시소러스 및 요약화일을 이용한 문서 검색시스템

정상철*, 신동욱
충남대학교 컴퓨터 공학과

The development of a document retrieval system using thesaurus and signature file

Sangcheol Jeong, Dongwook Shin
Department of Computer Engineering, Chungnam University

요 약

본 논문에서는 요약화일을 이용하여 복합명사를 효율적으로 처리하며 시소러스를 이용하여 검색하는 한글문서 검색시스템을 제안한다. 본 한글문서 검색 시스템은 한글문서를 대상으로 색인하는 자동색인기와 사용자의 질의를 받아 관련된 문서를 검색하는 검색기로 구성된다. 자동색인기는 우선 한글문서를 대상으로 최장일치 방법으로 명사들을 추출한 후 복합명사의 패턴을 분석하여 복합명사의 가능성이 높은 것들을 복합명사화한다. 두번째로 이들 복합명사들을 1+2SP방식으로 코딩한 후 요약화일 방법을 이용하여 요약화일을 작성한다. 검색기는 사용자 질의어를 받아 명사들을 추출한 후 시소러스를 이용하여 질의어를 확장한다. 다음 확장된 질의어를 1+2SP 방식으로 코딩한 후 관련된 문서를 검색한다. 본 논문에서는 한국통신에서 만든 코퍼스를 이용하여 제안된 방법의 성능을 평가하였는데 복합명사 처리 및 시소러스 이용방식이 효율적임이 입증되었다. 또한 KAIST에서 개발한 문서검색 시스템보다 동일한 코퍼스로 실험하였을 경우 재현률 및 정확률이 7~8% 정도 앞서 기존의 시스템보다도 성능이 우수하다는 것이 밝혀졌다.

1. 서론

정보검색이란 수집된 정보 또는 정보 자료의 내용을 분석한 뒤 적절히 가공하여 축적해 놓고 이용자의 정보요구에 적합한 정보를 탐색하여 찾아내는 일련의 과정을 말한다. 이와 같이 정보검색을 축적과 검색이라는 주요한 두가지 기능으로 구성되어 있으며 이 두 기능이 서로 조화를 이룸으로써 비로서 본래의 기능을 발휘하게 된다[12]. 따라서 정확히 표현하자면 정보의 축적과 검색이라고 해야 하지만 검색은 축적을 전제조건으로 한 것이므로 일반적으로 정보 검색 (Information Retrieval)으로 불리우고 있으며 약자로는 IR이라고 쓰인다.

색인은 정보검색의 한 분야로써 컴퓨터 데이타베이스에 저장되어 있는 수많은 정보에서 관련된 정보를 탐색하기 위하여 문헌을 분석한 후, 주요어를 추출하여 정보와 연관시키는 작업을 말한다. 이 색인과정에서 주요어를 추출하는 기존의 방식은 대개 색인자(indexer)가 정보 자료의 내용을 분석하여 중요한 개념을 적절한 색인어로 표현해 주는 것으로 색인자가 자신의 전문 지식에 기초하여 임의로 색인어를 부여하거나 시소러스(thesaurus)를 참조하여 미리 통제된 용어들 가운데서 가장 적당한 색인어를 선택한다. 그러나 1960년 이후에 사람이 관리할 수 없을 정도로 정보량이 급격히 증가하고 정보의 세분화 및 복합적 이용, 그리고 사람이 색인

을 하는 경우 선택하는 색인자가 종종 불일치 하다는 등의 문제점이 부각됨에 따라 컴퓨터에 의한 자동색인(automatic indexing)이 출현하게 되었다.

자동 색인 시스템은 수작업색인이 안고있는 문제점인 색인어 선택시 일관성이 결여되는 문제를 해결할 수 있고 정보를 대량으로 신속하게 처리할 수 있다. 지금까지 한글문서에 대하여 여러가지 자동색인 방법이 제시되어 왔는데 대표적인것이 장재우[6]와 이창렬 방식[8]등이다. 장재우는 명사들을 1+2SP 방식으로 코딩한 후 요약화일을 작성하여 복합명사 처리문제를 어느정도 해결하였다. 이창열은 복합명사 패턴을 정의하고 추출된 명사들을 복합명사 패턴에 따라 키워드를 자동적으로 생성하였다. 그러나 이 연구들을 포함하여 대부분의 연구가 복합명사를 잘 처리하지 못하고 있다.

복합명사 처리는 띄어쓰기가 되어있는 경우와 되어있지 않은 경우를 동일하게 인식할 수 있어야 인식율을 높일 수 있는데, 대부분의 경우는 이들을 처리하지 못하거나 일부분밖에 처리하지 못한다. 장재우 방식은 띄어쓰기 문제는 일부분 해결하였으나 문서에는 복합명사가 띄어쓰기로 되어있으나 사용자가 붙여서 질의한 경우에는 아래의 예에서 보듯이 이 문서를 추출하지 못한다.

문서1 : 정보검색, 시스템 -->
2SP : 정보,보검,검색,시스,스템

질의어 : 정보검색시스템 -->
2SP : 정보,보검,검색,색시,시스,스템

그 이유는 문서 1과 질의어의 2SP 방법을 비교해 볼 때 문서 1에서 "색시"에 대한 요약 부분이 없으므로 질의어에 대해서 이 문서가 검색되지 않기 때문이다.

본 논문에서는 이 문제를 해결하기 위해서 한글문서에서 명사 추출시 복합명사 생성규칙을 이용하여 복합명사를 생성하는데 문서에서 명사와 명사가 띄어져 있는 경우에는 이를 붙여 복합명사로 처리하여 색인한다. 보통 복합명사는 명사+명사, 명사+'의'+명사, 명사+명사+명사의 패턴이 90% 정도를 차지한다고 통계적으로 나와 있는데 이 세가지 패턴을 대상으로 하더라도 대부분의 복합명사를 처리할 수 있다.

또한 본 논문에서는 검색하는 과정에서 재현률을 높이기 위하여 시소러스를 이용하여 질의어를 확장시킨다.이때 질의어의 용어들을 시소러스에 일치시킨후 깊이 1정도의 자식노드들을 ORing하여 질의어를 확장시키고,만약 시소러스에 일치된 용어가 단말노드라면 형제노드들을 ORing하여 질의어를 확장시킨다.

본 논문에서는 한국통신에서 만든 코퍼스를 이용하여 제안된 방법의 성능을 평가하였는데 복합명사 처리 및 시소러스 이용방식이 효율적임이 입증되었다. 또한 KAIST에서 개발한 문서검색 시스템보다 동일한 코퍼스로 실험하였을 경우 재현률 및 정확률이 7~8% 정도 앞서 기존의 시스템보다도 성능이 우수하다는 것이 밝혀졌다.

본 논문은 다음과 같이 구성되어 있다. 제 2절에서는 관련연구로 본 논문과 관련된 연구사례를 소개하고 각각의 특징을 분석한다. 제 3절에서는 본 색인 시스템의 구성과 색인과정 및 검색과정을 설명한다. 마지막으로 질의어 검색과정을 설명한다. 제 4절에서는 한국통신(KT)에서 개발한 데이타 집합을 가지고 성능을 평가하였다. 성능평가를 위해 27개의 질의어를 이용하여 명사를 추출한 색인, 복합명사를 추출한 색인, 그리고 복합명사를 추출한 색인을 이용하여 시소러스를 사용한 질의어 확장시의 검색결과를 각각 재현률과 정확률의 측면에서 비교, 분석한다. 마지막으로 제 5절에서는 결론과 앞으로의 연구방향에 대하여 논의한다.

2. 관련연구

장재우[6]는 복합명사 처리문제를 어느정도 해결하기 위하여 요약화일 기법을 설계하였는데 이때 한글 텍스트를 코딩하는 방법을 세가지 제시하였다.

(1) 한 음절을 해쉉 단위로 코딩하는 방법
(1 Syllable Pattern : 1SP)
(2) 두 음절을 해쉉 단위로 코딩하는 방법
(2 Syllable Pattern : 2SP)
(3) 한 음절과 두음절을 혼합하여 코딩하는 방법
(1+2 Syllable Pattern : 1+2SP)

가장 간단한 1SP 방법의 장점은 사용자에 의해 필요로 하는 모든 부분 매치 질의가 지원될 수 있다는 점이다. 그러나 한음절에 의한 코딩 때문에 한 어절을 검색하는데 높은 false match를 야기시키는 단점이 있다. 반면 2SP 방법은 평균 어절의 길이가 2.75 음절이라는 점을 감안하여 설계한 것으로 찾고자 하는 어절이 2 이상일 때는 매우 효과적인 방법이다. 그러나 한글에는 한 음절이 하나의 어절을 구성하는 경우가 많으므로(예를 들면, 소, 말, 종, 길, 돌, 들, 산, 꽁, 배, 총, 손, 발, 램, 롬 등) 이를 검색할 수 없는 단점이 있다.

(추출 예)
한글문서 : 정보검색시스템
2SP : (정보,보검,검색,색시,시스,스템)

질의어 : (정보검색 AND 시스템)
2SP : (정보,보검,검색) AND (시스,스템)

한편 두 개의 방법을 결합한 세번째 방법을 사용하면 위의 두 방법의 장점을 모두 지닐 수 있기 때문에 효율적인 코딩 방법을 설계할 수 있다. 이 방법은 두 방법의 요약을 저장했기 때문에 부가 저장 공간이 다소 증가하지만 다음과 같은 장점이 있다.

첫째, 1음절어를 포함하는 부분매치(partial match)를 처리 할 수 있다. 둘째, 띄어쓰기 형태에 무관하게 질의에 해당하는 문서를 검색할 수 있다. 세째, 질의에 포함되는 색인어의 수가 적을 경우에는 탈락 착오가 낮다.

그러나 두번째 장점은 질의어가 띄어쓰기 형태에 무관하게 해당하는 문서를 검색하여 복합명사 문제의 일부분은 해결하지만 반대로 아래의 예와 같이 질의어가 복합명사의 형태로 나타나고 문서에서 추출한 색인어가 띄어쓰기 형태로 나올 경우는 검색되지 않는다.

문서 : 정보검색, 시스템 -->
2SP : 정보,보검,검색,시스,스템

질의어 : 정보검색시스템 -->
2SP : 정보,보검,검색,색시,시스,스템

문서와 질의어의 2SP 방법을 비교해 볼 때 문서에서 "색시"에 대한 요약 부분이 없으므로 질의어에 대한 해당 문서가 검색되지 않는다.

복합명사 처리문제를 해결하기 위해서 한글문서에서 명사 추출시 이창렬[7]씨의 복합명사 패턴에 의하여 색인어인 명사들을 복합명사의 형태로 추출한다. 복합명사의 패턴중 명사+명사, 명사+'의'+명사, 명사+명사+명사의 패턴이 90% 정도를 차지한다고 통계적으로 나와 있으므로 이 세가지 패턴을 대상으로 하더라도 대부분의 복합명사를 처리할 수 있다.

복합명사 처리는 정보검색에 있어서 정확률을 높이기 위해서도 필요하며 한글에서의 명사의 띄어쓰기 문제도 해결이 가능하다.

KAIST에서 개발한 한국어 정보검색 시스템[10]은 한국어 문서를 대상으로 색인하는 색인기와 사용자의 질의를 받아 관련된 문서를 보여주는 검색기로 구성된다. 색인기는 한국어 자연언어 문장을 형태소 해석하는 형태소해석기와 추출된 색인어에 가중치 할당 및 불용어를 제거하는 색인기, 문서검색시 사용하는 색인구조 생성기로 나누어진다. 검색기는 사용자의 질의를 해석하여 해당 문서를 가져오는 질의 해석기와 검색된 문서를 질의와의 관련성을 계산하여 순위를 매기는 문서랭커로 구성된다. 이 시스템은 불리안 검색외에 키워드가 문서의 어느 항목에 포함되어 있는지를 제한하는 조건검색 및 키워드간의 위치에 제한을 두는 인접검색 기능을 제공한다. 문서에서 추출된 색인어의 가중치를 구하기 위해서는 그 색인어가 문서내에서 갖는 표현력과 다른 문서와의 구별을 나타내는 분별력을 고려 하였다. 표현력 정보로는 문서내 빈도정보와 자연언어처리기법을 사용한 가중치가 있으며 분별력 정보로는 역문서빈도 정보가 있다. 질의와 문서와의 관련도 계산은 Extended Fuzzy Boolean Model을 이용하였으며 사용자는 관련도가 높은 문서부터 볼 수 있으므로 검색효율이 좋아진다.

3. 한글 자동 색인 시스템의 설계

3.1 한글문서 검색시스템의 구성

본 절에서는 한글문서 검색시스템에 관한 전체적인 구성요소 및 구조를 설명하겠다.

한글문서 검색시스템의 전체구조를 보면 그림 1과 같다.

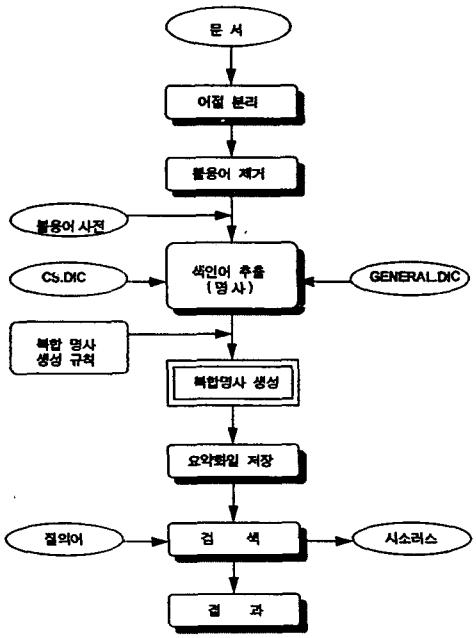

그림 1. 한글문서 검색시스템의 전체구조

본 시스템은 전산분야의 문서를 대상으로 구축되었는데 우선 이들 문서를 입력받아 어절단위로 분리한 후 불용어 사전을 이용하여 불용어를 처리한다. 두번째로 불용어가 제거된 어절들을 명사 사전에 비교하여 최장일치법을 사용하여 색인어인 명사를 추출한다. 이때에는 두 종류의 명사사전을 가지고 색인어를 추출한다. 전산분야에 관한 명사사전과 일단계로 비교하여 전산 분야에 관한 색인어를 추출하고 나머지 어절들을 일반명사사전과 비교하여 일반적인 색인어를 추출한다. 이렇게 두단계로 명사를 추출하는 이유는 정보검색시 일단 사용자가 전산분야의 문서이므로 주로 전산용어에 대한 질의를 할것이므로 정확률(precision)을 좀더 높힐 수 있다.

세번째로 복합명사의 생성규칙을 이용하여 복합명사를 생성하는데 이로인해 한국어의 문제점 중의 하나인 명사들의 띄어쓰기 문제를 해결 할 수 있다.

네번째로 색인어로 추출된 정보를 유지하기 위해 요약화일 구조로 저장한다. 정보검색 시스템에서 많이 사용되고 있는 문서 검색 기법으로는 우수한 검색 성능을 지닌 역화일 구조가 많이 사용되고 있지만, 역화일 기법은 데이타 화일의 50 ~ 300% 정도의 많은 부가 공간을 필요로 하는 단점이 있으며 반면에 요약화일은 부가 저장 공간의 크기가 데이타 화일의 20% 내외로 대규모의 텍스트 처리에 용이하나 검색 속도는 뒤떨어 진다. 그러나 요약화일을 이용할 경우 한글의 문제점 중의 하나인 복합명사 처리를 일부 해결할 수 있고 하드웨어의 성능이 나날이 향상되므로 본 논문에서는 정보 저장 하부구조로써 요약화일을 사용하였다.

다섯번째는 검색과정으로 사용자 질의가 들어오면 위에서 만든 요약화일을 탐색하여 적합한 문서들을 검색한다. 이때 사용자 질의로부터 명사들을 추출하고 각 명사들에 대하여 시소러스 용어들과 부합시켜 부합되는 부용어(subterm)들로 확장시킨다. 본 연구에서 사용된 시소러스는 CRCS를 한국통신 연구소에서 한글로 번역한 것으로 기본적으로 협의어와 광의어들의 관계가 표시되어 있다. 질의어에 있는 명사를 확장하는 방법은 만약 시소러스에서 이 명사와 일치되는 명사가 있으면 이 명사와 협의어 관계에 있는 명사들을 전부 ORing시켜서 확장시킨다. 이때 일치되는 바로 밑의용어들만 확장시키고 그 이하의 용어들은 확장시키지 않는데 그 이유는 모든 협의어들을 전부 확장한 경우에는 확장되는 용어의 수가 너무 많고 또 경우에 따라서는 너무 자세한 용어들까지도 확장되기 때문에 정확률의 현저한 저하를 초래하기 때문이다.

3.2 불용어 처리

불용어 처리란 색인의 역할을 할 수 없는 어휘들을 리스트로 가지고 있고 이들을 초기 단계에서 삭제시킴으로써 불필요한 시간을 절약하는 것이다.

러스버겐[3]에 의하면 불용어를 제거하여 중요하지 않는 단어들을 제거함으로써 검색시스템 전체 문헌파일의 크기를 30 ~ 50%까지 줄일 수 있다고 한다. 그러나 불용어를 너무 과다하게 선정할 경우에는 색인어로 선정되어야 할 단어가 제외되고 너무 적게 선정할 경우에는 불필요한 단어가 색인어로 선정되므로 불용어 선정에 신중을 기해야 한다.

본 논문에서는 입력된 한글 문서를 공백을 중심으로 어절을 분리하는 어절분리 단계를 거치고 불용어를 제거한다. 불용어 리스트를 작성하는 과정은 Fox[1]가 제시한 바와 같이 다음과 같은 기본적인 과정을 거친다.

1. 대량의 문서들로 부터 어휘들을 추출한다.
2. 자주 나타나는 어휘들을 선택한다.
3. 고빈도의 단어들 중 주요 어휘를 제거한다.
4. 중, 저빈도의 단어중 불필요한 어휘를 첨가한다.

이와 같은 과정을 거친후 생성된 불용어들로 불용어 리스트를 만든 후 문서에 일치시켜 불용어를 제거한다.

본 논문에서는 경제기획원에서 만든 220여개의 불용어[11]와 국내문헌에 나타나 있는 불용어를 중심으로 불용어 리스트를 작성하여 약 400여개의 불용어를 이용하였다. 또한 어미 변화에 따른 불용어 수의 증가를 효율적으로 처리하기 위해 우측절단 (right truncation) 방법을 많이 사용하고 있다. 즉 "전개"라는 단어가 불용어가 될 경우 "전개*"로 표시해 주면 "전개되는", "전개하는" 등의 단어들이 불용어로 간주되는 것을 말한다.

불용어 리스트의 예)

안고, 어려운, 해왔*, 왔*,.....

cf) * 표시는 뒤에 어떤 어절이 오더라도 불용어의 범위에 포함시킨다는 우측절단(right trancation)표시

3.3 색인어 추출

자동색인이란 인간이 색인어를 추출하는 대신 컴퓨터가 자동으로 추출하는 것을 말하며 색인결과는 검색시의 검색효율, 즉 정확률과 회상률을 좀 더 높히는데 목표를 두고있다. 본 논문에서는 색인어로서 명사만을 추출한다. 명사만을 추출하는 이유는 한국어의 특성상 의미있는 단어는 주로 명사를 가지고 표현하기 때문이다. 그리고 사용자가 본 논문에서와 같이 질의어로 불리안식을 사용한다면 거의 명사를 가지고 검색을 하기 때문에 색인어로 명사를 추출하는 것이 더욱 적당할 것이다.

본 논문에서는 효율적인 색인어의 추출을 위해 두 종류의 사전을 사용하는데 약 3000여 단어로 된 전산분야의 명사사전과 40000여 단어로 구성된 일반 명사사전이 있다. 사전을 두 종류로 나누는 이유는 본 논문의 실험 데이타가 한국통신에서 개발한 정보과학에 대한 1000개의 논문의 요약부분이므로 일반 명사사전에서 보다는 좀 더 정확한 색인어의 추출을 위해 전산분야의 명사사전이 필요하며 사용자가 질의어를 사용할 시에도 주로 전산분야의 색인어들을 사용할 것이기 때문이다. 또한 이 두종류의 사전들은 단어수가 많고 앞으로도 계속 확장을 할 것이므로 탐색시 많은 시간을 필요한다. 따라서 사전 데이타를 좀 더 빠르게 탐색해야 할 필요성이 있다. 이를 위해 이들 사전들을 트라이 구조[8]로 구성하였다.

보통 색인어를 어휘사전을 이용하여 비교추출하는 방법으로는 최장일치법과 Head_Tail법 등이 있는데 본 논문에서는 최장일치법이 조사사전이 필요없고 간단하므로 최장일치법을 사용하였다.

3.4 복합명사 추출

본 논문의 자동색인에서는 한글처리의 가장 큰 문제점중의 하나인 명사들의 띄어쓰기 문제를 해결하고 정확도를 높이기 위해 복합명사를 색인시 추출한다.

한글문서 검색 시스템의 문서 저장 하부구조는 요약화일로 구성된다. 이 때 요약추출 기법으로는 1음절과 2음절의 혼합(1+2 Syllable Pattern)의 코딩을 사용한다[8].

1+2SP 방법을 사용하면 위의 1음절의 코딩과 2음절어의 코딩의 장점을 모두 지닐 수 있기 때문에 효율적인 코딩방법을 설계할 수 있다.

복합명사 처리문제를 해결하기 위해서 한글문서에서 명사 추출시 복합명사 생성규칙에 의하여 색인어인 명사들을 복합명사의 형태로 추출한다. 복합명사 처리는 정보검색에 있어서 정확률을 높이기 위해서도 필요하며 한글에서의 명사의 띄어쓰기 문제도 해결이 가능하다.

한글에서의 복합 명사의 패턴은 다음과 같이 분류될 수 있다[9].

명사 2개(2개): 명사+명사(타입 1)
명사+'의'+명사(타입 2)

명사 3개(4개): 명사+명사+명사(타입 3)
명사+'의'+명사+명사(타입 4)
명사+'명사+'의'+명사(타입 5)
명사+'의'+명사+'의'+명사(타입 6)

명사 4개(8개):
명사+명사+명사+명사(타입 7)
명사+'의'+명사+명사+명사(타입 8)
.....
.....

명사 5개(5개):
명사+명사+명사+명사(타입 15)
.....
.....
명사+명사+명사+명사+'의'+명사(타입 19)
기타(타입 20)

복합명사의 패턴 통계 분석을 해보면 타입 1이 전체 복합명사 후보에서 58.4%, 타입 2가 22.6%, 그리고 타입 3이 7.0%를 차지하므로 이 세 타입을 합한 비율은 거의 90%가 된다. 따라서 이 세가지 패턴을 대상으로 하더라도 대부분의 복합명사를 처리할 수 있다.

본 논문에서는 이 세타입의 복합명사 패턴을 채택하여 한글문서 자동색인시 복합명사를 생성하여 한글의 문제점 중의 하나인 명사들의 띄어쓰기 문제를 해결하였다.

복합명사 생성규칙의 예)

1. 명사 + 명사
예)정보 검색을 --> 정보검색
2. 명사 + '의' + 명사
예)정보의 검색 --> 정보검색
3. 명사 + 명사 + 명사
예)정보 검색 시스템이--> 정보검색시스템

3.5 요약화일의 검색

역화일과 요약화일의 비교시 요약화일의 단점중의 하나인 검색속도가 다소 떨어진다는

단점을 다소 보완하기 위하여 본 논문은 검색시 Bit-Sliced Signature Files(BSSF)방법을 사용하고자 한다.

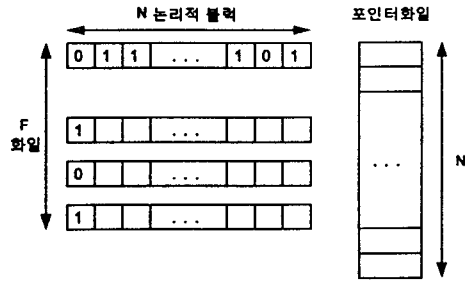

그림 2. Bit-Sliced 요약화일 구조

그림 2는 BSSF 방법[2]을 나타낸 것이다. 요약화일은 한 단일어를 탐색하기 위해 이 단일어를 요약화일 형태로 바꾼 후 검색시 F(요약화일의 bit 수) X N(논리적 블록수)번의 비트 비교를 해야함으로 검색속도가 매우 느리다. 반면에 BSSF방법은 N개의 논리적 블록들을 각각 첫 비트부터 검색하여 질의어와 일치되는 블록들은 계속 검색하고 그렇지 않은 블록들은 검색을 중지하므로 검색속도의 향상을 기대할 수 있다.

3.6 시소러스를 이용한 질의어 확장

본 논문에서 이용된 시소러스는 CACM에서 출판한 출판물을 색인하기 위해서 사용되는 CRCS(Computing Reviews Classification Structure)[4]를 한국어로 번역한 것으로 계층적 구조로 구성되어 있으며 1000여개의 노드들을 가지고 있는데 부모노드와 자식노드의 관계는 광의어와 협의어의 관계이다.

시소러스의 용어들은 구 중심으로 되어 있는 반면에 질의어는 단일명사와 복합명사로 되어 있어 이들을 일치시키기 어려우므로 시소러스를 단일명사와 복합명사 단위로 가공하였다.

예)

1)B.0 일반 --> 삭제

2)B.1.1 제어설계 유형
--> B.1.1 제어설계

3)B.1.2제어구조 및 마이크로프로그래밍
--> B.1.2 제어구조
B.1.2 마이크로프로그래밍

4)B.1.3 입/출력 형식
--> B.1.3 입력형식
B.1.3 출력형식

시소러스에 있는 용어들 중 예 1)과 같이 '일반'은 아무 의미가 없으므로 삭제하고, 2)와 같이 간단한 구구조로 되어 있으면 의미가 없는 단어를 제거하여 단일명사나 복합명사 형태로 가공하고 3)이나 4)와 같이 ' 및'이나 '/'가 중간에 있는 구조는 양쪽에 있는 용어들이 대등한 관계이므로 두개의 시소러스 용어로 분리하였다.

질의어에 있는 명사를 확장하는 방법은 만약 시소러스에서 이 명사와 일치되는 명사가 있으면 이 명사와 협의어 관계에 있는 명사들을 ORing시켜서 확장시킨다. 이때 일치되는 바로 밑의 용어들만 확장시키고 그 이하의 용어들은 확장시키지 않는데 그 이유는 모든 협의어들을 전부 확장한 경우에는 확장되는 용어의 수가 너무 많고 또 경우에 따라서는 너무 자세한 용어들까지도 확장되기 때문이다. 또한 만약에 시소러스에 일치되는 명사가 단말노드라면 확장할 협의어가 없으므로 그 명사의 형제노드들을 ORing하여 질의어를 확장시킨다.

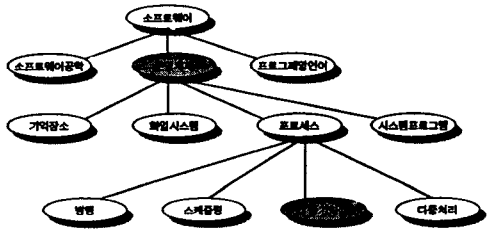

그림 3. 시소러스를 이용한 질의어 확장

예를들어 질의어가 (운영체제 AND 동기화)라면 그림 3과 같이 시소러스에서 질의어의 용어와 일치되는 용어를 찾는다. '운영체제'의 바로 밑의 협의어들은 '기억장소', '화일시스템', '프로세스', '시스템프로그램'이므로 이들을 ORing한다. 다음 질의어 용어인 '동기화'는 바로 밑의 협의어가 없으므로 형제 노드인 '병행', '스케쥴링', '다중처리'등을 ORing한다. 그 결과 질의어의 확장은 ((운영체제 OR 기억장소 OR 화일시스템 OR 프로세스 OR 시스템프로그램) AND (동기화 OR 병행 OR 스케쥴링 OR 다중처리))와 같이 된다.

4. 실험 및 분석

4.1 실험 데이타

실험 데이타 집합은 한국통신(KT)에서 개발한 1000개의 정보과학회 논문 요약부분, 27개의 불리언 질의어, 시소러스, 그리고 관련정보로 구성되어 있다. 1000개의 문서는 표 1에서 보는 바와 같이 한글과 영문으로 구성되어 있는데 한글색인이 본 논문의 요지이므로 문서번호(id), 한글제목(title), 한글 요약부분을 추출하여 자동색인을 하였다.

<id>0908
<title>파일조직 방법에 따른 검색성능의 비교연구
<author>송미련
<affiliation>
<language>한국어
<journal>정보관리학회지
<issn>1013-0799
<year>1986
<volume>3
<number>1
<pages>17-39
<abstract> 정보검색시스템에 대한 관심이 날로 증대하고 또 온라인 정보 검색 시스템의 발달로 이용자는 더욱 효과적이고 빠른 탐색을 기대하게 되었다. 여기서 중대한 문제의 하나가 파일조직방법의 선택이다. 본 논문에서는 파일조직방법이 검색성능에 영향을 미칠 것이라는 가설하에 여러가지 파일조직방법 중 도치파일과 클러스터파일을 선택하여 그에 따른 성능을 비교하였다.
<etitle>An Experimental Study on the Retrieval Performance of File Organization Methods
<eauthor>
<eabstract>In this thesis, inverted file and clustered file were compared in terms of their retrieval performance under the hypothesis that file organization method will influence the retrieval performance.
<classification>H.3.2.1
X.5.11
<keywords>파일조직 방법
검색성능
온라인 정보검색시스템
빠른 탐색
도치파일
클러스터파일
<notes>

표 1. 문서 견본

4.2 분석

본 자동검색 시스템은 SUN Workstation에서 C언어를 사용하여 구현하였다. 색인과정은 문서들을 우선 어절단위로 분리한 후 불용어 리스트를 이용하여 불용어를 제거한다. 불용어를 제거한 문서를 명사사전에 비교하여 색인어로써 명사를 추출한다. 그 다음 단계로 문서를 검색하기 위한 문서저장 하부구조로 요약화일로 저장한다.

본 정보검색 시스템의 성능평가는 다음과 같이 삼 단계로 나누어서 평가하였다. 1 단계는 질의어를 요약화일로 바꾸어 명사를 색인어로 추출한 1+2sp의 요약화일과 비교하여 재현률과 정확률을 구한다. 2 단계는 한글의 문제점중의 하나인 명사들의 띄어쓰기 문제를 해결하기 위해 복합명사를 색인어로 추출한 자동색인을 수행한 후 요약화일 형태로 저장을 한다. 그 다음 질의어를 요약화일로 바꾸어 문서가 저장되어 있는 요약화일과 비교하여 재현률과 정확률을 구한다. 3 단계는 재현률을 높이기 위해 질의어를 시소러스를 이용하여 질의어를 확장시킨다. 그 다음 질의어를 요약화일로 바꾸어 복합명사를 색인어로 추출한 1+2sp의 요약화일과 비교하여 재현률과 정확률을 구한다. 마지막으로 세가지 실험을 비교하여 복합명사를 색인어로 추출한 자동색인과 시소러스를 이용한 질의어 확장이 어느정도의 성능을 향상시키는지 평가하였다.

본 연구에서 사용된 KT Test Set의 30개의 질의어중 영문으로된 질의어와 논문의 작가를 찾는 질의어는 본 논문의 연구방향과 거리가 있으므로 이들 3개의 질의어를 제외한 27개의 질의어를 대상으로 실험하였다.

색인방식 / 검색효율	1+2SP	((1+2SP) + 복합명사)	((복합명사 + 질의어확장)
재현율	0.12	0.15	0.44
정확율	0.26	0.35	0.23

표 2. 실험결과

표 2는 위에서 설명한 단일명사, 복합명사, 시소러스를 이용한 질의어 확장에서 검색한 재현률과 정확률을 나타내며 복합명사를 이용한 색인과 시소러스를 이용한 질의어의 확장이 얼마만큼의 검색효율을 향상시켰는지를 보여준다.

예를 들어 18번 질의어 '병렬처리 AND 알고리즘'을 보면 명사를 색인어로 추출하였을 시 '병렬처리'는 '병렬'과 '처리"가 각각 개별적으로 색인어로 처리되므로 관련된 문서는 한건도 추출되지 않는다. 그러나 복합명사를 추출할 시에는 문서상에 '병렬처리"로 있던지

혹은 '병렬'과 '처리'로 분리되어 있어도 모두 '병렬처리'로 처리해 주므로 이런 문서들은 관련 문서로 검색할 수 있다. 그러나 문서에는 '병렬처리'나 '알고리즘'이 없고 관계있는 단어가 그 문서에 존재할 경우 관련정보에서는 이 문서가 질의어에 관련되었다고 하므로 검색효율은 그다지 높지않으나, 시소러스를 이용한 질의어 확장을 한 결과 관련 용어들이 질의어에 포함되므로 재현률을 향상시킬 수 있었다. 그러나 반면에 어느정도 정확률이 떨어졌음을 알 수 있다.

1+2SP 방식만을 이용한 방식의 검색효율에 비하여 복합명사를 이용한 방법은 검색효율이 재현률에서 0.15로 3%정도 증가했을 뿐만 아니라 정확률에서도 0.35로 9%정도 증가하였다. 또한 복합명사를 추출하여 자동색인을 하고 질의어 확장을 한 경우는 재현률 0.44과 정확률 0.23을 얻었는데 이는 1+2SP 방식에 비해서는 거의 비슷한 정확률을 유지하면서 질의확장의 본래목적인 재현률을 3.6배이상 향상시켰으며 복합명사를 이용한 검색효율에 비해서도 재현률이 거의 3배 정도 상승하여 좋은 결과를 보이고 있다. 정확률은 단일명사 색인에 비해 3%, 복합명사 색인에 비해 12%정도 떨어졌으나 회상률의 향상에 비하면 그리 크지 않다고 하겠다.

참고로 본 논문의 실험결과를 똑같은 실험데이타를 사용하여 실험한 김명철[6]씨의 방법에서는 코퍼스(Corpus)에 의존적인 상호관계를 시소러스내에 있는 용어간의 관련도로 이용하였고 관련도를 가진 시소러스를 3차원 구조로 구성한 다음 이 시소러스를 질의확장시에 사용하였다.

실험결과는 재현율 36%와 정확률 0.16%를 기록하였는데 이 시스템과 비교해 보아도 본 시스템은 재현율과 정확률이 모두 앞서 있어 좋은 결과를 보이고 있다.

실험에서 어떤 질의에 대해서는 질의어를 확장한 경우나 확장하지 않은 경우나 상관없이 재현률과 정확률 모두 0으로 나타나는 질의들이 있는데 이런경우는 질의어 포함된 용어들이 문서에 존재하지 않아 색인이 되지 않거나 시소러스에 관련용어들이 제대로 구성되지 않았기 때문이다.

5. 결론

본 논문에서는 한글문서를 대상으로 색인 및 검색을 할 수 있는 정보검색 시스템을 개발하였다.

한글처리의 문제점중의 하나가 복합명사 처리문제인데 이를 위해 1+2SP 방식의 요약화일과 복합명사 생성규칙을 이를 해결하였다. 또한 재현률을 향상을 위해 시소러스를 이용하여 질의어를 확장하였다.

실험으로는 한국통신에서 제공하는 정보과학분야의 코퍼스를 이용하였다. 실험결과 1+2SP 방식의 검색효율에 비하여 복합명사를 처리한 방식의 검색효율은 재현률에서 0.15로 3%정도 증가했을 뿐만 아니라 정확률에서도 0.35로 9%정도 증가하였다. 또한 복합명사를 추출하여 자동색인을 하고 질의어 확장을 한 경우는 재현률 0.44과 정확률 0.23을 얻었는데 이는 1+2SP 방식에 비해서는 거의 비슷한 정확률을 유지하면서 질의확장의 본래목적인 재현률을 3.6배이상 향상시켰으며 복합명사를 이용한 검색효율에 비해서도 재현률이 거의 3배 정도 상승하여 좋은 결과를 보이고 있다. 정확률은 단일명사 색인에 비해 3%, 복합명사 색인에 비해 12%정도 떨어졌는데 이는 시소러스를 이용하여 질의어를 확장할 때 관계없는 용어들이 첨가되었기 때문이다.

향후 연구과제로는 사전의 어휘수가 현재로는 빈약하므로 사전의 어휘량을 보강이 필요하다. 또한 불리언 질의어 뿐만 아니라 자연어 질의 처리등 여러 종류의 질의를 처리하고 좀더 다양한 검색기능의 제공과 시소러스의 확장에 따른 동적재구성을 가능하게 하도록 한다.

참고문헌

[1] Christopher Fox, "A Stop List for General Text", *SIGIR FORUM*, Vol.24, No.1-2, pp. 19-35, Fall 89/Winter 90.

[2] William B.Frakes, Information Retrieval : *data structure & algorithms*, Prentice Halll, 1992.

[3]Van Rijsbergen, C.J., *nformation Retrieval*, 2nd ed., London : Butterworths, 1979

[4] J. Sammet and A. Ralston, "The new computing reviews classification system-Final version,"Communications of the ACM, Vol.25,No. 1, January pp.13-25, 1982

[5] 김명철, 권오욱, 최기선, 김재군, 김영환, "시소러스와 상호정보를 이용한 정보검색 모델",*94년도 한국정보과학회 봄 학술발표논문집*, 한국정보과학회, pp.837-840, 1994

[6]장재우,"한글 텍스트를 위한 요약 화일 기법의 설계," *제3회 한글 및 한국어 정보 처리 학술대회 논문집*,

한국 정보 과학회, pp. 247-256, 1991
[7] 이승선, 송주원, 황규영, 최기선, "TRIE 구조를 이용한 한국어 전자사전을 위한 데이타베이스 인덱스 구조", *94 봄학술발표논문집*, 한국 정보과학회, pp. 849-852, 1994
[8] 이정기, 김철완, 장재우, "요약화일 기법을 이용한 한글문서 검색시스템의 설계", *제 5회 한글및 한국어 정보처리 학술대회 논문집*, 한국 정보 과학회, 1993. pp.47-77
[9] 이창열, 강현구, 장호욱, 박세영, " 자동 키워드 제작기 시스템 설계", *93년도 5회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 71-77
[10] 최기선, "정보검색 시스템 개발연구 ", *한국과학기술원 인공지능연구센터 중점 및 기초연구과제보고서*, pp.17-31, 1993.6
[11] 경제기획원 조사통계국 자료관리과, 한글 자동 인덱싱 작성에 관한 연구, 1990
[12]"*도서관 전산화 입문*" 1981