□특집□

정보 검색

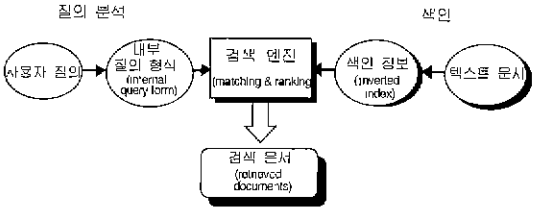
강 현 규† 박 세 영††

◈ 목 차 ◈

1. 서 론

현대 정보화 사회에서 정보는 사람의 관리가 불가능 할 정도로 쏟아져 나오고 있다. 따라서 이러한 수 많은 정보들을 체계적으로 저장하고 검색하는 시스템이 더욱 필요하게 되었는데, 이러한 시스템을 정보 검색 시스템(information retrieval system)이라고 한다. 정보 검색 시스템이란 시스템의 사용자가 필요로 하는 정보를 수집하여 정보 자료의 내용을 분석한 뒤 찾기 쉬운 형태로 조직하여 두었다가 정보에 대한 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템을 말한다.

전형적인 정보 검색 시스템은 (그림 1)과 같다. 먼저 색인(indexing)은 문서의 내용을 분석하여 색인어(index term, term, keyword)들로 구성된 문서와의 이차적인 색인 정보를 생성한다. 일반적으로 역 색인(inverted index)으로 구성한다. 검색에서 사용자의 질의(user query)에 대하여 내부의 질의 형식(internal query form)을 형성한 후 역 색인 정보와 검색 엔진(retrieval engine)을 통하여 매칭(matching) 및 순서화(ranking)함으로써 문서를 검색하게 된다.



(그림 1) 전형적인 정보 검색 시스템

정보 검색의 분야에는 색인(indexing) 및 요약화일, 역색인화일, PAT 트리 등의 색인 구조를 비롯하여 불리안 모델, 벡터 공간 모델, 확장된 불리안 모델, 퍼지 집합 모델, 확률 모델, 지식 기반 모델 등의 검색 모델들, 적합성 피드백, 데이터 퓨전, 질의 확장, 문서 분류 등의 기법과 데이터베이스, 분산, 멀티미디어 정보 검색의 잇슈들 그리고 최근에 정보 여과, 정보 시각화, 구조 정보 검색 및 응용으로서 디지털 라이브러리 등 다양한 분야가 있다. 본 고는 한글공학 이라는 특집으로서의 정보 검색 잇슈와 관련한 한글 색인어 추출의 문제를 중심으로 다룬다. 전반적으로도 한글과 관련한 부분을 중점적으로 다룰 것이다.

본 고의 2장에서는 자동 색인에 대하여 간단히

† 정회원 : 한국전자통신연구원 선임연구원
†† 정회원 : 한국전자통신연구원 책임연구원

언급할 것이며 제 3장에서는 한글과 색인어 추출의 문제를 전반적으로 다룰 것이다. 제 4장에서는 정확한 색인어 추출과 관련하여 정보 검색을 위한 형태소, 구문, 의미 분석의 장단점 및 기타 몇 가지 정확한 색인어 추출의 방향에 대하여 언급한다. 제 5장에서는 한글과 관련한 정보 검색의 향후 동향에 대하여 기술하고 마지막으로 6장에서 결론을 맺는다.

2. 자동 색인

색인이란 문서의 내용을 대표하는 용어(주제어)를 색인어로 추출하거나 부여하는 것이다. 색인은 특정한 정보가 필요한 사람에게 그 정보의 위치를 지시해 주는 역할과 방대한 정보원으로부터 가장 유사한 내용의 정보 자료만을 선별해 주는 역할을 한다. 이러한 색인 방법으로는 우선 전문화된 사람에 의한 색인어 추출을 들 수 있다. 색인자가 문서의 내용을 정확히 분석하여 중요한 개념들을 추출하여 색인어로 표현해 주는 것으로 색인자는 자신의 전문 지식에 기초하여 임의로 색인어를 부여하거나 아니면 대개 통제어휘집을 참고하여 미리 통제된 용어들 가운데에서 가장 적당한 색인어를 선택하게 된다.

그러나, 정보가 기하 급수적으로 늘어나고 있는 현실을 감안한다면 사람에 의한 색인어 추출 방법은 비용이 많이 들므로 경제적이라 할 수 없으며 색인어 선택시 일관성 결여 및 색인어 누락이 될 가능성이 있으므로 컴퓨터에 의한 자동 색인(automatic indexing)이 필요하다.

컴퓨터에 의한 자동 색인은 저장된 문서들을 미리 분석하여 그 문서의 주요 단어 또는 주요 어구를 추출(색인어 추출)한 후 찾기 편리한 형태로 저장(색인 정보 구성)하는 것을 말하며, 검색시 문서의 내용 전부를 검색하지 않고 주요어만을 검색함으로써 빠른 시간에 사용자의 요구를 만족시킬 수 있다.

3. 한글과 색인어 추출의 문제

색인어란 정보 검색 시스템에서 어떤 문서에 대해 그 문서의 전체적 내용을 나타내거나, 그 문서를 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어구 등을 의미하며 문서를 유사한 것들끼리 묶을 수 있는 능력을 가지고 있다. 자연언어 검색시 일반적으로 자동으로 색인어를 추출하여 색인하고 검색을 행한다. Salton에 의하면[17] 특정한 단어가 문서 집단 속의 상호 관련 없는 문서들을 분리시키는 능력치가 큰 것이 좋은 색인어가 되고 나쁜 색인어일수록 상호 관련 없는 문서들을 묶어 준다고 하였다.

영어-유럽어의 경우 일반적으로 단어가 공백으로 분리되므로 단어의 분리가 비교적 쉽다는 특징이 있다. 그래서 형태소 분석은 주로 동사의 과거형이나 명사의 복수형과 같이 굴절현상이 일어나는 단어에 대한 원형복원 문제 정도로 인식되어 왔다.

그러나 한글은 영어와 달리 단어의 개념이 명확하지 않다. 한글의 띄어쓰기 단위는 어절이며 한글의 한 어절은 해당 어절의 중심적인 의미를 표현하는 어근과 여러 가지 문법적인 기능을 표시하는 여러 가지 접사가 결합된 형태이다. 접사는 문법적인 의미만을 전달하므로 불용어에 해당한다. 따라서 한글 색인어 추출의 첫 단계는 어절을 분석하여 어근과 접사를 분리하는 것이다. 그러나 한글은 첨가어의 특성상 접사의 종류가 매우 다양하고, 또 여러 가지 형태로 접사가 결합할 수 있어서 어절에서 접사를 분리하는 것이 매우 복잡하다[1].

3.1 띄어쓰기 / 철자 / 맞춤법 오류 문제

신문기사나 웹문서 등과 같이 색인어 추출의

입력으로 들어오는 정보 자료들은 띄어쓰기 오류가 포함되거나 일관성이 결여된 경우가 많다. 오류가 포함되거나 일관성이 결여된 단어는 색인어 추출에 실패하여 사전 미등록어로 추정되므로 색인어로 선택될 가능성이 높아진다. 따라서 잘못된 색인어 추출이 될 가능성이 많다. 그 예는 다음과 같다.

일 외무성, 일외무성
상해 프랑스, 상해프랑스
문서내의, 문서 내의
광역자치단체장후보윤곽
지방자치단체장선거연기검토문서

또한 색인어 추출시 철자나 맞춤법 오류가 포함되어 있는 단어의 처리 문제이다. 정보 검색 시스템에서 처리해야 하는 정보 자료의 종류 및 근원지는 매우 다양하며 문서에 오류 어절이 포함되어 있는 경우도 많다. 이러한 철자나 맞춤법 오류의 문제를 처리 하기 위해서 미리 전처리 작업으로 오류 복원을 해야 하지만 이 또한 그리 쉽지 만은 않다.

3.2 어근 중 정확한 명사 추출 문제

한글은 형태적으로 첨가어에 속하며, 단어의 중심적인 뜻을 나타내는 어근과 문법적인 역할을 표시하는 접사가 결합하여 어절을 형성한다. 한글에서 띄어쓰기 단위는 어절이며, 색인어로 사용되는 체언 어절은 명사와 조사가 결합된 형태이다. 조사는 영어의 전치사와 같이 의미적인 기능보다는 문법적인 기능을 하므로 불용어에 해당한다. 따라서 한글 색인어 추출의 첫 단계는 체언 어절 판별과 체언 어절에서 조사를 제거하는 처리이다. 체언 어절 판별은 한 어절이 용언 어절인가, 수식언 어절인가, 체언 어절인가를 판단하는 과정이다.

체언 어절은 먼저 조사 부분을 불용어로 제거한다. 그러나 한글에서는 백여개의 조사가 있고, 이들이 결합한 복합조사인 경우 5천여개가 가능하다. 또한 대부분의 조사 음절은 체언 내에서도 사용될 수 있으므로, 체언과 조사의 경계 결정에 많은 모호성이 발생한다. 예를들어 원자로는이라는 체언 어절에서 조사를 분리하고자 할 때 이 어절의 분석으로 원자로+는과 원자+로는이라는 두가지 분석이 가능하여 모호성이 발생한다. 이와 같은 예는 다음과 같다.

소라도 => 소, 소라
말과 => 말과, 말
종이 => 종이, 종
지구의 => 지구의, 지구
경기 => 경기, 경
법의 => 법의, 법
강원도 => 강원도, 강원

3.3 복합어 인식 / 분리 / 결합 문제

복합어는 한국어에서 빈번하게 나타나는 색인어의 한 형태이다. 복합명사의 경우 2개 이상의 단일 명사들의 조합으로 이루어져 있고, 그 형태 또한 다양하게 나타나기 때문에 색인 및 검색시 문제를 일으키곤 한다. 이들 복합명사를 적절한 단위 명사로 분석 해야 하지만, 이 복합명사 분석 과정 역시 매우 어렵고 여러 가지 모호성을 발생시킨다. 색인시에는 인식의 문제를 야기시키며, 검색시에는 가중치 부여 등의 문제를 발생시킨다.

- 복합어 인식 / 결합
  가장 강수량이 많은 => 가장강수량
  원자로 등 => 원자로등
  국제적 언어 => 국제언어
  군대가 해산 당하다 => 군대해산
  인도가 독립을 하다 => 인도독립
  연대가 미상이다 => 연대미상
  개발되기 시작되었다 => 개발시작
  정보의 검색 => 정보검색
  정보를 검색하다 => 정보검색

정보를 효율적으로 검색하는 => 정보검색
문헌 정보 검색 시스템 => 정보검색
정보를 색인하고 검색하는 => 정보검색
정보를 색인한다. 그리고 이를 검색하는 => 정보검색

- 복합어 분리
  문서내의 => 문서 + 내의
  거주지도 => 거주 + 지도
  한국대학생선교회 => 한국 + 대학 + 생선 + 교회
  대구지하철공사장 => 대구 + 지하 + 철공 + 사장

3.4 기능어와 관련된 미등록어 추정 문제

고유명사, 전문 용어, 왜래어, 회사명, 지명, 신조어 등의 미등록어에 기능어가 결합된 단어에서 미등록어를 추정할 때는 기능어(조사, 어미, 접미사 등) 를 제외한 나머지 부분을 미등록어로 추정한다. 이때 미등록어의 끝부분이 기능어의 첫 부분으로 사용되는 음절로 구성되는 특수한 경우에 미등록어를 잘 못 추정하는 경우가 발생한다. 이러한 오류는 특수한 음절로 끝나는 미등록어가 일부 조사와 결합할 때에 드물게 발생함으로 무시할 수도 있지만 오류를 방지하려면 오류를 유발하는 미등록어를 사전에 등록하여야 한다. 그러나 이러한 종류의 미등록어를 모두 사전에 등록하기는 불가능하다.

벨기에는 -> 벨기 + 에는
홋가이도 -> 홋가이 + 도
오페라는 -> 오페 + 라는
미테랑 -> 미테 + 랑

3.5 동음이의어 (어의 모호성) 문제

정보검색에 있어서 동음이의어 처리는 중요한 문제 중의 하나이다. 한국어 정보 검색 시스템에는 한국어의 특성상 동음이의어로 인한 단어 자체에서 나오는 어의 모호성으로 자연언어 색인어 추출시나 검색시의 잘못된 결과를 초래한다. 예를 들어 동음이의어인 가장, 국화, 자전, 공전, 다리, 자수, 조선, 전기, 모자, 지구, 피아노, 인도, 말, 산, 배, 신. 실, 차, 한, 등, 눈 등의 색인어들로 인하여 검색시 정확한 검색을 하지 못하는 문제점이 있다.

말1 : 동물
말2 : 언어
말3 : 측량의 단위

3.6 아라비아 숫자나 영문자가 포함된 개념 색인어 추출 문제

일반적으로 한글에 아라비아 숫자나 영문자가 포함된 단어는 색인어로서 가치가 없다. 그러나 비타민-A, 3.1절, 6.25전쟁, 8.15광복 등 특정 개념이나 사건을 의미하는 것은 색인어로 추출 되어야 한다. WINDOWS, WIN95 등과 같은 영문자나 영문자에 아라비아 숫자로 이루어진 단어의 경우에도 정보 자료의 유형과 응용 분야에 따라 색인어로 선택되기도 하고 비색인어로 간주될 수 있다. 그 예는 다음과 같다.

헨리 7세
4중주
제 3차 십자군 전쟁
제 2차 경제 개발 5개년 계획
화엄사 4사자 3층 석탑

3.7 다어절, 구분자, 숫자, 년도, 단위 포함 색인어 문제

- 다어절 형식
  해에게서 소년에게, 허블의 법칙, 헨젤과 그레텔, 혈의 누, 홍경래의 난, 힘의 분해 등과 같이 하나 이상의 어절이 모여 의미를 표현하는 것이다. 색인어 추출시에 다어절에 대한 색인어를 추출할 수 있어야 한다.

따라서 이와 같은 다어절 형식을 사전에 미리 등록할 수 있으나 모든 종류의 다어절을 등록하기란 불가능 하다.

- 구분자 형식
  프로이센오스트리아 전쟁, 625 전쟁, 4-2-4 전법, 1,500미터 달리기 등 용어 사이에 구분자(deliminator)가 들어가 있는 경우를 말한다. 이러한 구분자 형식은 중간점(), 중간바(), 콤마(,), 피리어드(.), 공백( ) 같은 것들로 어떠한 것으로도 바뀔 수 있고 생략되어도 의미가 통하는 것들이다.
- 윗첨자 표제어
  내용 중의 다의 단어를 어떻게 매핑 시킬 것인가 ?
  나는 눈을 보았다. 눈1 / 눈 2
- 내용 중의 년도, 숫자, 도수(65도 C), 율(%), 단위 (km), 수식, 약자(DNA), 원소기호, 한글 자소
  1930, 40 년대의 암흑기를 거쳐..., 1,800 일(5년)이었으며, ...
  75부 1,335 권에 이르렀다., 141.7 km
- 시간 (연대) 추출
  1974년, BC 24년, 세종 14년, 장수왕 5년, 조선선조 25년
  1964년 4월 10일, 19세기, 일본의 소화 45년
  고려 인종, 조선 인종: 인종 2년... (?)

3.8 구(phrase) 색인어 문제

일반적으로 정보 검색에서 단일 색인어 보다는 구 색인어가 검색 효율을 높이는 것으로 알려져 있다. 그러나 단순히 명사형 + 명사 구 색인어만을 보더라도 집사람은 일반적으로 통용되는 반면 아파트사람이나 주택사람은 통용되지 않는다. 이러한 구 색인어를 구별해 내기는 쉽지 않다. 또한 형용사형 + 명사의 구 색인어의 경우 질의자나 문서에서의 가장 높은 빌딩은 아마도 엠파이어 스테이트 빌딩을 염두에 두고 문서에서나 질의에서 표현 하였을 것이다. 그러나 일반적인 정보 검색 시스템들에서는 형용사형 + 명사의 색인어를 고려하지 않기 때문에 색인어 추출시 고려가 되지 않으며 따라서 정확한 정보를 검색하기란 어려울 것이다. 마찬가지로 동사형 + 명사의 구 색인어의 경우 풀을 먹는 동물의 경우 아마도 토끼와 같은 채식 동물을 기대하고 표현 되었을 것이다. 그러나 마찬가지로 현 정보 검색 시스템들은 동사형 + 명사의 구 색인어를 고려하질 않기 때문에 정확한 검색을 기대하기가 어려울 것이다.

- 명사형 + 명사 (*는 비명사구를 의미함)
  집 + 사람, 아파트 + 사람(*), 주택 + 사람(*), 가옥 + 사람(*)
  증명 + 사진, 증명의 + 사진 (*), 칼라의 + 사진 (*)
- 형용사형 + 명사
  가장 높은 빌딩, 가장 빠른 비행기, 긴 잎을 가진 식물
  아름답고 웅장한 산, 다람쥐와 비슷한 동물
- 동사형 + 명사
  풀을 먹는 동물, 지구가 도는 이유는, 속도를 겨루는 경기
  흔들어 뛰어내리기, 매달려 흔들어 팔걸치기

3.9 중심 의미의 해석 문제

단순히 복합명사의 표현이더라도 그 중심 의미가 어디에 있는지 판별하기가 쉽지 않은 경우가 많다. 예를 들면 태양 흑점 변화 주기의 경우 이 표현의 중심의미가 태양에 있는지 흑점에 있는지 변화에 있는지 아니면 주기에 있는지 쉽게 판정을 할 수가 없다. 또한 빙상 선수단 경무대 방문의 경우는 빙상선수단, 경무대 방문은 색인어로서 가치가 있는 것 같지만 선

수단경무대는 색인어로서 무의미한 것 처럼 느껴진다. 아울러 도둑 고양이 와 고양이 도둑은 명사의 순서만을 도치하였음에도 불구하고 그의 중심 의미는 판이하게 다름을 알 수 있다. 제시한 이러한 종류의 색인어들은 그의 중심 의미를 어디에 두고 색인어로서 추출을 할 것인가 하는 것이 과제로 남겨져 있다.

- 태양 흑점 변화 주기 : 1) 태양, 2) 흑점, 3)변화, 4)주기
  십자군 전쟁 발발 원인 : 1) 십자군, 2) 전쟁, 3) 발발, 4) 원인
  세계에서 가장 높은 산은 : 1) 세계, 2) 가장 높은 산
- 빙상 선수단 경무대 방문
  빙상선수단, 선수단경무대(*), 경무대방문
- 도둑 고양이 / 고양이 도둑
  이승만 대통령 / 대통령 이승만
  경찰 수사 / 범인 수사

4. 정확한 한글 색인어 추출

4.1 정보 검색을 위한 형태소 해석

한국어와 같은 교착어에서는 형태소들이 결합하는 과정에서 활용이나 축약 현상이 빈번히 일어나므로 형태소를 분석하는 과정에서 형태소의 분리문제와 원형 복원 문제를 동시에 처리해야 하는 어려움이 있다. 또한 신조어가 빈번히 발생하고 전문용어 등이 방대하여 형태소 분석에 필수적인 사전정보의 구축이 어렵다는 문제가 있다. 따라서 사전에 등록되지 않은 단어 즉 미등록어에 대한 인식 기능이 요구되고 있다.

정보 검색의 색인을 위한 형태소 해석을 통한 색인어 추출은 문장 중의 각 어절을 명사, 조사, 부사 등의 형태소 단위로 분리한 후, 문서나 질의의 내용 표현에 적절한 명사류나 복합어들을 인식/분리/결합함으로써 적절한 색인어를 추출할 수 있다. 그러나 형태소 해석을 위한 규칙이 복잡하고, 형태소 해석 결과의 애매성, 사전 내에 등록되어 있지 않은 미등록어, 비문법적인 어절들, 띄어쓰기나 철자, 맞춤법 등의 오류로 인하여 부정확한 색인어가 추출되는 문제점을 지니고 있다. 따라서 정보 검색을 위한 형태소 해석은 일반 기계 번역이나, 담화 분석 등의 상세한 문장 분석에 필요로 하는 형태소 해석이 아닌 제 3장에서 언급한 형태소 분석에서 해결할 수 있는 정보 검색의 색인어 추출을 위한 형태소 해석기로서 가볍고, 빠른 형태소 해석기가 필요로 된다.

형태소 해석을 이용한 방법은 구현이 비교적으로 간단하고, 한글에도 쉽게 적용하여 사용할 수 있는 장점이 있다. 그러나 이 방법은 형태소 해석을 각각의 단어에 대해 한 다음, 하나의 단어에 대해 빈도수 및 가중치를 계산(태거, tagger)하므로 구 단위(term phrase)의 색인어 추출이 어렵다.

4.2 정보 검색을 위한 구문 분석

문장 내의 특정한 구문적 의미를 가지는 단어, 단어구가 문서의 내용을 나타낼 수 있다는 가정하에서 구문 해석을 자동 색인어 추출에 이용한다. 구문 해석을 이용한 자동 색인어 추출 시스템은 먼저 형태소 해석을 하고 그 결과를 가지고 구문 해석을 한 다음, 구문적 의미를 지니는 특정한 단어 및 단어구를 색인어로 추출한다.

구문 해석을 이용한 자동 색인어 추출 방법은 형태소 해석을 이용한 자동 색인보다는 색인어를 훨씬 잘 추출할 뿐 아니라 구 단위의 색인어도 훌륭히 추출할 수 있다. 그러나 구문 해석을 이용한 자동 색인은 자연언어 구문 해석 결과에서 나오는 필연적 애매성(ambiguity)과 단어 자체에서 나오는 애매성이 있고, 실제적으로 구문 해석기 구현이 매우 복잡하다는 단점이 있다.

4.3 정보 검색을 위한 의미 분석

문서의 의미를 대표하는 색인어를 추출하기 위해서 의미 분석을 통해 색인하는 것이 정공법이라 할 수 있다. 일반적으로 의미해석은 단어, 구, 절, 문장 등을 정규 의미구조로 표현하는 과정이며 여기서 모호성을 해결하는 것이 그 주된 작업이라고 볼 수 있다. 색인어 추출은 이렇게 분석된 구조로부터 색인어의 의미적 비중을 계산하는 것이라고 할 수 있다. 그러나 의미해석을 좀더 포괄적으로 보면 문장이나 문서 내의 의미객체(semantic object)들간의 관계를 찾아내는 작업으로 정의할 수 있고, 다양한 종류의 의미구조가 있을 수 있다. 궁극적으로는 이해 시스템의 기술을 부분적으로 필요한 만큼 수용하여야 한다는 것이 원칙이다. 한정된 분야와 응용범위 안에서는 가능할 수도 있겠지만 투자에 비하여 그 결과에 커다란 변화를 가져오지 않을 수 있다. 따라서 실용성이 본격적인 의미해석의 경우보다 높다고 할 수 있으므로 구현하기 위해서는 많은 노력이 필요하다. 어떤 형태의 의미해석도 지식 베이스(knowledge base)나 지식 표현(knowledge representation), 시소러스(thesaurus) 등 대량의 지식을 필요로 하기 때문이다. 따라서 점진적인 연구와 실험을 통하여 개발하여야 할 것이다.

4.4 정보 검색을 위한 키팩트(keyfact)

정보 검색의 수준을 검색의 정확도에 따라 다음과 같이 세 가지로 나누어 볼 수 있다. 예를 들어 정보를 검색하는 데 사용되는 편리한 도구는 무엇입니까 ?라는 질의어에 대하여 각각의 정보검색 시스템에서 대답할 수 있는 수준은 다음과 같다.

- 키워드기반 정보검색
  정보, 검색 그리고 도구 등의 단어가 들어있는 문서를 검색해 주는 수준
- 키팩트기반 정보검색
  정보를 검색하고, 정보의 검색 그리고 편리한 도구라는 문장이 들어있는 문서를 검색해주는 수준 또는 Netscape, Explorer등의 검색도구가 들어있는 문서를 검색해주는 수준
- 지식기반 정보검색
  구체적인 정보검색 도구인 Netscape, Explorer 등을 대답해 주는 수준
  그러나 키워드기반 정보 검색 시스템은 정확한 검색 결과를 내주지 못할 수 있으며, 지식기반 정보검색 시스템은 가장 이상적인 시스템이긴 하지만 현재의 기술로는 해결하기 어려운 여러 가지 문제점들을 가지고 있다. 따라서 정보 검색을 위한 키팩트기반의 형태소, 구문, 의미 분석을 가미할 수 있을 것이다. 이를 위하여 다음과 같은 경우를 해결 할 수 있어야 할 것이다.
- 같은 의미를 가지는 문장이 서로 다른 통사적 형태로 존재하고 있는 경우
  예를 들어 인터넷은 발전하고, 인터넷의 발전 등의 같은 내용이 인터넷 발전 혹은 발전된 인터넷이라는 질의를 사용해서 찾을 수 없다. 이러한 4가지 표현이 하나의 사실(fact)를 나타내기 때문에 어떠한 표현이라도 같은 의미로 색인되어 찾아져야 할 것이다.
- 같은 의미를 가지는 문장이 서로 다른 표현방법으로 존재하고 있는 경우
  예를 들어 새로운 기술이라는 내용이 들어있는 문서를 첨단 기술, 차세대 기술, 신 기술이라는 질의어로 찾을려고 할 때 현재의 키워드기반 정보검색 시스템에서는 불가능하다. 이를 위해서는 의미적으로 같은 표현들은 의미표준화를 통하여 같은 의미로 취급되어져야 할 것이다.
- 같은 의미는 아닐지라도 의미적으로 매우 가까운 키워드를 가지는 경우
  예를 들어 의사라는 내용이 들어있는 문서

를 찾을 때 정확하게 그 의사라는 단어가 없거나 의미적으로 주위에 있는 단어들이 필요한 경우 간호사나 병원 등의 의미적으로 관련된 질의어를 사용해서도 그 문서를 찾을 수 있어야 한다.

4.5 색인어의 재 분석

문서와 색인어 추출과의 관계에 있어서 한가지 가정은 색인어 추출을 위해서 완전한 이해가 반드시 필요한 것은 아니라는 점이다. 색인어 추출은 근본적으로 어떤 어휘가 높은 가능성을 가지고 문서를 대표한다는 것을 입증하는 것으로 반드시 문서 전체적인 부분에 대하여 완전한 이해 및 전체에서 골고루 색인어가 추출될 필요는 없을 수도 있다. 일반적으로 두 사람이 잘 알려진 객체에 대해 같은 키워드를 선택할 확률은 20%가 되지 않는다는 사실이나, 일반 도서의 경우 저자가 추출한 색인어는 평균적으로 1페이지당 1.6 ~ 2.5개의 색인어라는 사실에서 근본적으로 문서에서 색인어를 추출하는 경우 문서에서 페이지 전체를 완전하게 이해하기 보다는 핵심적인 어휘의 추출과 페이지 당 수개의 색인어만을 추출하여 분석 실험하는 것도 고려 해 볼만하다.

4.6 문맥에 의한 색인어 선정

일반 문서의 의미는 문장 보다도 문맥(context)이 보다 더 중요하다. 문맥 분석은 그 분석의 심도와 관점에 따라 복잡성이 다르다. 가령 자연언어 이해의 관점이라면 그 어려움이 가장 심하다고 볼 수 있지만, 문맥의 단순한 구조 정도를 목표로 한다면, 비교적 쉽게 접근할 수 있다. 이상적이라면 문서의 내용을 최대한 이해해서 색인어를 가려낼 수 있어야 하지만 이는 현재의 기술 여건에는 비현실적일 뿐이다. 따라서 문서의 중심문단(centric paragraph)을 위주로 색인어를 추출한다거나 문서를 토픽(topic) 단위로 나누고(passage) 이 단위에 기반하여 색인어를 추출 색인 할 수 있을 것이다. 또한 문서가 갖는 구조적인 정보를 이용하여 그 구조에 기반 해 단위(passage)를 정하고 이를 기반한 색인어 추출을 고려해 볼 수 있다. 아울러 만일 문서를 요약하는 기술이 발달되었다면 이 요약된 정보를 바탕으로 색인어 추출을 하고 이를 실험 분석 할 수 있을 것이다.

5. 향후 동향

5.1 한글 정보 검색의 동향

- 다국어 정보 검색

웹과 인터넷 기술의 급속한 발전에 따라 여러 언어로 작성된 문서가 급격히 증가하고 이들 문서에 대한 검색 요구에 따라 다국어 정보 검색 관련 연구가 활발히 진행되고 있다. 다국어 정보 검색이란 질의어의 언어와 상관없이 여러 언어로 만들어진 정보를 검색하는 것이다. 현재 다국어 문서 검색은 단일언어 문서 집합에 대하여 여러 언어로 질의하여 검색하는 단계에 와있다. 하지만 향후에는 여러 언어로 구성된 문서 집합들에 대하여 여러 언어로 질의하는 것은 물론이고 하나의 문서가 여러 언어로 구성되어 있는 경우에도 검색하는 단계로 발전해 나갈 것으로 예상된다.

- 검색 결과 자동 요약

자동 요약이란 문서의 내용을 축약하는 것이라 할 수 있다. 즉, 본래 문서가 가지고 있는 의미를 유지하면서 문서의 길이나 정보의 복잡도를 줄이는 것이라 할 수 있다. 최근에 인터넷과 정보 서비스 기술의 발달로 인해 유통되는 정보의 양이 기하 급수적으로 늘어남에 따라 자동 요약의 필요성이 대두되고 있다. 인터넷의 정보 검색 엔진을 포함한 정보 검색 시스템에 질의를 하는 경우 검색된 문서의 수가 대량(1000개 이상)인 경우가 보통인데, 이런 경우 일일이 모든 문서를 검토한다는 것은 거의 불가능하다. 따라서 문서에 대한

자동 요약을 통해 검색 정보의 검토 시간을 줄이는 것이 필요하다.

● 자동 번역

웹 공간에는 다양한 언어로 된 다양한 내용의 문서가 산재해 있다. 정보의 바다에서 언어의 장벽을 뛰어넘게 도와주는 것이 웹 자동 번역이다. 자국의 질의를 통하여 외국의 검색된 정보를 바로 번역해서 보게 함으로서 전 세계의 정보를 쉽게 습득할 수 있게 도와준다. 현재의 언어간 번역 기술로는 특정한 언어에 대해 내용에 제한 없는 불특정 문서를 만족스러울 정도로 번역하기는 어렵다. 하지만 일어의 경우 웹 페이지 서비스를 번역 서비스함으로서 어느 정도의 내용 전달이 가능하다.

● 자동 분류

최근에 인터넷과 정보 서비스 기술의 발달로 인해 유통되는 정보의 양이 기하 급수적으로 늘어남에 따라 자동 분류의 필요성이 대두되고 있다. 인터넷에서 수 많은 웹 문서 가운데 필요로 하는 문서를 신속히 찾기 위해 웹 문서를 미리 분류할 수 있다면 정보 검색의 시간이나 정확도가 향상될 수 있을 것이다. 따라서 새로이 생성되는 문서의 자동 분류의 필요성이 대두되고 있다.

● 전거어, 동의어, 시소러스 사전 구축

전거어란 하나의 뜻을 나타내는 다른 형태의 용어들(이형 동의어)을 말한다. 전거 인명이나 지명등의 고유 명사의 쓰임이 사람마다 다르기 때문에 여러 가지 변형된 단어가 하나의 표준으로 갈 수 있도록 구성한 사전을 말한다. 예를 들면 표준 형태인 네덜란드의 경우 네델란드, 네델랜드, 네덜랜드 등의 다양한 변형을 가질 수 있다. 검색시 전거어는 색인어로 사용된 표준 형태로 바꾸어 준 후 색인이나 검색을 해야 정확하게 검색될 수 있다.

동의어는 미국(아메리카), 국제 축구 연맹 (FIFA), 크레디트 카드 (신용 카드) 등과 같이 일반적으로 사용되는 동의어가 존재하는 것을 말한다. 동의어는 (배, 선박), (식당, 음식점) 등 동의의 개념을 담고 있거나, (기차, 철도), (유명지, 관광지) 등의 관련어 개념을 가지고 있다.

시소러스는 색인 작업시에는 적절한 색인어의 선택과 색인어의 통제를 위해 필요하며, 검색시에는 적절한 탐색 용어의 선택을 지원한다. 그리고 시소러스는 용어 통제 뿐만이 아니라 검색어의 확장이나 축소를 통하여 검색 효율을 조절하는 데에도 사용된다. 즉 용어간의 계층 관계 및 연관 관계를 이용하여 포괄적인 검색을 하거나, 특정한 용어를 사용하여 보다 한정된 검색을 함으로써 검색 문서 수를 적절하게 조절할 수 있다.

● 온토로지(Ontology) 구축

온토로지란 원래 철학에서는 절대적 진리를 의미한다. 이러한 개념은 근래에 들어 지식기반 시스템을 만드는데 있어 응용분야마다 다른 지식을 구축하고 그러한 지식들은 다른 분야에 전혀 사용될 수 없기에 진리에 가까운 지식을 구축한다는 의미에서 사용하기도 한다.

정보 검색에서의 온토로지의 의미는 언어 종속적이지 않은(language independent) 단어들간의 관계 설정으로 정의되어 질 수 있다. 정보 검색에서도 의미와 다국어에 대한 관심이 고조되면서 이제까지의 어휘 중심의 정보 검색을 위해 시소러스가 필요하였듯이 온토로지에 대한 연구가 진행되고 있다. 그러나 아직 가장 작은 단위의 의미 혹은 개념에 대한 정의가 불분명하고 그 범위를 설정하기도 어렵거니와 그 개수가 많아서 실질적으로 정보 검색에서 사용될 수 있는 온토로지의 구축은 그리 쉬워 보이지는 않는다.

5.2 한국어 테스트 컬렉션

정보 검색에 대한 관심이 고조되고 정보 검색 시스템에 대한 개발이 국내의 기술로 이루어지고

있음에도 불구하고, 지금까지 정보 검색에 대한 연구 및 개발에 투자된 시간과 노력이 외국에 비하여 상대적으로 적었다. 지금까지 개발된 한글 테스트 컬렉션은 KT SET I(1,053개의 논문 초록들과, 30개의 질의), KT SET II(4,414개의 논문 초록 및 신문기사, 50 개의 질의), KRIST(13,515개의 연구보고서 초록과, 30개의 질의), ETRIKEMONG SET(23,113개의 백과사전 항목과, 46개의 자연언어 질의)이다. 향후 정보의 양이 보다 급속히 증가할 경우 현재 국내의 정보 검색 기술로서 외국의 시스템들과 경쟁하는 데는 많은 어려움이 예상된다. 따라서 대용량의 데이터를 대상으로 하는 정보 검색 기술에 대한 확보가 시급히 요구되며, 이를 위해서는 정보 검색 시스템의 성능들을 비교할 수 있는 평가 항목들에 대한 조사와, 정보 검색 시스템의 성능을 평가할 수 있는 테스트 컬렉션의 개발이 선행되어야 할 것이다.

6. 결 론

지금까지 간단하나마 한글과 색인어 추출 문제, 정확한 한글 색인어 추출 및 향후 동향에 대하여 다루었다. 현재 국내에도 한글 검색엔진을 비롯하여 자체 정보 검색 시스템들이 다수 개발되고 있으나 아직까지 한글 색인어를 정확하게 추출하여 주는 시스템은 드물다. 이는 가장 기본이라 할 수 있는 명사, 조사, 형용사, 동사 등, 정보 검색용의 한국어 사전이 여기저기서, 초보 수준에서 어느 정도 수준에 올라있는 형태로 만들어지고 있으나, 검증 및 공유라는 측면에서 아직 미흡한 면이 있다. 또한 정보 검색을 위한 전거어, 동의어, 시소러스, 온토로지 사전 구축이 막대한 시간과 비용이 드는 관계로 모든 응용 분야를 한 기관에서 자체 구축하기는 매우 어렵다. 따라서 이러한 구축 작업을 여러 기관에서 각 분야별로 공동 제작하여 공유할 수 있는 여건이 시급하다. 또한, 한국어 정보 검색 시스템들이 서비스 및 실험적인 수준에서 만들어지고 있으나, 서로 시스템을 비교 평가하고 서로의 장단점을 논하면서 한층 진보된 연구를 하는 마인드 부족과 정보 검색을 위한 한국어 표준 테스트 컬렉션(test collection)의 부족 때문에 검색 시스템들을 검증할 수 없는 문제점을 가지고 있다.

앞으로 정보 검색용의 형태소 분석, 구문 분석 및 의미 해석 및 진전된 연구들이 다양하게 이루어지고 서로 공유할 수 있는 사전 및 시소러스 등을 공동 제작하며, 한국어 정보 검색 시스템들을 표준 테스트 컬렉션에 의해 비교 및 평가 함으로써, 궁극적으로는 한국어 정보 검색 서비스를 이용하여 정보화 사회의 밑거름이 될 수 있기를 기대한다.

참고문헌

[1] 강승식, 권혁일, 김동렬, "한국어 자동 색인을 위한 형태소 분석 기능", 1995 봄 학술 발표회, 한국정보과학회, pp. 929~932, 1995.

[2] 강현규, 장호욱, 이승률, 박세영, "옥서에서의 표제어와 지연어 검색의 설계 및 구현", 1994 가을학술발표회, 한국정보과학회, 연세대학교, 서울, pp.633~636, 1994. (우수논문)

[3] 강현규, 박세영, 최기선, "자연언어 정보검색에서 상호정보망을 이용한 2단계 문서순위 결정 방법", 한국정보과학회 논문지, 제 23권, B편, 8호, pp. 852~861, 1996.

[4] 강현규, 왕지현, 김영섬, 서영훈, "정보검색에서 질의 형식화를 도와주는 개념마법사의 설계", 제 9회 한글 및 한국어 정보처리 학술대회, pp. 23-27, 1997.

[5] 김민정, 권혁철, "한국어 특성을 이용한 자동 색인 기법", 가을학술발표논문집, 한국정보과학회, pp. 1005-1008, 1992.

[6] 남기심, 고영근, 표준 국어문법론, 탑출판사, 1994.

[7] 장동현, 맹성현, "자동 요약 시스템", 한국정보과학회, 정보과학회지, pp. 42-49, 1997.

[8] 장명길, 김영길, 박영찬, "다국어 정보 검색", 한국정보과학회지, 제 16권, 제 8호, pp. 21-31, 1998.

[9] 박영찬, 최기선, "통계적 명사패턴 분류를 이용한 복합명사 검색 모델", 제 8회 한글 및 한국어 정보처리 학술발표 논문집, 1996.

[10] 박혁로, "효율적인 정보 검색을 위한 복합명사 처리", 과학기술정보워크샵(KOSTI96), pp. 429-440, 1996.

[11] 신중호, 박혁로, 이강혁, "과학기술문헌 자동색인을 위한 어절분석 시스템", 과학기술정보워크샵(KOSTI96), pp. 415-428, 1996.

[12] 이창열, 강현규, 장호욱, 박세영, "자동 키워드 제작기 시스템 설계", 제 5회 한글 및 한국어 정보처리 학술발표 논문집, pp. 71-77, 1993.

[13] 조영환, 박혁로, 이준호, "정보검색 시스템 평가 및 테스트 컬렉션 개발", 한국정보과학회지, 제 16권, 제 8호, pp. 48-55, 1998.

[14] 최기선, "한국어 정보 검색", 한국정보과학회지, 제 12권, 제 8호, pp. 24-32, 1994.

[15] 한국전자통신연구원, 내용기반 멀티미디어 정보검색 기술 개발, 보고서, 정보통신부, 1997.

[16] 한국전자통신연구원, ETRIKEMONG SET, 자연어처리연구실, 한국전자통신연구원, 1997.

[17] G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company, 1988.

강 현 규

1985년 홍익대학교 전자계산학과 (학사)
1987년 한국과학기술원 전산학과 (석사)
1992년 정보처리 기술사
1997년 한국과학기술원 전산학과 (박사)
1987년-현재 한국전자통신연구원 선임연구원
관심분야 : 정보 검색, 자연언어 처리, 멀티미디어 처리, 디지털 라이브러리

박 세 영

1980년 경북대학교 전자공학과 (학사)
1982년 한국과학기술원 전산학과 (석사)
1989년 프랑스 파리 7대학 (박사)
1995년~1996년 미국 미주리 주립대 객원연구원
1982년~현재 한국전자통신연구원 책임연구원, 자연어처리연구부장
관심분야 : 자연언어 처리, 정보 검색, 멀티미디어 처리, 디지털 라이브러리