한글 텍스트 검색을 위한 요약 화일 기법에 관한 연구*

송 병호, 이 석호
서울대학교 컴퓨터공학과 데이타베이스연구실

A Reasearch on Signature File Methods for Korean Text Retrieval

Byoungho Song, Sukho Lee
Database Lab., Dept. of Computer Eng., Seoul Nat'l University

요 약

텍스트에 대한 내용 본위 검색 기법으로서 요약 화일(signature file) 기법은 역화일(inverted file)이 허용되지 않을 때 매우 유용하다. 그러나 한글은 영문과 달리 어절의 형성이 복잡하고 띄어쓰기 형태가 고정되지 않음에 따라 기존의 단어 위주 영문 본위 요약 화일 기법을 그대로 적용시킬 수 없다. 본 논문에서는 이를 위하여 띄어쓰기를 무시하고 중복된 2음절 패턴을 도출하여 요약 화일을 구성, 검색하는 기법을 제안한다. 이 기법은 일본어, 중국어 등 비슷한 문제를 가진 외국어에도 적용될 수 있다.

I. 서론

점증되는 사무 자동화나 멀티미디어(multimedia) 정보 처리의 필요성에 따라 사무 문서(office document), 특히 그 중에서도 문자 정보로 이루어진 텍스트(text)에 대한 검색 기법의 연구가 많이 진행되어 왔다.

* 본 연구는 한국 과학재단 목적기초연구 89-0103-12의 지원금에 의한 것임.

가장 간단하게 구현할 수 있는 방법으로는 텍스트에 제목, 분류번호나 키워드 리스트를 별도의 애트리뷰트로 첨부하여 그에 대한 조건 검색을 생각할 수 있으며, 기존의 대다수의 광화일 시스템들이 이러한 기법을 이용하고 있다. 그러나 이러한 애트리뷰트로써 텍스트의 내용을 충분히 표현하기 어려우므로 결국은 텍스트 자체의 내용을 조건 검색하는 내용본위(content-based) 검색 기능이 필요하다.

텍스트(text) 또는 문서(document)에 대한 내용본위(content-based) 검색 기법은 오래전부터 연구되어 왔으며 근래에 멀티미디어 데이타베이스에 대한 연구가 활발해지면서 더욱 중요시되고 있다.

텍스트에 대한 내용본위 검색으로는 텍스트 전체를 순차적으로 탐색하는 전 텍스트 주사(full text scanning), 유효한 단어 각각에 대한 인덱스를 전부 만드는 역 화일(inverted file), 텍스트를 대표하는 요약으로 가능성 없는 텍스트를 미리 골라 내는 요약 화일(signature file) 및 관련되는 텍스트끼리 물리적으로 가까운 위치에 모아 두는 클러스터링(clustering) 기법 등이 있으나 역 화일이 허용되지 않을 때 요약 화일이 매우 우수한 것으로 평가되고 있다.[4,6,7]

그러나 기존의 요약 화일 기법은 단어를 단위로 처리하고 있어 단어의 구분이 모호한 한글에 적용하기는 어렵다.

본 논문에서는 이러한 한글의 특성에 맞는 요약 화일 기법에 대해서 논한다. II장에서는 요약 화일의 기본 개념 및 한글 텍스트에 적용시 문제점을 살펴 보고 III장에서는 부 스트링을 처리할 수 있는 요약 화일의 변형 기법을 살펴 보며 IV장에서는 한글 요약 화일 기법을 제안한 뒤 V장에서 성능을 분석한다.

II. 요약 화일의 문제점

요약 화일의 개념 자체는 오래되었으나[10] 그 개념을 정립하게 된 것은 1983년 Faloutsos에 의해서였으며 그 이후 효율의 개선에 대한 연구는 많았으나 그 골격은 계속 유지되고 있다.

그 개념은 원 텍스트의 내용 중에서 중복되지 않고(distinct) 관사등 많이 쓰이는(common) 것을 제외한 단어들을 하나씩 해싱하여 단어 요약(word signature)을 만든 뒤 이들을 중첩하거나(superimposed coding) 접합하거나(concatenation) 중첩한 뒤 압축하여 원 텍스트보다 훨씬 작은(약 10%) 요약 화일을 만든 뒤 검색시 요약 화일을 이용하여 가능성 없는 대다수의 텍스트들을 배제하고 나머지를 전 텍스트 주사하는 것이며[6,7] 다음과 같은 특성을 목표로 한다.[8]

1) 요약은 원 텍스트보다 훨씬 작으므로 이에 대한 탐색은 효율적이다.

2) 검색 조건에 부합되는 텍스트는 요약으로도 발견된다.(즉, 선택 착오(false hit)가 없다.)

영문은 형태상 굴절어(inflectional language)로서 실사와 허사의 구별이 뚜렷하지 않고 어형이나 어미의 굴절로 어법 관계를 나타내므로, 띄어쓰기가 용이할 뿐 아니라 띄어쓰기 단위(word)가 곧바로 검색조건 키워드로 사용되기 용이하다.

그러나 한글은 교착어(agglutinative language)로서 조사, 어미, 보조어간 등 어법 관계를 나타내는 허사가 있어, 이것이 의미를 나타내는 명사등 실사에 붙음으로써 어법 관계를 나타내므로 띄어쓰기 단위인 어절의 구성 형태가 복잡하다. 또한 복합어나 한자어, 외래어의 경우 등에 있어 띄어쓰기가 자유롭다. 따라서 어절을 단위로 해싱하여 요약 화일을 생성하는 것은 무의미하다.

국내에서 현재 쓰이고 있는 DACOM의 천리안, IBM의 HAIRS 등의 한글 문서 인덱싱 기법은 이러한 특성을 무시한 채 띄어쓴 어절 단위로 인덱싱을 함으로써 전방 일치 기능만 제공하고 있다.

한글에서의 정보 검색은 다음과 같은 특성을 갖는다.

1) 어절의 일부분도 의미가 있을 수 있다.
   예) 주택청약통장은...
2) 띄어쓰기의 분리자(delimiter)인 빈칸이 무의미할 수 있다.
   예) 주택 청약 통장 = 주택청약통장

따라서 한글 텍스트 검색을 위해서는 기존의 단어 위주가 아니라 부단어(subword)와 복합 단어(compound word)를 찾아 낼 수 있는 요약 화일을 구성하여야 한다.

III. 기존 연구

이러한 문제와 관련하여 영문의 경우 부 스트링을 처리할 수 있는 기법들이 몇몇 발표되었다.

Faloutsos[5]는 한 단어의 일부를 찾을 수 있도록 단어 전체를 해싱하지 않고 각 단어를 연속되고 중복되는(successive,overlapping) 세 문자(letter triplet)로 나누어 해싱하는 기법을 제안하였다. 이에 따르면 'free'라는 단어는 ' fr', 'fre', 'ree', 'ee '의 네 조각으로 나뉘어진다.

Gebhardt[8]는 Faloutsos의 세 문자 단위가 불균일 빈도를 가지며 서로 다른 단어에 있는 것끼리 조합되어 해당 단어가 있는 것처럼 검색되는 조합 착오(false combination) 확률

이 높으므로 네 문자 단위(letter quadruplet)가 보다 효율적이라고 하였다.

이상의 방법들은 분할이 한 단어 내에서만 이루어져 복합 단어를 처리하지는 못하고 있다.

텍스트를 빈칸에 의하여 경계지워진 단어의 모임으로 보지 않고 일종의 단일 스트링으로 취급함으로써 복합 단어까지 처리할 수 있는 기법이 발표되었다.

Harrison[3,9]은 스트링 S_1이 스트링 S_2의 부 스트링이라면 S_1의 부 스트링의 집합은 S_2의 부 스트링의 집합의 부분집합이 됨을 이용하여 원 텍스트와 검색 키워드 모두 길이가 k인 부 스트링으로 나누어 k-요약(k-signature)을 만들어 비교하는 방법을 제안하였다.

Barton등[2]은 고정 길이보다 출현 빈도가 균일한 가변 길이의 256개 키(key)로써 모든 영문 스트링을 효율적으로 표현할 수 있다고 하였다.

그러나 이러한 방법들도 복합 단어를 처리할 수는 있으나 원 텍스트의 띄어쓰기 대로 검색 키워드가 주어져야 하므로 한글 텍스트에는 부적절하다.

IV. 제안 기법

송[1]은 Harrison의 기법과 유사하게 한글 텍스트를 2음절 패턴으로 중복되게 분할하여 그 요약을 중첩하는 기법을 제안하였다. 그러나 이 방법은 한 음절 내에서 분할하였기 때문에 띄어쓰기의 문제를 완전히 해결하지는 못하였다.

본 고에서 제안하는 기법의 기본 개념은 한글의 띄어쓰기를 무시하고 모두 붙여 쓴 것으로 간주하여 연속되고 중복되는 2음절 패턴의 집합을 생성해 내는 것이다. 예를 들어 다음과 같은 원 텍스트가 주어지면

"유가와 이론은 유가와입자를 예언하였다"

텍스트의 요약은 그림 1과 같이 중복된 2음절 패턴의 요약을 중첩해서 만든다.

검색 키워드가 "유가와 입자"로 주어지면 같은 방법으로 키워드 요약이 만들어져(그림 2) 키워드 요약에 '1'로 세팅된 위치에 모두 '1'이 있는 텍스트 요약을 찾음으로써 검색이 이루어진다. 단, 이 방법으로는 한 음절로 이루어지는 키워드는 검색 조건으로 사용할 수 없다.

이와 같은 방법을 사용하면 어절의 부분 탐색뿐만 아니라 원 텍스트와 키워드의 띄어쓰기 방식이 다르더라도 찾을 수 있다.

유가	0010 0000 0000 0000 0000 0000
가와	0000 0100 0000 0000 0000 0000
와이	1000 0000 0000 0000 0000 0000
이론	0000 0000 0001 0000 0000 0000
론은	0000 1000 0000 0000 0000 0000
은유	0000 0000 0000 0001 0000 0000
와입	0001 0000 0000 0000 0000 0000
입자	0010 0000 0000 0000 0000 0000
자를	0000 0001 0000 0000 0000 0000
를예	0000 0000 0000 0000 0010 0000
예언	0000 0000 0000 0000 1000 0000
언하	0000 0000 0100 0000 0000 0000
하였	0000 0000 0000 0000 0000 0001
였다	0000 0000 0000 0010 0000 0000
	1011 1101 0101 0011 1010 0001

<그림 1> 원 텍스트 요약 예

유가	0010 0000 0000 0000 0000 0000
가와	0000 0100 0000 0000 0000 0000
와입	0001 0000 0000 0000 0000 0000
입자	0010 0000 0000 0000 0000 0000
	0011 0100 0000 0000 0000 0000

<그림 2> 키워드 요약 예

V. 성능 분석

정보검색(IR)분야에 있어서 검색 조건에 부합되지 않는 텍스트가 요약화일상으로는 부합될 가능성이 있는 것으로 나타나는 확률을 지칭하는 '배제 착오'(false drop) F_d는 다음과 같이 정의된다. 이 F_d는 텍스트 검색 기법의 효율성을 따지는 가장 중요한 요소이다.

F_d = 요약화일상으로 부합되는 것으로 나타나는 텍스트의 수
/ 실제로 부합되지 않는 텍스트의 수

본 기법에서의 F_d는 각 텍스트에 대한 요약의 길이를 m-bit, 검색조건 키워드 내에 나타

나는 2음절 패턴의 갯수를 l_1, 텍스트 내에 나타나는 2음절 패턴의 갯수를 l_2라고 했을때 [9] 및 [10]에 의하여 다음과 같은 수식으로 표현된다.

$$F_d \cong (1 - e^{-l2/m})^{l1}$$

실험적으로는 1KB의 한글 텍스트에서 발생하는 2음절 패턴의 갯수는 평균 370.8개, 중복을 배제한다면 328.6개가 나타나며 요약의 길이를 800 비트(원 텍스트의 10%)로 할 경우 평균 240.5 비트가 '1'로 세팅됨을 확인하였다. 이것은 같은 크기의 영문 텍스트에서 해싱 대상 단어 갯수 D가 평균 40개[6,7]임과 비교하면 상당히 큰 숫자처럼 보인다. 그러나 영문에서도 부 스트링 검사를 위해 3문자 단위나 4문자 단위로 나눈다면 해싱 대상이 몇 배나 커지게 되므로 결국 영문과 비교해 볼 때 그리 큰 숫자는 아니다.

이와 같은 실험적 통계치에 의하여 m=800, l_2=328.6을 대입하면

$$F_d = 0.3368^{l1}$$

으로서 각 l_1값에 따른 F_d의 변화는 표 1과 같다.

l_1	F_d	l_1	F_d	l_1	F_d	l_1	F_d
1.0	0.3368	2.0	0.1135	3.0	0.0382	4.0	0.0129
1.1	0.3021	2.1	0.1018	3.1	0.0343	4.1	0.0115
1.2	0.2710	2.2	0.0913	3.2	0.0307	4.2	0.0104
1.3	0.2430	2.3	0.0819	3.3	0.0276	4.3	0.0093
1.4	0.2180	2.4	0.0734	3.4	0.0247	4.4	0.0083
1.5	0.1955	2.5	0.0659	3.5	0.0222	4.5	0.0075
1.6	0.1753	2.6	0.0591	3.6	0.0199	4.6	0.0067
1.7	0.1573	2.7	0.0530	3.7	0.0178	4.7	0.0060
1.8	0.1411	2.8	0.0475	3.8	0.0160	4.8	0.0054
1.9	0.1265	2.9	0.0426	3.9	0.0144	4.9	0.0048

<표 1> l_1값에 따른 F_d의 변화(m=800, l_2=328.6)

그러나 한글에서도 영문에서와 마찬가지로 일반적으로 많이 쓰이는 단어들이나 조사, 동사 등을 분리해 낼 수 있다면 l_2값을 상당히 줄일 수 있으므로 효율을 크게 향상시킬 수 있을 것이다.

VI. 결론

한글 텍스트에서 어절의 부분 탐색과, 띄어쓰기 방식의 다름에 상관없는 탐색을 함으로써 실제 업무에 내용 본위 검색 기법으로 사용될 수 있는 한글 요약 화일 기법을 제안하였다.

현실적으로 검색 조건이 한 단어 또는 한 어절로 이루어지는 경우는 거의 없으므로 본 기법은 검색조건 키워드가 복잡해질수록 유용할 것으로 기대된다.

이 기법은 비슷한 띄어쓰기 문제를 가진 일본어, 중국어뿐 아니라 'Data Base'와 'Database'로 병기하는 일부 영문의 경우에도 적용 가능하다.

앞으로 이 기법에 대한 세밀한 정량 분석이 필요하며, 근본적으로는 한글 텍스트에 있어서 사용자들의 검색 경향에 대한 분석이 필요할 것으로 본다. 아울러 해싱 대상 패턴 수를 줄이기 위하여 키워드가 될 가능성이 없는 형태소 또는 단어들을 분리해 내는 기법에 대한 연구도 필요하다.

* 참고 문헌 *

[1] 송병호,이석호, "요약 화일을 이용한 한글 사무문서 인덱싱 기법에 관한 고찰," '90 봄 학술발표 논문집, 한국정보과학회, 17, 1, 265-268, Oct. 1990.

[2] Barton I.J., Creasey S.E., et. al., "An Information Theoretic Approach to Text Searching in Direct Access Systems," CACM, 17, 6, 345-350, June 1974.

[3] Bookstein A., "On Harrison's Substring Testing Technique," CACM, 14, 12, 180-181, Dec. 1971.

[4] Burkowski F.J., "Surrogate Subsets: A Free Space Management Strategy for the Index of a Text Retrieval System," 13th ACM-SIGIR, 211-226, Sep. 1990.

[5] Tsichritzis d., Christodoulakis S., et. al., "A multimedia Office Filling System," 9th VLDB, 2-7, Nov. 1983.

[6] Faloutsos c., "Signature Files: Design and Performance Comparison of some Signature Extrection Methods," ACM-SIGMOD, 63-82, May 1985.

[7] Faloutsos c., "Access Methods for Text," ACM Computing Surveys, 17, 1, 49-74, Mar. 1985.

[8] Gebhardt F., "Text Signatures by Superimposed coding of letter triplets and quadruplets," Information Systems, 12, 2, 151-156, 1987.

[9] Harrison M.C., "Implementation of the Substring Test by Hashing," CACM, 14, 12, 777-779, Dec. 1971.

[10] Knuth D.E., "The Art of Computer Programming," Vol. 3, "Sorting and Searching," 550-567, A-W Pub. Co., 1973.