점증색인과 복합명사 처리를 위한 한국어 텍스트 색인 저장구조[1]

김평ㅇ, 주원균, 장동현, 맹성현
충남대학교 컴퓨터과학과

An Index Storage Structure for Incremental Indexing and Compound Nouns in Korean Texts Retrieval

Pyung Kim, Won-Kyun Joo, Dong-Hyun Jang, Sung-Hyun Myaeng,
Dept. of Computer Science, Chungnam National University

요 약

텍스트를 내용에 기반하여 검색하는 기법은 DBMS 의 발전과 병행하여 독립적으로 발전해 왔다. 본 논문에서는 정보검색 시스템과 DBMS 의 통합을 전제로 하여 새로운 문서의 점증적 추가를 효율적으로 처리할 수 있는 색인 구조를 기술한다. 또한 한국어 텍스트에 자주 출현하는 복합명사를 색인어로 사용할 때 대두되는 부분정합(partial matching)을 효율적으로 처리할 수 있는 새로운 색인구조의 개발 결과를 소개한다.

1. 서 론

정보검색 시스템에서 사용되는 색인어는 대부분 문서에서 명사를 위주로 추출되고 그에 따라 색인정보가 만들어지게 된다. 특히 한국어의 경우 복합 명사가 많은 비중을 차지하고 있으며 새로 창출되는 단어들도 역시 기존 명사의 조합으로 이루어진 복합명사의 경우가 많다. 그에 따라 정보검색 시스템에서 복합명사를 처리하는 문제가 관심을 끌게 되었고 그에 따라 문서에서 복합명사를 추출하는 방법은 물론 추출된 명사를 이용하여 색인정보를 효율적으로 구축하고 복합명사의 부분정합(partial matching)을 가능하게 검색하여 주는 시스템이 요구되었다. 또한 인터넷의 보편화로 인한 정보망 사용의 확대로 정보의 공유가 쉬워지면서 새로운 정보가 생성되고 축적되는 속도가 급속히 증대함에 따라 전체 DB 의 재색인 없이 기존의 저장 구조에서 새롭게 추가되는 정보만 처리할 수 있는 시스템이 요구되었다. 본 논문에서는 근본적으로 색인 데이터를 저장하는 구조로서 점증적으로 유입되는 문서를 동적으로 색인하여 저장하면서 한국어 텍스트의 특성 중의 하나인 복합명사를 효율적으로 처리할 수 있는 데에 주안점을 두고 있다.

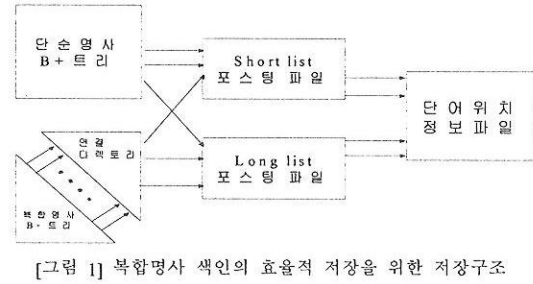
2. 점증색인과 복합명사를 고려한 저장구조

동적으로 유입되는 문서정보를 기존 역파일 저장 구조에 효율적으로 추가하는 방법이 근래에 제안되었는데 대표적인 것으로 미국 매사추세츠 대학에서 개발한 MNEME 라는 영속 객체 저장 구조를 사용하는 연구[1]와 스텐포드 대학에서의 연구[2]를 들 수 있다. 한국어 텍스트의 검색에 있어 고려되어야 할 다양한 언어 종속성[3] 중 가장 중요한 요소 중의 하나는 복합 명사의 처리에 대한 필요성이라고 할 수 있다. 한국어 텍스트에 있어 복합명사는 그 출현 빈도가 매우 높을 뿐만 아니라 복합명사의 띄어 쓰기가 자유롭기 때문에 색인어를 추출하는데 있어 복합명사를 분리하거나 단순명사 열을 복합명사로 합성하는 방법에 대해 많은 연구가 이루어져 왔다[4][5][6]. 한국어 정보 처리에서 질의어와 색인어의 불일치에 의한 검색효율의 저하를 방지하기 위해 복합명사와 명사구에 대한 색인어 생성에 대한 연구[7]도 있었으나 텍스트에서 색인어를 추출한 후 이를 효율적으로 저장하는 방법이나 검색에 있어서의 복합명사의 부분 정합에 관한 연구 결과는 거의 전무한 상태이다. 영어의 경우도 역시 구(phrase)의 추출이나 이를 이용한 검색 기법에 대한 연구[8]가 되어 왔으나 역시 이러한 기능을 지원하기 위한 저장 구조에 대해서는 아직 거의 알려져 있지 않다. 본 논문에서는 스텐포드 대학에서 제안한 이중구조(dual-structure)를 기반으로 하여 복합명사 혹은 영문의 경우 구(phrase) 검색을 효율적으로 지원하는 저장 구조 방식을 제안한다.

1 본 연구는 정보통신부 산학연 공동기술개발사업의 연구비 지원으로 수행되었음

2.1 저장구조

본 논문에서 제시하는 저장구조는 복합명사와 구성 명사를 색인어로 저장하고 각각에 따른 문서 리스트를 포스팅 파일에 저장함으로써 야기되는 저장 공간 사용의 비효율성을 제거하고 검색 효율을 동시에 향상시키는데 주안점을 두고 있다.



[그림 1] 복합명사 색인의 효율적 저장을 위한 저장구조

[그림 1]에서 볼 수 있는 것과 같이 복합명사를 위한 색인 저장 구조는 크게 4 부분으로 이루어져 있다.

- 검색 용어 사전으로 사용되고 포스팅 파일의 효율적인 접근을 허용하는 B+ 트리
- 복합명사의 특성을 고려한 새로운 구조의 포스팅 파일
- 복합명사와 구성명사간의 연결 정보를 알려주는 연결 디렉토리
- 용어의 문서내 발생 위치 정보를 저장하고 있는 단어위치 정보파일

1) B+ 트리

단순명사와 복합명사를 위한 B+ 트리가 각각 존재한다. 단순명사만을 위한 트리의 노드에는 용어, 이 용어 자체 혹은 이를 포함한 모든 용어를 고려한 문헌빈도수(DF), 이 용어 자체에 대한 문헌 빈도수, 포스팅 파일로의 포인터 등이 저장되어 있다. 복합명사만을 위한 트리의 노드에는 용어, 연결 정보파일의 포인터 등이 저장되어 있다. 단순명사는 단순명사 B+ 트리를 이용하여 포스팅 파일에 직접 접근하게 되고 복합명사는 B+ 트리를 통해서 연결 정보파일에 접근한 후 포스팅 파일로 접근하게 된다.

2) 연결 정보파일

연결 정보파일은 복합명사들간의 구성 관계를 표시하기 위한 정보 저장 파일이다. 즉, 단순명사가 모여서 복합명사를 구성하고, 또 그 복합명사들이 모여서 다른 복합명사를 구성하게 되는데 그 구성관계를 저장하기 위해 사용된다. 연결 정보파일에는 복합명사를 구성하는 복합명사들의 저장 위치와 복합명사들간의 유사도 정보는 물론 그

Copyright (C) 2005 NuriMedia Co., Ltd.

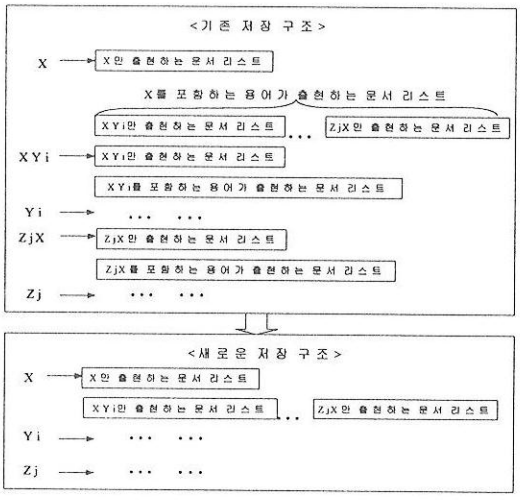
복합명사로 구성된 다른 복합명사들의 저장 위치와 유사도 정보를 저장하게 된다.

3) 포스팅 파일

점증 색인을 고려하여 포스팅 파일은 색인 정보의 크기에 따라 short list 포스팅 파일과 long list 포스팅 파일로 구분되어 저장된다. Short list 포스팅 파일에서 용어들은 일정한 크기의 버켓을 단위로 하여 여러 용어가 같이 관리되므로 long list 포스팅 파일로의 빈번한 이동을 줄일 수 있으며 추가되는 정보의 크기에 따라 long list 포스팅 파일로 이동도 하게 된다. Long list 포스팅 파일은 추가되는 정보를 저장하기 위한 공간을 기존에 색인된 정보의 크기에 비례적으로 할당하여 관리하게 된다.

4) 단어위치 정보파일

색인어가 발생한 문서내의 위치정보를 저장하기 위해 사용된다. 색인어에 대한 위치 정보는 질의 결과를 사용자에게 제시할 경우와 단어가 위치 정보를 검색에 이용할 경우에 사용되며 포스팅 파일과 별도로 관리되므로써 공간과 검색에 따른 효율을 기할 수 있다.



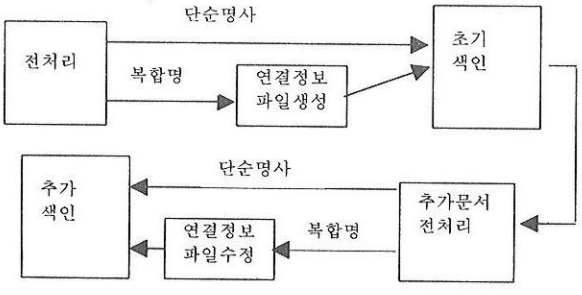
[그림 2] 저장구조의 비교

[그림 2]는 저장 공간을 줄이기 위해 설계된 포스팅 파일의 구조를 보여주고 있다. 이렇게 각 용어에 해당하는 포스팅 정보를 저장함으로써 각 용어에 해당되는 문서리스트를 별도로 저장하는 일반적인 경우보다 저장 공간상의 효율을 향상시킬 수 있음을 알 수 있다.

2.2 색인 정보의 저장

점증색인과 복합명사의 부분정합을 가능하게 하기 위해 수행되는 색인과정은 다음과 같이 세단계로 나누어 생각할 수 있다.

- 색인 대상 문서집합을 가지고 색인구조를 구축하는 과정에 앞서 색인정보를 필요한 정보 형태로 가공하는 색인정보 전처리 단계
- 역파일 구조를 처음으로 생성하는 단계
- 추가된 색인문서를 대상으로 기존에 생성된 색인 구조를 이용하여 추가된 정보를 전체적인 재색인 없이 추가하는 단계



[그림 3] 역파일 구축 과정

[그림 3]은 초기 색인 대상문서에서 역파일 구축을 한 후 동적으로 추가되는 문서를 대상으로 추가 색인을 실시하는 과정을 보여주고 있다.

1) 전처리

색인 문서를 대상으로 색인 정보를 추출한 후 추출된 용어를 대상으로 복합명사 분리작업[9]을 거쳐 색인 구조 생성에 적합한 구조를 생성하게 된다.

2) 초기 색인

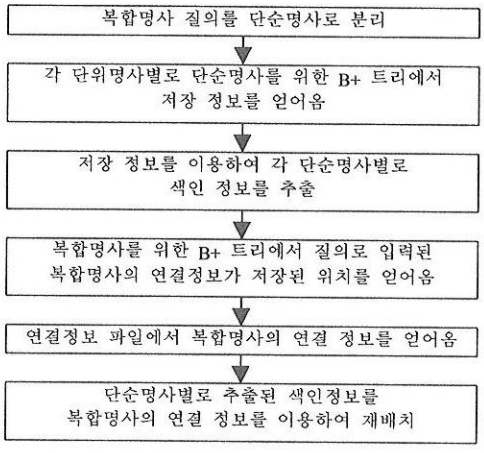
연결 정보 파일과 역파일 구조에 색인 정보를 저장하게 된다. 이때 단순 명사와 복합명사를 따로 저장하여 관리하게 되며 점증 색인을 위한 여유 공간도 색인 정보에 따라 비례적으로 할당하게 된다.

3) 추가 색인

기존에 구축된 색인 구조를 이용하여 동적으로 유입되는 문서를 대상으로 추가 색인을 실시한다. 전체적인 재색인없이 추가되는 정보만을 저장하게 된다.

2.3 색인 정보의 검색

검색과정은 질의어로 단순명사가 들어온 경우와 복합명사가 들어온 경우로 나누어 생각할 수 있다. 단순명사인 경우 B+ 트리를 통해 포스팅 파일에 접근하여 저장정보를 얻어온다. 복합명사인 경우는 부분정합을 고려하여 검색되므로 그 과정은 [그림 4]와 같이 수행된다.



[그림 4] 복합명사 검색 과정

3. 실험 및 평가

시스템 평가의 주안점은 기존 방법과 제안한 방법간의 역파일 생성에 따른 저장 공간의 크기를 비교하고, 색인어 검색에 따른 검색 시간을 비교하는데 있다. 여기에서는 동일한 실험 문서들의 집합을 대상으로 동일한 환경에서 역파일을 생성하고 검색을 실시하였다.

복합명사를 단위명사로만 색인하여 검색하는 시스템은 복합명사 질의를 처리하기 위하여 단어의 위치정보를 이용하여야 하므로 복합명사의 검색에 따른 오버헤드가 크다. 따라서 여기에서 비교 대상은 복합명사와 구성명사의 모든 조합을 중복 저장함으로써 부분정합을 가능하게 하는 시스템으로 한다.

실험 문서집합으로는 KT-SET[10] 과 KRIST[11]를 대상으로 하여 단어의 구성비를 기준으로 역파일 생성에 따른 저장공간의 크기와 복합명사 검색에 걸리는 시간을 측정하고 비교하기로 한다. 구현환경은 SUN Ultra Sparc 1E/200 이고, 운영체제는 Solaris 4.1.3 이며, 컴파일러는 C 이다.

[표 1] 실험대상 문서 분석

문서집합		KT SET	KRIST
내용		정보과학회논문	과학문서
건수		937 건	13515 건
크기		0.9Mb	26Mb
명사 조성비	단순명사	9163 개	138726 개
	복합명사	3986 개	90706 개

Copyright (C) 2005 NuriMedia Co., Ltd.

[표 1]은 실험 문서로 쓰인 KT-SET 과 KRIST 에 속한 문서들의 내용과 크기에 대한 비교이다. 표에서 나타난 명사 조성비중 단순명사는 실험 문서중 단순명사로만 쓰인 경우와 복합명사의 구성명사로 쓰인 용어의 수이고, 복합명사의 경우에는 두개 이상의 단순명사로 구성된 용어의 수이다. 명사 조성비에서도 알 수 있듯이 많은 문서에서 복합명사의 발생 빈도가 높음을 알 수 있다.

3.1 저장공간의 평가

색인 대상문서에 추출되는 색인어중 복합명사가 차지하는 비중이 커지면 동일 정보가 중복저장 되는 경우가 증가하고 그에 따른 저장 공간이 증가하게 된다.

[표 2] KT-SET 역파일 구축에 따른 공간 오버헤드

파일	기존 방법	새로운 방법	감소율
B+ 트리	3.9 Mb	3.4 Mb	10 %
포스팅 파일	0.8 Mb	0.7 Mb	
단어 위치정보 파일	0.7 Mb	0.6 Mb	
연결정보 파일	0 Mb	0.2 Mb	
총계	5.4 Mb	4.9 Mb	

[표 3] KRIST 역파일 구축에 따른 공간 오버헤드

파일	기존 방법	새로운 방법	감소율
B+ 트리	67 Mb	56 Mb	12 %
포스팅 파일	19 Mb	17 Mb	
단어 위치정보 파일	15 Mb	12 Mb	
연결정보 파일	0 Mb	4 Mb	
총계	101 Mb	89 Mb	

색인어 추출시 동일한 복합명사의 분리방법을 기존 방법과 제안한 방법에 적용하였다. [표 2]는 KT SET 을 대상으로 색인정보 구축에 따른 저장 공간의 크기 비교 결과를 보여주고 [표 3]은 KRIST 를 대상으로 색인 정보 구축에 따른 저장 공간의 크기 비교를 보여준다. 위의 표들을 보면 실제적으로 많은 공간 차이를 보이지 못하고 단지 10% 정보의 공간만 감소시킬 수 있었는데 그 이유는 다음과 같다.

- 한국어의 띄어쓰기를 고려하지 않아 실제로 추출된 복합명사의 수가 정확하지 못하다.(띄어쓰기를 고려하면 복합명사의 수 증가)
- 실험을 위해 추출된 복합명사중 구성명사의 수가 2 인 복합명사의 수가 많아 공간적 효과를 별로 보지 못하였다.
- 부분정합을 허용하기 위하여 연결정보 파일이라고 하는 부가적인 저장공간을 할당하였다.
- 같은 크기의 노드로 구성된 B+ 트리를 복합명사와 단순명사에 사용하였다. (복합명사의 노드 정보는 단순명사의 노드 정보에 비해 현저히 적음)

[표 2]와 [표 3]을 보면 역파일 생성에서 B+ 트리가 차지하는 크기가 매우 큰 것을 알 수 있다. 본 실험에서 사용된 B+ 트리는 Marcus J. Ranum 에 의해 개발된 것으로 가변 길이의 레코드를 처리하고 캐쉬 기능을 사용할 수 있으나 B+트리의 구조 유지에 상당한 부가 정보를 필요로 하고 있다. 또한 색인어를 저장하기 위한 노드의 크기를 상당한 크기로 일률적으로 할당, 사용하여 B+ 트리의 크기가 커지게 되었는데 이를 수정하기 위해서는 B+ 트리의 구조적 분석을 통한 재설계를 통하여 크기를 줄이며 효율성을 증대하기 위한 작업이 필요하다.

3.2 검색시간의 평가

복합명사의 검색에 따른 시간을 비교하기 위해 복합명사와 구성명사를 중복 저장하는 방법과 복합명사 검색에 따른 검색 시간을 측정하였다. 색인어중 구성명사가 2, 3, 4 개인 복합명사 각 5000 개를 임의로 추출하여 검색을 실시하였다. 검색어로 사용된 복합명사는 부분정합을 고려하여 검색을 실시하였고 그에 소요되는 시간을 기준으로 두 방법을 비교하였다.

[표 4] 검색시간의 비교

복합명사	기존 방법	새로운 방법	감소
2 구성명사	15 초	13 초	13 %
3 구성명사	32 초	21 초	38 %
4 구성명사	69 초	30 초	56 %

[표 4]는 구성명사 2, 3, 4 개로 구성된 복합명사를 가지고 검색을 실시한 결과로, 기존의 방법보다 복합명사의 구성명사수가 많아지면서 시간의 차이가 많이 났다. 4 개의 구성명사로 이루어진 복합명사의 검색시간은 기존의 방법보다 50% 가 넘는 차이를 보이는데 그 이유는 다음과 같다.

- 복합명사의 구성명사 수가 많아지면 부분정합을 가능하게 하기위해 B+ 트리에서 검색될 색인어의 수가 증가한다.
- 복합명사의 구성명사의 수에 따라 색인 정보 추출을 위한 포스팅 파일의 디스크 액세스가 증가한다.
- 추출된 정보의 관리를 위한 메모리의 관리에 따른 시간이 소요된다.

4. 결론 및 향후연구

텍스트 정보를 대상으로 문서를 검색하는 정보검색시스템의 성능을 평가하는 일반적인 지표로 검색결과에 대한 신뢰도(effectiveness)와 검색 속도(efficiency)를 들 수 있다. 한국어 텍스트 검색에 있어서 신뢰도 및 검색속도의 향상을 위해서는 한국어의 특성을 고려하는 것이 중요하다는 것은 이미 잘 알려진 사실이고, 이에 대한 연구가 색인 과정의 언어처리 관점에서 이루어져 왔다.

본 연구에서는 저장 구조에 초점을 맞추어 한국어 텍스트 검색에 있어서의 속도와 신뢰도를 향상시킬 수 방안을 제시하였다. 구체적으로 살펴보면, 복합명사 처리를 효율적으로 지원해 주는 색인 저장 구조를 설계함으로써 저장 공간의 효율성과 검색속도를 동시에 향상시켰고, 뿐만 아니라 부분정합(partial matching)을 효과적으로 수행할 수 있게 함으로써 검색 순위를 체계적으로 결정하여 검색의 신뢰도 향상도 가능하게 하였다. 또한 새롭게 제안하는 이 저장 구조는 점증색인을 가능하게 하는 기능도 통합 발전시켜, 지속적으로 추가되는 문서를 보다 효과적으로 처리할 수 있다.

비록 이 구조가 한국어 텍스트의 특수성에 적합하게 설계되었지만, 타 언어를 사용한 문서 검색에도 활용될 수 있다. 예를 들어 영어 문서의 경우 구(phrase) 검색에 대한 요구가 있는데, 본 연구에서 제안한 저장 구조가 이 목적으로도 사용될 수 있다. 본 논문에서 제안된 저장구조는 현재 정보검색 시스템 전용 색인 구조로 파일의 구조를 가정하고 있으나 향후 DBMS 와 정보검색 시스템의 밀결합을 위해 일반화되고 모듈화된 형태로 발전되어야 한다.

참고문헌

[1] E. Brown 외 2 인 , "Fast Incremental Indexing for Full-Text Information Retrieval", Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.

[2] K. Shoens 외 2 인, "Synthetic Workload Performance Analysis of Incremental Updates", Proceedings of SIGIR '94, Dublin, Ireland, 1994.

[3] 맹성현 외 1 인, Proceedings "On Language Dependency in Indexing", of the International Workshop on Information Retrieval with Oriental Languages, Taejon, 1996.

[4] 이창열 외 3 명, "자동 키워드 제작기 시스템 설계", 제 5 회 한글 및 한국어 정보처리 학술발표 논문집, 1993.

[5] 박영찬 외 1 인, "통계적 정보를 이용한 복합명사 검색모델", 인지과학, 6 (3), 1995.

[6] 윤보현 외 2 인, "통계정보와 선호 규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지(B), 24 (8), 1997.

[7] 윤보현 외 2 인, "A Korean Information Retrieval Model Alleviating Syntactic Term Mismatches", Proceedings of NLPRS' 97, Phuket, Thailand, 1997.

[8] J. Fagan, "The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval", J. of American Society for Information Science, 1989.

[9] 장동현 외 1, "효율적인 색인어 추출을 위한 복합명사 분석방법", 제 8 회 한글 및 한국어 정보처리 학술대회, 1996.

[10] 맹성현, "대용량 통신처리에서의 다자간회의를 위한 멀티캐스팅 연구", 최종 연구보고서, 한국전자통신연구소, 1995.

[11] 김성혁 외, "자동색인기 성능시험을 위한 Test Set 개발", 정보관리학회지 제 11 권 1 호, 929-932, 1994.

[12] 이준호 외 4, "정보 검색 연구를 위한 KRIST 테스트 컬렉션의 개발", 정보관리학회지 제 12 권 2 호, 1991.

Copyright (C) 2005 NuriMedia Co., Ltd.