구문 및 의미분석을 통한 한국어 자동색인

최 기 선*

□목 차□

초 록

통계적 정보 혹은 백터 모델을 이용하는 자동색인은 색인어와 문서간에 관계성을 간접적으로 혹은 상대적으로 계산하기 때문에 검색의 정확도를 높이는데 한계가 있다. 이 보다는 적극적인 방법으로 언어학적 정보와 인공지능의 기술을 이용하여 색인어의 관계성을 계산하는 방법론을 소개한다. 동사의 격틀을 이용하여 개발된 시스템으로부터 언어적 분석 방법의 가능성을 찾아볼 수 있다. 미래의 정보검색은 사용자 중심으로 구성이 되어 사용자에 대한 연구가 깊게 반영이 되어야 할 것이다.

키워드 : 정보검색, 자동색인, 구문분석, 의미해석, 전자사전, 형태소 해석, 지능형 검색, 문서요약, 흑판모델

ABSTRACT

The inherent limitation of the conventional approaches in automatic indexing lies in the fact that they compute the relevancy between index terms and documents rather indirectly or relatively. As an alternative the anlaysis of document texts seeks a means of establishing a direct relevancy of the terms. More rigorous linguistic analysis will ensure better chance of relevancy. Various semantic topologies among terms may suggest the sufficient quality for relevancy.

The enhanced and guaranteed relevance will allow the high precision of retrieval. Along with this line the on going project in KAIST pursues the user oriented retrieval system that spawns still may other issues that are not common in traditional perspective.

1. 정보 검색시스템

대량의 정보를 저장하고 검색하기 위한 노력이 제1세대 컴퓨터부터 시작될 만큼 정보검색 분야의 중요성은 일찍부터 인식되어 왔다. 더구나 오늘날에는 그 어느때보다도 정보검색 시스템에 대한 필요성이 높지만 높은 욕구를 충족시킬 만큼 충분히 정교한 시스템의 개발이 이루어지지 못하고 있다. 정규화가 잘된 정보에 대해서는 데이타 베이스 분야의 발전에 힘입어 어느

* 한국과학기술원 교수

정도 조작이 가능하게 되었으나 정규화할 수 없는 더 많은 정보에 대해서는 효율적인 검색이 쉽지 않다. 일반적으로 정보검색 시스템에서 다루는 것은 정규화할 수 없는, 혹은 한다해도 의미가 없을 자연어 문장들을 대상으로 한다. 자연어 문장들이 모여 하나의 문서를 형성하고 문서는 보통 검색 단위가 된다.

사용자가 필요로 하는 문서에 대한 질의표현과 저장된 문서들을 구분하는 표현간의 유사성을 계산해서 관련되는 문서를 찾아내는 것이 정보검색의 문제라고 할 수 있다. 따라서 기본적인 정보검색 시스템은 적어도 질의 표현과 문서 표현에 대한 정의와 함께 문서로부터 문서를 대표하는 표현을 도출하는 방법 그리고 유사성 계산을 통한 검색 방법등을 구현함으로써 설계되어진다. 여기에서 가장 어려운 문제로서 문서를 대표할 수 있는 표현을 도출하는 색인과정이라고 할 수 있는데 이 부분의 발전은 직접적으로 검색 시스템의 성능 향상을 가져올 것이다.

다음 장에서는 일반적인 정보 색인 및 검색 방법에 대해서 살펴본 후 이들 방법들의 한계점에 대해서 논의한다. 이어지는 3장에서는 KAIST에서 개발해 온 검색시스템에 대하여 그리고 4장에서는 KAIST에서 계획하고 있는 새로운 색인방법과 사용자 위주의 지능형 검색 시스템의 구성에 대하여 논의한다.

2. 기존의 색인 방법과 검색 방법 및 그 한계

문서나 질의의 표현 단위로 하나의 단어나 혹은 여러개의 단어를 복합적으로 사용하는 것이 일반적이다. 이를 색인어 혹은 키워드(Key Word)라고 부르듯이 문장을 색인하는데 쓰기 위한 것으로 문서의 내용 혹은 질의문의 내용을 대표할 수 있어야 한다. 문서를 나타내는 정확한 단어나 어휘를 뽑아 낸다는 것은 현재까지의 기술로는 용이하지 않아 단어의 빈도수나 특정 단어가 전체 문서들 안에서 색인어로써 갖는 상대적 가중치, 혹은 단어의 확률적 가중치 등에 의하여 문서를 대표하는 색인어를 선택하는 것이 대표적인 방법들이다 [Salton 89].

검색 시스템을 평가하는데 일반적으로 두가지 비율을 사용한다.

$$\text{정확도(Precision)} = \frac{\text{검색된 관련성이 있는 문서의 수}}{\text{검색된 문서들의 수}}$$

$$\text{회상도(Recall)} = \frac{\text{검색된 관련성이 있는 문서의 수}}{\text{관련성이 있는 문서의 전체 수}}$$

그 하나는 검색되는 문서들의 관련성 여부에 따른 정확도이고 또 하나는 검색된 관련성 있는 문서들의 수에 따른 회상도이다.

이상적인 시스템은 이 두가지 비율이 높은 경우이겠지만 현재까지의 기술은 두 변수간의 심한 반비례 관계를 바꾸지 못하고 있다. 이 반비례 관계는 근본적으로 검색된 문서들의 수와 정확도간의 반비례관계에 기인한다. 따라서 검색된 문서들의 수의 증가와 함께 검색되는 관련된 문서가 같이 비슷한 비율로 증가한다면 정확도가 항상 높게 유지될 것이고 두 비율간의 반비례 관계가 완화될 것이다. 색인어는 불용어가 될 수 없으며 너무 특정적이어서도 안되고 너무 일반적이어서도 안된다. 너무 특정적인 경우 사용자가 색인어를 사용할 가능성이 희박한 경우로서 그 단어의 의미보다 상위개념의 단어를 찾아 색인어를 쓰는 것이 바람직한데 이를 위해 시소러스를 사용한다. 이와는 대조적으로 너무

일반적인 경우 그 단어가 특정 문서를 대표하는 기능이 떨어져서 검색효율을 저하시키게 되는데 이를 위해서는 몇 개의 단어를 모아 하나의 색인어로 간주함으로써 일반성을 낮출 수 있다.

어떤 정보검색 시스템안에서 사용되고 있는 색인어가 n개가 있다면 각 문서는 n차원의 벡터로 색인되는 것으로 볼 수 있고 따라서 문서간의 관계는 벡터 공간의 위치에 의해 이해된다. 마찬가지로 질의 표현도 벡터값으로 해석하여 색인 벡터와의 유사성 계산에 의하여 문서의 선택여부를 결정할 수 있다. 이외에 많은 다른 방법이 있으나 이중 가장 많은 연구가 이루어진 것으로 확률적 검색 모델을 들 수 있다.

기존의 정보검색 방법론에 있어서 몇가지 문제점들은 다음과 같다.

1. 색인어가 문서의 내용을 충분히 대표하지 못할 뿐만 아니라 색인어와 문서간의 관계성에 대한 확신을 가질 수 있다.

2. 단어만으로 문서의 내용을 충분히 표현하는데 어려움이 있다.

3. 확률적 방법에서는 시스템의 completeness를 보장할 수 없다.

이러한 문제들은 문서의 내용을 파악해서 색인어를 뽑고자 하는 노력의 부족에 기인한다. 이를 위해서는 인공지능의 기술이 성숙되어야 하는데 그렇지 못한 것이 문서 분석방법을 기피하게 하는 직접적인 원인이 되었을 것이다. 그러나 문서의 내용에 대한 분석을 하지 않고서는 색인어의 관계도에 대한 확신을 할 수 없게 되고 검색시스템의 성능은 한계에 부딪칠 것이다.

3. 자연언어 처리기술을 이용한 색인

이 절에서는 KAIST에서 개발해온 구문분석을 이용한 색인어추출기를 중심으로 자연언어 처리기법을 이용한 색인어 추출에 대하여 논한다. 기본적으로 자연언어 처리기술을 색인어 추출시스템에 도입하는 목적은 문서의 문맥을 반영하는 언어학 정보를 이용해서 적절한 색인어를 찾고자 하는데 있다. 언어학 정보는 구문, 의미 등에 걸쳐 다양한 형태로 존재한다.

3.1 구문 정보

문맥상 중요한 어구를 가려내는데 필요한 언어학 정보에는 여러가지 단계가 있을 수 있다. 예를 들어 FASIT(A Fully Automatic Syntactically Based Indexing System)은 색인어를 추출하는데 완벽한 구문분석이나 의미정보를 이용하고 있지 않다 [Dillon 83]. 이 시스템은 색인어로서의 경향을 나타내는 특정 구문유형을 이용하고 있다. 그러나 PHRASE [Ear 173] 같은 시스템은 완벽한 구문분석에 의존한다. 색인어추출방법에 구문정보를 이용하면 상당한 이득이 있다고 실험결과를 토대로 주장하는 사람도 있다[Smeaton 86].

최근의 연구결과에 따르면 언어학 정보의 유형은 어구의 문맥 중요도를 결정하는데 도움이 되는 것들에 제한된다. 그러한 한 유형으로 주로 명사구인 구문 유형을 들 수 있다.(그림 3-1). 그림 3-1과 같은 유형은 많은 문서들을 분석하여 작성될 수 있지만 많은 예외 상황이 발견된다. 이러한 구문만으로는 어구의 문장에서의 역할을 충실히 설명할 수 없으므로 더 많은 정보가 필요하다.

(det) (art) noun

(det) (art) noun conj. verb

(det) (art) adverb adjective noun
(det) (art) noun of

〈그림 3-1〉 간단한 영어 구문 유형

어떤 어구는 다른 어구를 강하게 강조하는 경향이 있다(그림 3-2). 이런 단어 의존적 유형표를 이용하면 그러한 유형에 대하여는 색인어의 질을 높일 수 있는 반면 여러 문서에 대한 일반성을 보장하지는 못한다. 그래서 단어 의존적 유형을 이용한 제한분야에 이용되는 정보검색 시스템을 많이 발견할 수 있다는 것은 놀라운 일이 아니며 색인 시스템에서도 이같은 유형을 이용할 수 있다.

좀더 복잡한 색인은 각 항목에 대한 의미 명세를 포함한 광범위의 정보를 포함한 사전을 필요로 할 수 있다. 각 항목에 대한 의미명세는 시간과 노동이 많이 투입되는 작업이지만 어느정도의 일반성 및 포괄성을 제공해 줄 것이다. 의미표현 방식은 매우 다양하기 때문에 그 응용에 따라 적절하게 취해야 한다.

the importance of X	emphasize X
the influence of X	be surprised at X
the significance of X	depends on X
the advantage of X	X consists of
the power of X	X decides

〈그림 3-2〉 강조를 나타내는 영어 단어 유형

KAIST에서 개발하고 있는 색인 추출기는 의미정보와 심층격 정보를 포함하는 사전을 이용하고 있다. 심층격 정보는 동사가 취할 수 있는 보어의 수와 격에 관한 정보를 담고 있다. 동사는 다른 어구의 중요도를 결정하는 척도가 된다는 것이 심층격 정보를 이용하는 우리의 철학이다. 심층격 정보는 단어 의존적이고 또한 모든 영역에 적용할 수 있을 정도로 일반적이다.

3.2 심층격과 한국어 문서의 색인

심층격이란 동사가 반드시 필요로 하는 필수격을 의미한다. 한국어는 부분 자유어순어에 속하므로 문장내에서의 어구의 역할인 '격'은 조사에 의해 결정된다. 심층격을 나타내는 조사가 동시에 문장 구문성분을 나타내기도 하기 때문에 영어에서와 같이 표층격과 심층격의 구분이 한국어에서는 명확하지 않다. 필수격 어구가 문장내에서 중요한 역할을 한다는 것은 사람이 문서에서 추출한 색인어에 수의격 어구가 얼마나 포함되는지를 보면 알 수 있다. 그림 3-3은 수의격이 신문기사에서 추출되는 색인어로 부적절하다는 것을 나타낸다 [최기선 91].

기사수	수의격	색인어인 수의격	고유명사인자유격색인어
150	474	71(15%)	41(57%)
147	400	58(14.5%)	42(72%)
154	433	60(13.6%)	37(66%)

〈그림 3-3〉 수의격의 색인어로서의 가치

필수격의 판정은 격조사와 보조사를 통해 이루어진다. 각 필수격에 대해서는 대표 조사군들이 규정되어 있다. 그러나 조사의 격판정 애매성으로 인해 대응 어구의 역할을 충분히 보장하지는 못한다. 이로 인해 사전에 부가적인 정보가 적재되어야 한다. 그러나 조사가 나타나는 같은 범위의 동사가 한정하는 격을 조사함으로써 조사가 내포하는 많은 경우의 애매성은 해소될 수 있다. 필수격을 판정하는 단계를 살펴보면 다음과 같다.

길동-은 27일 한국-에서 떠났다.

길동-은 27일 철수-를 한국-에서 만났다.

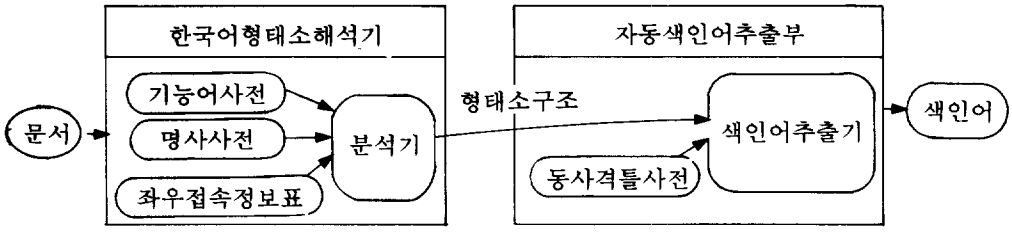
조사 '에서'는 '시발'이나 '장소'를 나타내는 대표조사로 등록되어 있다고 가정한다. 이 두 문장에서의 애매성은 동사에 의해 쉽게 해소된다. '떠났다'는 과거 형태로 사전에 필수격으로 '시발'을 취하도록 등록되어 있다. '만났다' 역시 과거 형태로 경험자와 목적어를 필요로 하고 '시발'이 '장소'를 포함하지 않는 것으로 등록되어 있다. 이것은 두번째 문장에서 '장소'를 나타내는 '에서'는 '만났다'에 대하여 수의격임을 의미한다. 그래서 첫번째 문장의 '한국'은 두번째 문장에서의 '한국'보다 더 높은 가중치를 갖는다. 둘째 문장의 '한국'은 색인어로서의 자격이 낮다고 보는 것이다.

3.3 실험과 그 결과

이 항목에서는 KAIS(Kaist Automatic Indexing System)에 대해서 살펴본다. 개발된 시스템은 한국어의 특성을 최대한 이용하려 했다. 첫째로 한국어의 특성은 영어와 달리 어미변화와 조사체계가 복잡하다는 것을 들 수 있다. 이러한 복잡한 어미변화 체계는 어구에 대한 형태소구조 분석을 어렵게 한다. 또 다른 복잡한 문제는 한국어의 자유어순성이다. 이로 인해 한국어의 구구조 및 문구조에 대한 분석이 필요하였다.

3.3.1 시스템 구성

KAIS는 크게 형태소해석기와 자동색인부로 구성되어 있다 [그림 3-4]. 형태소해석기는 기능어사전, 명사사전, 좌우접속정보표를 이용해 한 어절씩 읽어들여 그 형태소 구조를 밝혀낸다. 기능어사전은 동사와 형용사의 어간, 어미, 조사, 접두사, 접미사 등을 포함하며 각 항목에 대하여 접속정보를 가지고 있다. 한국어는 용언의 어미변화가 많으므로 이에 대한 정보가 필요한데 이를 위하여 좌우접속정보표를 이용한다. 이것은 어절내의 각 형태소가 접속가능한지를 결정하는데 이용된다. 명사에 대한 접속정보는



〈그림 3-4〉 KAIS의 구성도

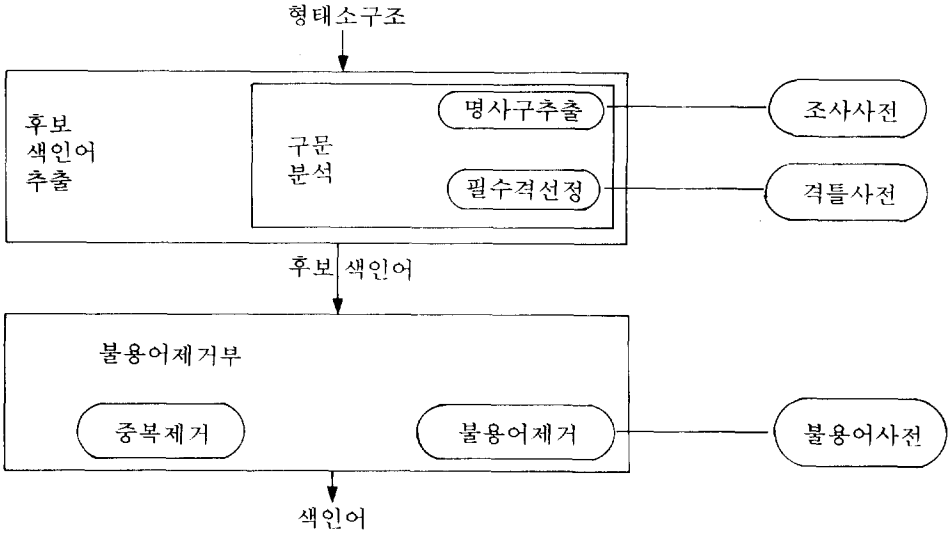
기능어의 그것에 비해 단순하다. 이런 이유로 시스템의 효율성을 위해 사전을 구분하여 구현하였다. 현재 기능어는 2,600항목, 명사사전은 65,000항목을 수록하고 있다.

한국어 어절을 분석하는 방식에는 여러가지가 있을 수 있지만 실험결과를 통하여 대부분의 경우에 가장 적절한 형태소결과를 효율적으로 얻어낼 수 있는 방법은 좌로부터의 최장일치법이라는 가정으로 형태소구조를 분석하고 있다[최기선 91]. 이 방법에 부가적으로 다품사로 인한 애매성해소와 사전미등록어에 대한 처리를 위한 많은 경험적 정보를 이용하고 있다.

다음 구성요소는 형태소해석 결과와 동사격틀사전을 이용하여 문서의 내용을 대변할 수 있는 색인어를 추출할 수 있는 자동색인어추출부이다. 색인과정은 다음의 두 단계로 이루어져 있다.

1. 구문 분석
2. 표층격에 의해 불용어와 수의격어를 제거하고 필수격과 대명사를 색인어로 선택한다[그림 3-5].

구문분석 방식으로는 문장의 표층구조만을 생각하는 격틀분석법(Case Frame Analysis)을 이용하고 있다. 문장내에서 필수격의 역할이 색인어가 가져야 할 특성을 갖기 때문에 표층격에서 색인어를 추출하는 것이 충분하다고 기본적으로 가정하고 있다.



〈그림 3-5〉 색인부

후보색인어 추출과정에서는 주어진 입력문장에 대해서 간단한 구문분석을 하고 명사구의 격을 판정할 수 있는 조사와 보조사표를 이용해서 명사구를 추출한다. 이것이 한국에 적용된 격문법이다. 그리고나서 격틀사전을 이용하여 필수격을 뽑아낸다. 색인과정의 마지막 단계는 중복

된 어구를 제거하고 불용어사전을 이용해 불용어를 제거하는 것으로 이루어진다. 불요어는 문제의 의미를 나타낼 수 없는 보편어나 접속어 등을 포함한다.

3.3.2 성능 평가

성능평가기준으로 문서에 대한 적절한 색인어가 선정되었는가에 주안점을 주었다. 평가에 사용된 문서는 주요 일간지에서 사회, 경제 분야의 기사 90개를 뽑았다. 각 기사는 200~400어절로 이루어졌고 전문가가 선정한 40~60여개의 색인어가 각각 선정되어 있다. 우선 자동색인의 적절성을 평가하기 위하여 사람이 선정한 색인어와 일치한 비율을 측정하였고, 부적합한 색인어의 비율을 측정하였는데 이것은 사람이 선정하지 않은 색인어의 비율을 구하였다.

1. 적절성 = $|S \cap H| / |H|$

2. 부적절성 = $|S - H| / |H|$

H : 사람이 선정한 색인어 집합

S : 색인시스템이 선정한 색인어 집합

이상적인 시스템은 적절성이 높고 부적절성이 낮아야 하나 현재로서는 현실적인 시스템은 고적절성은 항상 고부적절성을 수반한다. 그래서 시스템은 두 측정치의 중간치에서 타협을 해야 한다. 실험결과는 그림 3-6에 나타냈다.

[Salton 88]에 의하면 자동색인시스템은 적절성이 66%보다 높고 부적절성이 100~200%이면 좋은 시스템이라고 하였다. KAIS는 수의격을 제거하고 필수격만을 색인어로 추출하기 때문에 부적절성이 뛰어나다. 이러한 우수성에도 불구하고 KAIS는 색인과정이 형태소해석기에 의존도가 높으므로 형태소해석기를 향상시키는 것이 과제로 남아 있다.

	사람선정	기계선정	일치색인	적절성	부적절성
기사1	1,148	1,424	903	79%	45.4%
기사2	972	1,417	795	82%	63.9%
기사3	1,365	1,638	1,066	78%	41.9%

〈그림 3-6〉 색인어 적절성 실험결과

4. 지능형 색인 및 검색

단순한 언어정보만으로는 문맥을 파악한다거나 특정단어가 그 문맥에서 갖는 중요성 등을 찾아내는 것은 쉽지 않다. 좀 더 정확한 색인을 위해서는 언어정보 뿐만이 아닌 지식베이스와 추론체계까지도 도입해서 문맥구조를 생성해서 각 후보 색인어의 중요성 혹은 관계성을 계산해야 한다. 그러나 이 부분의 연구는 아직 성숙되지 않은 상태여서 일반적인 응용시스템에 적용하기에는 무리가 있다. 실제로 일반적인 정보검색 시스템은 다양한 내용의 문서를 대상으로 하기 때문에 이들 문서를 이해해서 색인한다는 것은 현실성이 없다.

현재 KAIST에서 연구 개발을 목표로 하는 차세대 정보검색 시스템은 다음과 같은 사항을 염두에 두고 설정되었다(그림 4-1).

1. 새로운 추론 모델 및 사전구조 등의 개발을 통한 지능형 색인방법에 기초하여야 한다.

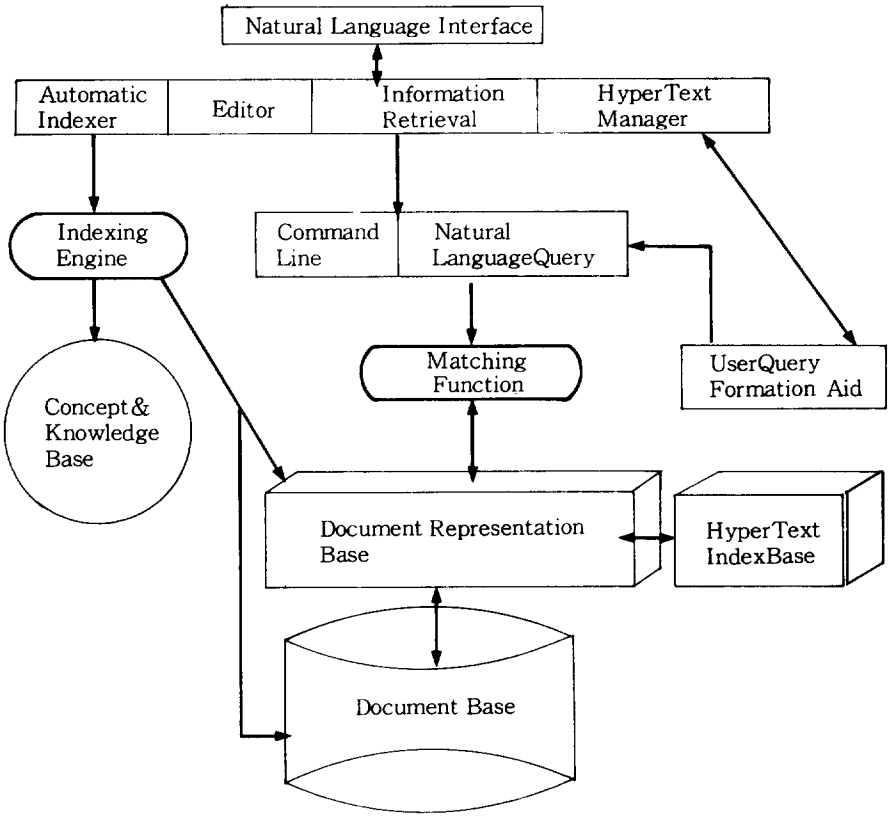
2. 새로운 색인방법은 현재의 하드웨어 및 소프트웨어 기술안에서 단순히 실험뿐만이 아니라 실용화가 될 수 있어야 한다.

3. 사용자를 위한 최대의 편의 기능을 갖추어야 한다.

4. 사용자 위주의 종합검색 시스템은 문서를 만들 수 있는 편집기, 문서 입력기능 및 문서요약기능을 통한 질의어 형성 도움기능, 색인기능, 자연언어 인터페이스, 그리고 사전관리 기능 등을 포함한다.

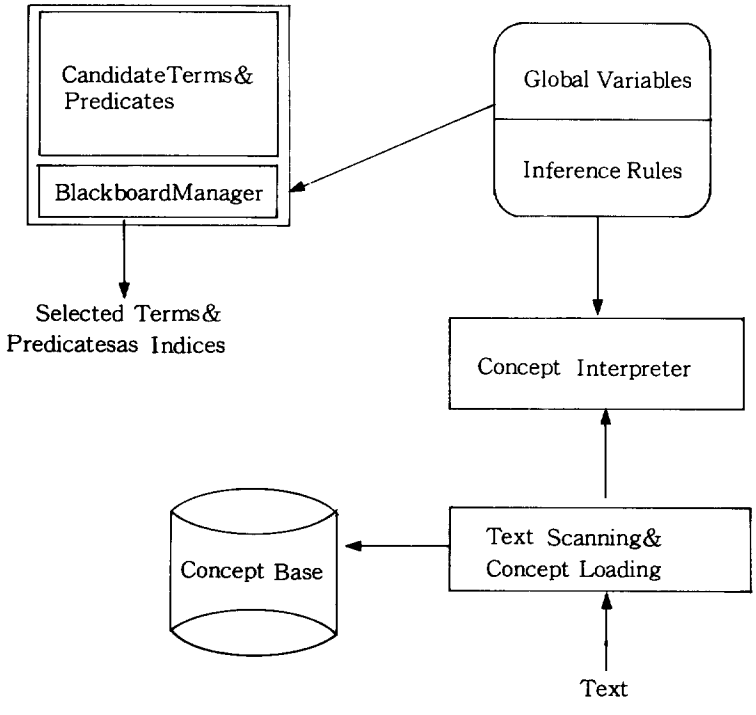
지능형 색인기

어떤 단어의 중요도는 그 단어가 문장내에서 갖는 구문적 역할, 다른 주변단어와의 의미적 관계, 그리고 앞서 분석된 문장들의 단어들과의 의미적 관계를 고려해서 계산될 수 있다. 문맥구조도 유용할 수 있지만 여기서는 고려가 되지 않는다. 무엇보다도 한 문장의 의미를 이루기 위한 구성단어들의 의미의 성공적인 결합은 동사에 의해서 이루어진다. 각 동사는 그 동사가 지배하는 단어들의 중요도에 직접적인 영향을 미친다. 동사뿐만이 아니라 어떤 품사의 어휘라도 주변어휘의 위상에 영향을 끼치는 것들은 모두 같은 방법으로 사전에 기록된다. 또한 어휘에 따라서는 문중에 나온 생물체의 여러가지 상태의 변화를 가져오게 되는데 이러한 정보도 사전에 기록하게 된다.



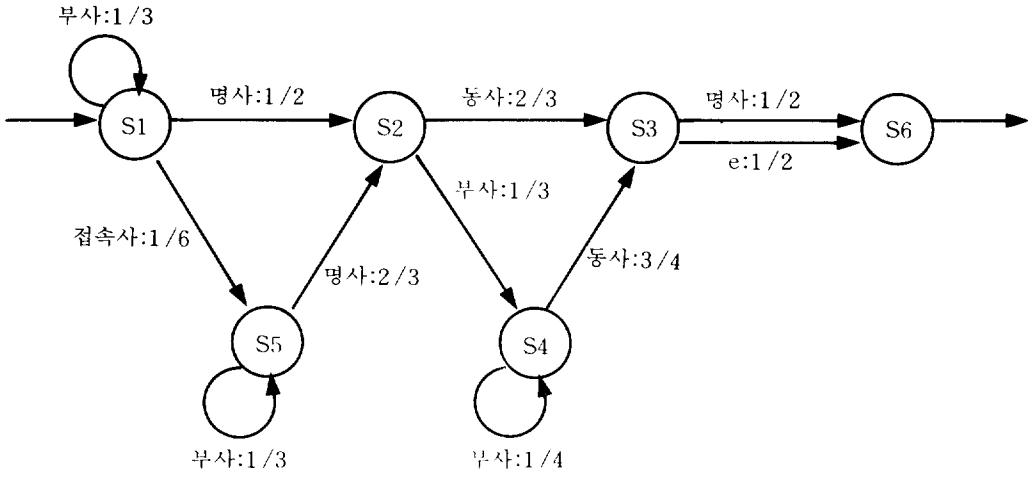
〈그림4－1〉 지능형색인및검색시스템구성

사전의 각 단어에는 품사와 같은 언어학적 정보 외에도 주변어휘와 관계하여 가지는 역할에 관한 정보가 기입되어야 한다. 그림 4.2에서 볼 수 있듯이, 사전의 개념들은 문장의 단어에 의해 발화되어 독자적으로 시스템의 Global변수를 변화시킨다. 여기에는 일정한 규칙이 있으며 BlackBoard Manager나 추론규칙에 담겨져 있다. 추론 규칙은 흑판위에 있는 개념들의 관계를 이해하기 위해서, 그리고 각 개념들이 시스템 변수에 영향을 가하기 전에 여과를 위한 것이다. 문서의 분석이 끝나고 난 후 흑판에 남아있는 단어나 논리적 표현중에서 높은 중요도를 가진 것들이 색인어로 뽑아지게 된다.



〈그림4－2〉 자동색인을위한흑판모델

위에서 설명한 것과 같은 색인기 구현의 커다란 문제점은 복잡하고 커다란 규모의 사전작성에 있다. 따라서 시스템의 성능은 사전의 정확도에 달렸다고 볼 수 있다. 색인작업은 각 단어의 품사를 필요로 하기 때문에 한 문장의 구문해석이 다 끝난 후에 하는 것이 바람직하다. 설계되고 있는 시스템은 확률적 문법에 의한 구문해석 방법을 고려하고 있는데 그림 4.3은 간단한 예를 보여준다.



〈그림4-3〉 확률적문법

그림 4.3의 예는 매우 단순화된 모습이며 실제는 ATN정도의 기능을 가지게 된다. 즉 state가 변할 때마다 조건이 만족되어야 하고 만족되는 것중 가장 가능성이 높은 것을 택하며 그다음 state로 넘어가기 전에 중간 구조를 만들 수 있다. 가장 높은 가능성의 경로를 계산하는 것은 경로의 확률값에만 의존하는 것이 아니라 전체에서 각 state가 갖는 비중을 고려해서 계산하는 회귀경로에 입각한 통계적 검색(Return Path Based Probabilistic Search)방법을 사용한다. 기존의 구문 분석 방법으로는 다양한 문장의 종류를 다루기가 쉽지 않으며 특히 문장의 길이가 길어지면 시간이 많이 걸릴 뿐만 아니라 실패할 가능성이 매우 높다. 따라서 매번 완벽한 구문해석을 보장은 못하지만, 빠르고 robust한 장점때문에 확률적 방법을 도입하는 것이다.

지능형 검색

색인어의 관계도(Relevancy)에 대한 어느정도의 확신을 가질 수 있다면 검색의 문제는 상대적으로 쉬워진다. 지금 현재 고려하고 있는 검색 방법은 단순한 유사성 비교만으로 충분할 것이라고 믿어진다. 우선 사용자는 자연언어로 필요한 문서에 대한 설명을 할 수 있어야 하며, 문서에 대한 설명은 색인때와 같은 방법으로 해석되고 해석결과는 각 문서의 색인어와 유사성 계산에 사용된다. 지능형 검색의 핵심은 자연언어 인터페이스나 유사성계산에 있는 것이 아니라 사용자가 그가 필요로 하는 문서를 찾도록 도와주는 기능에 있다. 많은 사용자들은 그들이 필요로 하는 문서에 대한 정확한 결정을 내리지 못하거나 설명하는데 어려움을 가지고 있다. 또 관련된 많은 문서들 중에서 적절한 것을 고르는데 시스템이 도와줄 수 있을 것이다. 사용자가 자신의 질의를 수정하기 위해서는 어떠한 문서들이 있는지 알 필요가 있다. 하이퍼 텍스트 기술은 사용자로 하여금 문서들을 검토하고 필요한 것을 고르게 할 수 있는 여러가지 기능을 제

공할 수 있다. 색인어나 어휘를 자연어표현으로 사용자에게 보여줌으로써 사용자가 문서의 내용을 쉽게 알 수 있게 하는 기능도 유용할 것이다.

5. 결 론

기하급수적으로 늘어나는 정보를 관리하기 위해 자연언어 처리기술을 이용한 응용시스템의 필요성이 점증하고 있다. 정보검색은 그 한 예로서 자연언어 처리기술이 유용하게 사용될 수 있는 분야이다. 이런 취지에서 한국어 정보 및 문서관리 그리고 다른 응용분야의 핵심이 되는 한국어 처리기술에 대한 지속적인 연구개발이 더욱 강조된다. 정보검색 시스템은 지금까지 이론적인 취지에서보다 필요성에 의한 응용적 각도에서 더 많이 연구되어 왔다고 할 수 있다. 본질적 문제의 복잡성을 고려할 때 통계적 방법은 문제를 쉽게 접근하게 하였고, 실용적인 시스템의 개발을 가능하게 하였다. 그러나 검색결과의 질에는 한계가 있으며 이 한계점의 극복을 위해서는 다른 방법을 찾지 않을 수 없다.

최근에 KAIST에서 추진되는 연구개발의 방향의 관점에서 정보검색에 대하여 살펴 보았다. 미래의 시스템은 사용자위주의 지능적 기능을 갖추게 될 것이고, 문서의 관리도 통계적 방법과 함께 지식기반 및 추론에 입각해서 할 수 있을 것이다.

참 고 문 헌

1. [최기선 91] 최기선, 한국어 문서로부터의 색인어 추출에 관한 보고서, DACOM, 인지과학회, 1991
2. [김덕봉 90] 김덕봉, 최기선, 강재우, 한국어 형태소해석기와 그 사전 : 접속정보를 이용한 철자검색기, 언어연구, vol.26, No.1, 1990.
3. [CHOI 91] CHOI, KEYSUN, Syntactic Analysis based Automatic Indexing for Korean, Tke 91, Shanghai, China.
4. [DAVIS 90] DAVIS, ROS SACKS Using syntactic analysis in a document retrieval system that uses signature files, 13th International Conference on Retrieval, 1990. p.179.
5. [DILLON 83] DILON, MARTIN Fully Automatic boole indexing, Journal of Documentation, vol.339, No.1, 1983, pp. 135−154.
6. [EARL 73] EARL L. L. Use of word government in resolving syntactic and semantic ambiguities, Information Storage and Retrieval, 9(12), 1973, pp. 6339−664.
7. [JORL 87] JORL, L. FAGAN Automatic phrase indexing for document retrieval : An examination of syntactic and non-syntactic methods, ACM, 089791-232-2, 1987.
9. [KLING 73] KLINGBIEL, PAUL H.Machine aided indexing of technical literature, Information Storage and Retrieval, vol.9, 1973, pp.79−84.
9. [KLING 73] KILINGBIEL, PAUL H. A technique for machine-aided indexing,

Information Storage and Retrieval, vol.9, 1973, pp.477－494.

10. [LOIS 72] LOIS, L.EARL The resolution of syntactic ambiguity in automatic language processing, ISR, vol.8, 1972, pp.277－308.
11. [MARTIN 85] MARTIN, DILLON, ANN S. GRAYFASIT : A fully automatic syntactically based indexing system, The American Society for Information Science, March 1985, pp. 477－494.
12. [OLNEY 76] OLNEY, JOHN A new technique for detecting patterns of term usage in text corpora, Information Processing and Management, vol.12, 1976, pp.235－250.
13. [SALTON 89] SALTON, GERARD Automatic text processing, Addison Wesley, 1989.
14. [SALTON 89] SALTON, GERARD On the application of syntactic methodologies in automatic text analysis, Cornell University TR, 1988.
15. [SALTON 88] SALTON, GERARD Automatic text indexing using complex identifiers, ACM, vol.12, 1988.
16. [SALTON 65] SALTON, GERARD Automatic information organization and retrieval, McGraw-Hill Book Company, 1965.
17. [SMEATON 86] SMEATON, ALAN F. Incorporating syntactic information into a document retrieval strategy, ACM conference on R & D in information retrieval 1986, p.103.
18. [SPARCK 74] SPARCK, JONES Automatic indexing, Journal of Documentation, 30(4), 1974, pp.393－422.