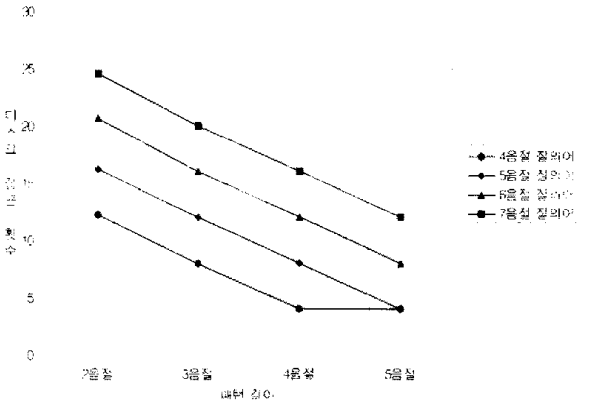
4. 성능평가

본 절에서는 패턴의 길이와 질의어의 길이가 검색성능 및 저장공간에 미치는 영향을 실험을 통하여 분석한다. 또한 제안된 색인기법의 검색성능 및 색인에 필요한 저장공간을 수학적 모델을 이용하여 기존의 역화일, 요약 화일과 비교한다. 본 실험에서는 불용패턴을 고려하지 않았다.

4.1 패턴 길이와 질의어 길이별 검색성능 및 저장공간 비교분석

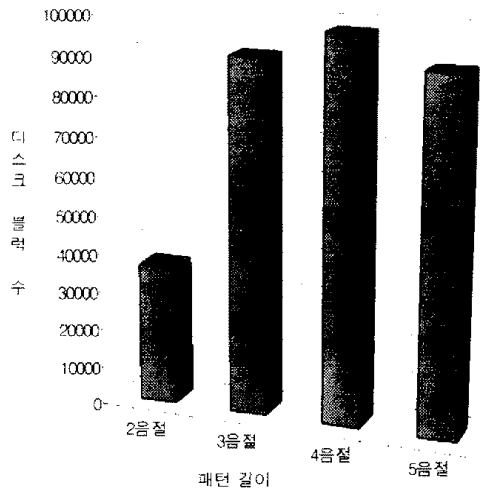
본 실험에서는 실제 데이타를 이용하여 제안된 색인기법으로 색인을 구성하고, 이에 대하여 검색시 평균 디스크 접근 횟수와 색인구성을 위해 요구되는 디스크 블럭 수를 조사하였다. 제시된 실험 결과는 메모리 64 MByte와 9GByte의 하드디스크를 갖춘 Sun Sparc station 5에서 실행한 결과이다. 시뮬레이션 프로그램은 C로 작성하였고 gcc로 컴파일 하였다. 실험데이타는 현재 한국통신 전화번호 안내 서비스에서 이용하고 있는 서울지역 상호 중 임의로 뽑은 50,245개를 이용하였고, 이들의 평균 길이는 5.65자이다. 한글은 완성형이며, 영문자, 숫자, 기호 등 한글 이외의 문자들은 제외하였다. 색인은 B+-tree로 구성하였고 디스크 블럭의 크기는 4Kbyte이다.

패턴 길이와 질의어 길이가 검색성능에 미치는 영향을 알아보기 위해 실험데이타에 대하여 2음절~5음절 패턴 기반으로 각각 색인을 구성하고, 각각에 대하여 4음절~7음절 길이의 질의어로 검색을 수행하였다. (그림 7)은 각각의 경우에 대하여 100회의 질의를 수행했을 때 평균 디스크 접근 횟수를 나타낸 것이다. 이때 질의어는 실험데이타에서 임의로 추출한 7음절 이상의 상호를 4음절~7음절로 분할한 것을 이용하였다. (그림 7)에서 나타난 결과를 보면 질의어 길이가 짧을수록, 패턴 길이가 길수록 검색성능이 우수해짐을 알 수 있다. 이것은 본 논문에서 제안한 색인기법이 질의어를 패턴 단위로 분할하였을 때 생성되는 패턴의 수만큼 검색을 반복 수행하기 때문이다.



(그림 7) 패턴 길이, 질의어 길이별 검색성능 비교

(그림 8)은 패턴 길이별로 실험데이타에 대하여 제안된 색인기법에 의해 색인을 구성하였을 때 요구되는 디스크 블럭 수를 나타낸 것이다. (그림 8)에 나타난 결과를 보면 패턴길이가 3음절 이상인 색인들은 2음절 패턴기반의 색인에 비해 월등히 많은 디스크 블럭이 요구되고, 4음절 패턴기반의 색인에서 최대값을 갖는다. 이와 같은 결과는 실험데이타로부터 생성되는 패턴 종류의 수와 관련이 있다. 완성형 한글이 2350자이므로 패턴의 길이가 n음절일 경우 최대 2350^n 종류의 패턴이 생성 가능하다. 그러므로 2음절 패턴기반의 색인이 3음절이상의 패턴기반 색인들에 비해 훨씬 적은 패턴 종류을 생성한다. 그러나 한글에는 자주 사용되지 않는 문자들이 상당히 존재하고, 패턴이 길어질수록 일반용어에서 사용되지 않는 패턴이 많아지기 때문에 패턴의 길이가 길어지더라도 생성되는 패턴은 한계를 가지게 된다.



(그림 8) 패턴 길이별 저장공간 비교

4.2 기존 색인구조와의 검색성능 및 저장공간 비교

본 실험에서 요약 화일, 역 화일, 제안된 색인기법의 검색성능 및 저장공간 관점의 수학적 모델을 이용하여 각 색인기법의 성능을 비교한다. 각 색인기법의 성능 분석을 위해 사용되는 입력 및 설계 인자들은 다음 <표 1>과 같다.

각 색인기법의 검색성능 및 저장공간은 다음과 같이 계산된다. 이때 요약 화일은 동적 요약 화일기법 중 하나인 HS 화일[16]을 사용한다. 제안된 색인기법에 대한 수식은 [14]의 역 화일 성능평가 수식을 2음절 패턴을 고려하여 수정한 수식이다. 다음 식에서 R은 검색시 평균 디스크 블럭 접근 회수를 나타내며, O는 색

〈표 1〉 입력 및 실계 매개 변수

기호	정 의
N	데이타의 총 레코드 수
P_y	디스크 블럭 크기(바이트)
A	질의를 만족하는 레코드 수
V	데이타 중 unique한 단어 수
h	B⁺-트리의 높이
c	한 단어(패턴)당 평균 포스팅 화일의 페이지 수
t	포인터 크기
w	단어의 평균 길이
n	질의 패턴 수
l	패턴 길이
S	데이타 중 unique한 패턴 수
hs	HS 화일의 높이
m	한 단어 요약에서 '1'로 설정되는 비트 수
n	한 단어당 frame 수
$\beta(i)$	HS 화일에서 단계 i의 평균 이용률
$p(x,i)$	HS 화일의 단계 i에서 x개의 위치가 '1'을 포함할 확률
b_1	HS 화일에서 단말노드의 blocking factor
b_2	HS 화일에서 내부노드의 blocking factor
lf	HS 화일에서 평균 페이지 load factor
f	HS 화일에서 frame 수

인을 위해 요구되는 디스크 블럭 수이다.

○ HS 화일

- 검색성능

$$R=\sum_{d=1}^{hs-2}\left(\prod_{i=1}^{d}\beta(i)p(m,i)\right)+\prod_{i=1}^{d}n\beta(i)p(m,i)+1+A$$

$\sum_{d=1}^{hs-2}\left(\prod_{i=1}^{d}\beta(i)p(m,i)\right)+1$는 내부노드를 위한 디스크 접근 회수, $\prod_{i=1}^{d}n\beta(i)p(m,i)$는 단말노드를 위한 디스크 접근 회수이며, A는 매칭된 레코드를 접근한 회수이다.

- 저장공간

$$O=\left\lceil\frac{N}{b_1\times lf}(1+f)\right\rceil+\left\lceil\frac{B_0}{b_2\times lf}\right\rceil+\sum_{i=1}^{hs-2}\left\lceil\frac{B_i}{b_2\times lf}\right\rceil$$

(단, $B_0=\left\lceil\frac{N}{b_1\times lf}\right\rceil$, $B_i=\left\lceil\frac{B_{i-1}}{b_2\times lf}\right\rceil$ 이고 i - 1, 2, …, $hs-2$ 이다.)

$\left\lceil\frac{N}{b_1\times lf}(1\times f)\right\rceil$는 단말노드 구성에 요구되는 디스크 블럭 수이며, $\left\lceil\frac{B_0}{b_2\times lf}\right\rceil$는 단말노드의 부모노드를 구성하는데 요구되는 디스크 블럭 수이고, $\sum_{i=1}^{hs-2}\left\lceil\frac{B_i}{b_2\times lf}\right\rceil$는 단말노드의 부모노드를 제외한 내부노드를 구성하는데 요구되는 디스크 블럭 수이다.

○ 역 화일

- 검색성능

$$R=h-1+c+A$$

(단, $h=\lceil\log_d V\rceil$, $d=\left\lfloor\frac{P_y-t}{w+t}\right\rfloor$,

$c=\left\lceil\frac{N\times t}{V\times p_y}\right\rceil$ 이다.)

$h-1$은 B⁺-트리 접근 회수이고, c는 포스팅 화일 평균 접근 회수이며, A는 매칭된 레코드를 접근한 회수이다.

- 저장공간

$$O=\left\lceil\frac{(w+t)\times V}{P_y}\right\rceil+c\times\left\lceil\frac{V}{u}\right\rceil$$

(단, $u=\begin{cases}\left\lfloor\frac{V\times p_y}{N\times t}\right\rfloor & \text{if } N\times t\le V\times p_y\\ 1 & \text{otherwise}\end{cases}$ 이다.)

$\left\lceil\frac{(w+t)\times V}{P_y}\right\rceil$는 B⁺-트리 구성에 요구되는 디스크 블럭 수이며, $c\times\left\lceil\frac{V}{u}\right\rceil$는 포스팅 화일 구성에 요구되는 디스크 블럭 수이다.

○ 제안된 색인기법

- 검색성능

$$R=n\times(h-1+c)+A$$

(단, $h=\lceil\log_d S\rceil$, $d=\left\lfloor\frac{P_y-t}{l+t}\right\rfloor$,

$c=\left\lceil\frac{N\times(w/2-1)\times t}{S\times p_y}\right\rceil$ 이다.)

$h-1$은 B⁺-트리 접근 회수이고, c는 포스팅 화일 평균 접근 회수이며, A는 매칭된 레코드를 접근한 회수이다. n은 질의 패턴의 수로 질의 패턴 수만큼 검색이 반복됨을 나타낸다.

저장공간

$$O = \lceil \frac{(l+t)\times S}{P_y} \rceil + c\times \lceil \frac{S}{u} \rceil$$

$$(\text{단, } u = \begin{cases} \lfloor \frac{S\times p_y}{N\times(w/2-1)\times t} \rfloor & \text{if } N\times(w/2-1)\times t \le S\times p_y \\ 1 & \text{otherwise} \end{cases} \text{이다.})$$

$\lceil \frac{(l+t)\times S}{P_y} \rceil$는 B^+-트리 구성에 요구되는 디스크 블럭 수이며, $c\times \lceil \frac{S}{u} \rceil$는 포스팅 화일 구성에 요구되는 디스크 블럭 수이다.

성능분석을 위해 사용된 실험데이타는 현재 한국통신 전화번호 안내 서비스에서 이용하고 있는 서울지역 상호 데이타 중 임의로 추출된 약 150만개를 이용하였다. 성능 비교를 위해 <표 2>와 같은 데이타베이스를 사용하며, 이는 실험데이타를 분석한 결과이다.

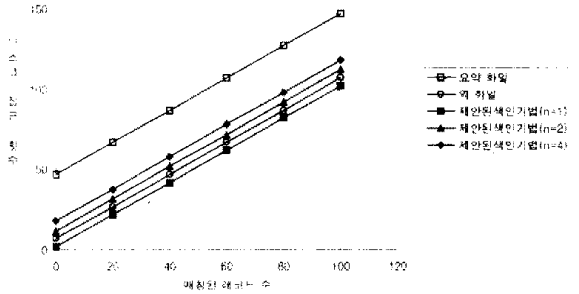
〈표 2〉 성능 비교에 사용된 데이타베이스

기호	값
N	1,538,829
P_y	4,096
m	17
V	327,796
t	4
w	12
l	4
S	96,573
A	1~100
b_1	NA
b_2	16
lf	0.5~0.8
f	NA

<표 3>과 (그림 9)는 <표 2>의 데이타베이스를 이용하여 각 색인기법들의 검색성능을 평가한 결과를 보여준다. 제안된 색인기법에 대한 질의어의 2음절 패턴 수(n)가 1, 2, 4개인 경우를 고려하였다. 실험결과를 보면 제안된 색인기법은 질의어의 길이별로 요약화일에 비해 31~95%, 30~91%, 27~83% 높은 검색성능을 보였고, 2음절 질의어에 대해서는 기존 역 화일보다 최고 25% 향상된 검색성능을 보였다. 이것은 완성형 한글의 2음절 패턴 수는 2350^2개이고 이들 중에서 사용되지 않는 패턴들이 상당수 존재하기 때문에 2음절 패턴으로 색인을 구성할 경우 색인 트리의 높이가 기존 단어에 의한 색인구성 방법보다 낮아진다. 따라서 검색을 1회만 반복하는 2음절 질의어에 대해서는 역 화일보다 향상된 검색속도를 나타낸다.

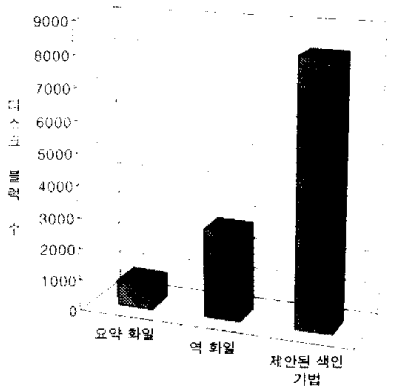
〈표 3〉 제안된 색인기법과 기존 색인구조와의 검색성능 비교 결과

	매칭된 레코드 수					
	0	20	40	60	80	100
요약화일	49	69	89	109	129	149
역화일	3	23	43	63	83	103
제안된 색인기법(n=1)	2	22	42	62	82	102
제안된 색인기법(n=2)	4	24	44	64	84	104
제안된 색인기법(n=4)	8	28	48	68	88	108



(그림 9) 제안된 색인기법과 기존 색인구조와의 검색성능 비교

(그림 10)은 각 색인기법의 저장공간을 평가한 결과를 보여준다. 동일한 데이타에 대해 제안된 색인기법으로 색인을 구성할 경우 기존 역 화일의 약 2.9배, 요약 화일의 약 10배로 저장공간의 오버헤드가 큰 것으로 나타났다. 그러나 제안된 색인기법은 문자열 부분검색을 완벽하게 지원하고 요약 화일에 비해 우수한 검색성능을 가지며, 2음절 질의어에 대해서는 역 화일보다 높은 검색성능을 보였다는데 그 의미가 크다.



(그림 10) 제안된 색인기법과 기존 색인구조와의 저장공간 비교

5. 결 론

본 논문에서는 최근 정보 검색 분야에서 널리 이용되고 있는 역 화일을 응용하여 문자열 부분 검색을 지원하는 색인 기법을 제안하였다. 이를 위해 역 화일과 요약 화일에 대해 간략하게 알아보았고, 이를 기반으로 문자열 부분 검색의 원리 및 이를 위한 색인 방법과 검색 방법을 제안하였다. 제안된 문자열 부분 검색을 위한 색인 기법에서는 역 화일을 형태소 분석 결과 단어로 구성했던 기존의 방법과는 달리 단어를 일정한 크기로 분할한 패턴으로 색인을 구성하였다. 검색할 때에는 분할된 질의 패턴별로 검색을 수행하고, 공통된 결과를 추출하여 최종 결과를 얻게 된다.

제안된 색인기법은 패턴 길이와 질의어 길이에 따라 검색성능이 차이를 보였다. 이것은 패턴의 길이가 생성되는 패턴 수를 결정하고 이것은 검색회수에 영향을 주기 때문이다. 저장공간 이용 효율에서도 패턴의 길이별로 큰 차이를 보였다. 또한 요약 화일에 비해 우수한 검색성능을 보였으며, 2음절 질의어에 대해서는 역 화일보다 다소 높은 검색성능을 보였다.

제안된 색인기법은 기존 색인 구조에 비해 저장공간의 오버헤드가 크지만 완벽한 문자열 부분검색을 지원하고, 기본적으로 역 화일기법을 사용함으로써 역 화일의 빠른 검색성능을 보였다. 향후 연구과제로 저장공간 문제를 해결하기 위해서는 효율적인 포스팅 화일 압축 기법이 연구되어야 할 것이다.

참 고 문 헌

[1] R. Elmasri and S.B. Navathe, "Fundamentals of Database Systems," Benjamin, 1989.

[2] William B. Frakes and Ricardo Baeza-Yates, "Information Retrieval Data Structures & Algorithms," Prentices Hall, 1992.

[3] 유재수, 한세기, 김명호, 이윤준, "HS화일:효율적인 정보 검색을 위한 동적 요약 화일 방법", 한국정보과학회논문지, pp.994-1004, 1994. 6.

[4] 유재수, 최한석, "효율적인 요약 화일 구성을 위한 경험적 결집 방법", 한국정보과학회논문지, pp. 1625-1632, 1995.12

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

[6] Salton G. and McGill M.J., "Modern Information Retrieval," McGraw Hill Book Company, 1983.

[7] Faloutsos C., "Signature-based Text Retrieval Methods," A Surbey IEEE Computer Society Technical Committee on Data Engineering, Vol. 13, No.1, pp.25-32, 1990.

[8] Dattola R, "FIRST : Flexible Information Retrieval System for Text," JASIS, Vol.30, No.1, pp. 207-212, 1979.

[9] Faloutsos C., "Access Methods for Text," ACM Computing Survey, Vol.17, No.1, pp.49-74, 1985.

[10] J. R. Files and Huskey, H.D., "An Information Retrival System Based on Superimposed Coding," Proc. of the Fall Joint Computer Science, pp. 423-432, AFIP press, 1969.

[11] Boyer R. and S. Moore, "A Fast String Searching Algorithms," CACM, Vol.20, No.11, pp.762-772, 1977.

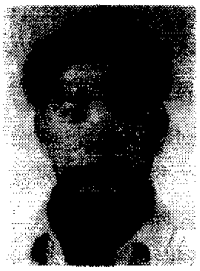
[12] Guibas L. and A. Odlyzko, "A New Proof of the Linearity of the Boyer-Moore String Searching Algorithm," SIAMJ on Computing, Vol.9, No.4, pp.672-682, 1980.

[13] Knuth D. and J. Morris, "Fast Pattern Matching in Strings," SIAMJ on Computing, Vol.6, No.2, pp.323-350, 1977.

[14] 장재우, "단어 분별도를 이용한 요약 다중키 접근 기법의 설계 및 평가", 박사학위논문, 1990.

[15] 송병호, 이석호, "2음절 패턴을 이용한 한글 요약 추출 기법", 한국정보과학회논문지, pp.1179-1190. 1993, 8.

[16] J.S. Yoo, K.S. Choi and M.H. Kim, "Analysing Performance Characteristics of Dynamic Signature File Methods," Journal of Electrical and Information Science, Vol.2, No.4, pp.37-45, 1997.



강 승 헌

e-mail : k1000@pretty.chungbuk.ac.kr

1997년 목포대학교 전산통계학과 졸업(학사)

1999년 충북대학교 정보통신공학과 졸업(공학석사)

1999년~현재 충북대학교 정보통신공학과 박사과정

관심분야 : 데이터베이스시스템, 정보검색, XML문서 구조 검색

유 재 수

e-mail : yjs@pretty.chungbuk.ac.kr

1988년 신북대학교 컴퓨터공학과 졸업(학사)

1991년 한국과학기술원 전산학과 졸업(공학석사)

1995년 한국과학기술원 전산학과 졸업(공학박사)

1995년~1996년 8월 목포대학교 전산통계학과 전임강사

1996년 8월~현재 충북대학교 정보통신공학과 조교수

관심분야 : 데이터베이스시스템, 정보검색, 멀티미디어 데이터베이스, 분산객체컴퓨팅