정보검색 기술의 현황과 발전방향

한국정보통신대학교 맹성현*

1. 서 론

정보검색 기술은 거의 40년 전부터 미국 코넬대학의 Salton과 영국 케임브리지 대학의 van Rijsjbergen 및 Sparck Jones 등 초창기 선구자들을 중심으로 기초가 세워지고 ACM에서는 약 30년 전에 SIG(Special Interest Group)로 인정 받은 후 독립된 학문 분야로 꾸준히 발전되어 왔다. 1968년에 출간된 Salton의 책 [1]에서는 정보검색을 "정보의 구조, 분석, 조직, 저장, 탐색(searching), 검색(retrieval)에 관한 분야"로 정의하고 데이터베이스 시스템과 질의응답 시스템도 포함시키고 있다.

인터넷과 더불어 디지털화된 텍스트의 양이 급증하고 정보 형태 및 정보에 대한 요구 형태가 다양해지면서 1990년에 들어 정보검색에 관한 관심은 폭발적으로 증가하였고, 웹 서치 엔진이 등장하면서 일반 대중에게도 널리 알려지게 되어 ***독립된 검색 엔진으로서 혹은 지식 관리 시스템이나 스팸메일 필터와 같은 다양한 지식정보 시스템의 요소 기술로서 매우 보편화되어 있다.***

국내에서 정보검색 기술에 대한 연구 개발이 본격적으로 시작된 것은 1990년대 초라고 할 수 있는데, 정보검색이 텍스트 검색 혹은 문서 검색으로 간주되는 상황에서 바로 해결되어야 할 문제는 한국어처리와 관련된 문제이었기 때문에 문서의 색인을 위한 형태소 해석 기술의 개발에 많은 시간과 노력이 집중되었다. 이후 웹 검색 엔진의 상업화 바람에 힘입어 색인, 저장, 검색 결과 랭킹과 같은 기초 기술 및 엔지니어링 기술에 급격한 발전이 이루어졌으나, 상업화에 치중한 결과 첨단 기술 개발은 등한시되었고 연구활동이 언어처리 혹은 데이터베이스 기술의 한 영역으로서 수행되어 집중적인 핵심 첨단 기술의 개발이 미흡한 실정이다. 다만 몇몇 연구자들이 세계와 어깨를 겨룰 수 있는 기술을 발표하고 있는 것은 집중적인 투자가 이루어지는 경우 이 분야의 핵심에서 선도적인 역할도 할 수 있다는 가능성을 보여 주고 있다.

본 논문의 목적은 이러한 시점에서 정보검색 기술에 관심을 갖는 연구자들이 이 분야를 이해하고 발전방향을 예측하여 국내에서의 정보검색 기술 개발 및 활용이 보다 체계적으로 이루어지도록 일조하는데 있다. 기술의 발전방향을 예측하는데 있어서의 객관성을 최대한 유지하기 위해 2002년 말에 미국 ARDA의 후원으로 매사추세츠 대학에서 열린 미래 정보검색 기술개발에 대한 워크샵의 토의 내용[2]을 반영하였다.

2. 정보검색 분야 별 기술 현황

2.1 검색 모델

검색 모델은 검색 프로세스 및 여기에 관여하는 엔티티(문서 내용, 질의, 사용자 등)를 어떻게 표현하는가를 결정하므로, 검색기의 기능 및 성능에 직접적인 영향을 준다. 즉 사용자의 질의가 찾는 정보를 개략적으로 표현하고 있고 문서의 내용도 색인 형태로 개략적으로 표현될 수 밖에 없으므로 검색 프로세스는 원천적으로 불확실성을 내재하고 있는데, 이를 고려한 검색 모델이 필요한 것이다. 정보검색 모델은 찾아야 할 정보를 명확히 기술할 수 있는 질의어를 가지고 있는 정형데이터 검색용 데이터베이스 분야와의 차별성을 분명하게 보여 준다.

초기 시스템의 기반이 되었고 현재 웹 서치 엔진에서도 반영되고 있는 불리언 모델은 질의 용어를 포함하고 있는 문서의 집합을 구성한 후 불리언 연산자를 사용하여 다수의 집합을 통합한 후 최종 검색 결과를 결정하는 모델이다. 그 후 1970년 대에 소개되어 꾸준히 발전되어 온 벡터공간 모델은 현재까지도 많은 시스템에서 사용되고 있는 모델로 색인어로 이루어진 좌표 공간에서 문서와 질의를 벡터로 표현한 후 그 벡터 간의 유사도에 의해 검색 결과를 결정하고 랭킹한다.

검색 모델에 있어서 이론적으로 가장 튼튼한 기반을 가지고 있는 것은 확률 모델이라 할 수 있다. 주어진 질의에 대해 특정 문서가 적절할 확률을 구해서 문서를 랭

* 종신회원

킹하는 방법에 기초를 둔 이 모델은 오랜 동안 연구가 되어 왔는데, 모델에 출현하는 확률값 및 파라메터를 어떻게 추정할 것인가에 따라 그 성능이 달라진다. 1990년도 초반에 개발된 추론망 기반 모델[3]과 1990년도 중반에 개발된 포아송 분포 기반 랭킹 모델[4]은 모두 TREC(Text Retrieval Conference)의 평가대회에서 매우 우수한 성능을 보여 주었다. 또한 최근에 많은 연구가 이루어지고 있는 언어 모델(Language Model)[5]도 문서와 컬렉션의 통계량을 체계적인 방법으로 반영하는 확률 기반 모델의 하나로서 다양한 데이터 소스나 검색 기능을 포함해야 하는 미래의 검색 시스템 개발에 가장 적절한 모델로 각광을 받고 있다.

이 외에도 퍼지(fuzzy)집합 기반 모델[6], p-norm 모델 [7], 개념그래프 (conceptual graph) 기반 모델[8], 논리 기반 모델[9], 잠재의미색인 모델[10] 등 검색 대상 및 목적에 따라 다양한 검색 모델이 제시되었다. 정보검색 프로세스 및 엔티티를 보는 관점에 따라 제안된 이 모델들은 나름대로의 장점을 가지고 있는데, 향후에는 현재보다 훨씬 다양한 정보 형태 및 검색 태스크를 만족시키기 위한 모델의 개발이 필요하게 될 것이다. 체계적인 검색 모델 없이 필요한 기능이 주먹구구식으로 기존 시스템에 반영된다면 그 성능을 예측할 수 없을 뿐만 아니라 확장 및 성능 향상에도 한계를 가지게 되는 것은 자명한 일이다.

2.2 색인 및 저장

정보검색에서 다루는 데이터는 모두 비정형성을 가지고 있다. 즉 DBMS에서와 같이 테이블 형태 등으로 구조화되어 있는 데이터를 대상으로 검색하는 것이 아니기 때문에, 문서나 이미지와 같은 비정형 데이터로부터 그 내용을 대변하는 색인어를 추출하여 효율적인 검색이 이루어지도록 해야 한다. 특정 정보항목을 대변하는 색인어 집합은 그 정보항목의 내용을 충실히 반영하면서 동시에 다른 항목과의 차별화가 되도록 선정되어야 한다.

텍스트를 색인하는 경우 현재 정보검색 기술의 수준에서는 대개 단어 혹은 구(phrase)를 사용하고 이들을 총칭하여 용어(term)라고 부른다. 로마자 기반 언어권에서의 용어 자동추출은 비교적 수월하여 단어를 단위로 할 때의 기술은 대부분 이미 안정화되어 있다. 다만 구를 추출하거나 단어의 중의성 해소를 통해 의미단위 색인을 하는 기술의 개발은 아직도 진행 중이다. 반면에 한국어, 중국어와 같은 비영어권에서는 텍스트를 분석하여 형태소와 같은 의미를 가진 최소단위를 추출하는 작업이 쉽지 않아 아직도 기술 개선의 여지가 많이 남아 있고 자연언어처리 기술이 가장 많이 공헌하고 있는 분야이다.

색인어 추출 외에 고려할 사항은 추출된 색인어의 가중치 계산 및 저장 기술이다. 가중치 계산은 검색 결과 랭킹 즉 검색 효과에 직접적인 영향을 주므로 지속적인 연구가 진행 중이다. 대부분 용어 빈도(TF: term frequency)와 역문헌 빈도(IDF: inverse document frequency)와 같이 용어 관련 빈도에 의존하고 있으나, 두개의 포아송 분포를 사용하는 방법 등 새로운 통계량을 사용하는 방법과 특정 용어가 가지고 있는 구문적 역할을 비롯한 언어 지식을 활용하는 기술도 개발되고 있다.

색인어 저장 기술은 검색의 효율성을 고려하여 역색인(inverted index) 구조가 가장 보편적으로 사용되고 있다. 웹 검색의 경우와 같이 대용량 자료를 색인해야 하는 경우가 많아지면서 역색인 구성 과정과 질의에 대한 검색을 효율적으로 수행하는 기술이 상용 시스템 개발 과정에서 중요한 자리를 차지하게 되었다. 역색인 구성시 파일 입출력을 최소화하기 위한 기술 혹은 역색인 검색시 디스크 접근 회수를 최소화하기 위해 압축/복원 방법 등은 특히 상용 시스템이나 대용량 실험을 위한 시스템 개발에 매우 유용한 기술이다.

2.3 사용자 시스템 상호작용

웹 서치 엔진을 비롯해서 모든 정보검색 시스템의 검색 결과에 원하지 않는 문서가 포함되어 있는 근본원인은 크게 두 가지로 볼 수 있다. 첫 번째는 색인이 원천적으로 문서 내용을 그대로 대변할 수 없다는 것이고, 두 번째는 시스템이 처리해야 할 질의가 사용자의 정보 요구를 제대로 표현하지 못하고 있기 때문이다. 이렇게 불충분하고 불확실한 정보를 어떻게 연결시켜야 하는가가 매칭 혹은 유사도 계산의 문제이다.

현재의 기술 수준에서 이러한 상황을 극복하는 방법은 정보검색의 문제를 단순히 컴퓨터 시스템이 모두 해결해야 하는 것으로 보지 말고, 컴퓨터 시스템과 사용자의 협업을 통해 해결해야 하는 것으로 보는 것이다. 즉 사용자가 시스템에 자신의 정보 요구를 전달하는 채널이 제한되어 있는 상태이기 때문에 상호작용을 통해서 점증적으로 시스템에게 원하는 바를 전달할 수 있도록 해야 한다.

관련된 대표적인 기술로 적합성 피드백(relevance feedback)을 들 수 있다. 이 기술은 사용자가 자신의 초기 질의에 대한 검색 결과를 보고 각 문서에 대한 적합성 여부를 표시하면 시스템이 초기 질의를 수정하여 다시 검색을 수행함으로써 적절한 문서의 수를 증가시키고 부적절한 문서의 수를 줄이도록 한다. 여기서 중요한 것은 적절한 문서와 부적절한 문서의 내용을 어떻게 분석하여 초기 질의를 어떻게 수정하도록 하는가이며 Roc-

chio의 방법[11]이 대표적이다. 사용자의 직접적인 적합성 피드백 정보 없이 질의를 확장하는 방법으로 의사 피드백(Pseudo-feedback) 기술이 개발되었다. 여기서는 검색 결과 중 순위가 높은 문서는 일반적으로 적합하다는 가정하에 문단으로 나눈 상태의 검색 결과에서 상위 문단에 출현하는 개념어의 가중치를 계산하여 차기 질의에 반영하는데, 이때 지역정보와 전역정보를 동시에 사용하여 계산한다[12].

이 외에 사용자 프로파일을 이용하여 사용자의 정보 요구를 간접적으로 파악함으로써 검색 효과를 향상시키는 방법이 있다. 사용자 혹은 사용자 그룹이 가지고 있는 관심사를 용어 형태로 저장한 후 시스템에 질의가 입력되었을 때 이를 관심사에 입각하여 해석함으로써 검색 효율을 향상시킬 수 있다[13]. 이 기술은 현재 웹 검색에 원용되고 있는데 실제 입력된 일반 사용자의 질의를 일정 기간 수집하여 분석을 한 후 성향을 파악하여 인기 있는 검색어를 추천한다. 개별 사용자가 아닌 전체 사용자 그룹을 모델링 한다는 면에서 매우 단순한 방법이지만, 나아가서는 특정 검색어(질의)에 대해 검색된 문서 중 어느 문서를 사용자가 클릭하였는지를 통계적으로 분석하여 랭킹을 하는 기술도 개발이 될 것이다.

2.4 웹 검색

웹 서치 엔진 기술 및 디렉토리 기반 정보 서비스 포탈은 정보검색 분야를 대중화시킨 요인이 되었고, 일반인을 위한 대용량 정보검색 기술을 발전시키는 주요 동인이 되었다. 웹 서치 엔진의 경우 매우 짧은 질의에 비해 상대적으로 많은 양의 정보를 대상으로 검색을 해야 하는 어려움이 있지만 일반 문서에는 존재하지 않은 링크 및 앵커 그리고 추가적인 구조 정보가 있다는 특성을 반영한 새로운 기술이 개발되고 있다. 특히 하이퍼링크 정보를 이용하는 페이지랭크 기술은 상업적인 성공을 가져다 주었다.

웹 검색과 관련된 새로운 기술 영역으로 크롤링(crawling) 기술을 들 수 있다. 분산되어 있는 무수히 많은 컴퓨터에 저장되어 있는 문서를 수집하여 색인하는 이 기술은 어느 부류의 문서를 얼마나 빨리 수집하여 검색 대상에 포함시키는가에 우위가 결정된다. 이 기술을 상업적인 면에서 볼 때 초대용량의 문서를 수집, 저장, 서비스하는데 있어 발생하는 엔지니어링 문제를 어떻게 해결하는가가 중요하다.

웹 검색에 국한되는 기술은 아니지만, 검색 결과 너무 많은 문서가 사용자에게 반환되기 때문에 이를 정리해서 결과를 제시함으로써 정보의 과부하를 줄이고 사용자가 원하는 문서에 신속하게 접근할 수 있도록 하는 기술이 각광을 받고 있다. 일반적으로 클러스터링 기술을 사용할 수 있으나 속도를 개선함과 동시에 정확성을 유지해야 하는 어려움이 있다[14].

웹 검색의 일환으로 개발된 메타 검색(meta-searching)은 사용자 질의를 웹 서치 엔진으로 보낸 후 검색된 결과를 받아 통합하여 사용자에게 제시하는 기술이다. 검색 서비스를 제공하는데 있어 웹 크롤러나 초대용량 문서에 대한 색인을 구축할 필요가 없다는 장점과 효과적인 통합이 이루어질 경우 다양한 원천으로부터 검색될 결과를 활용한다는 면이 매력적이긴 하나, 기존 시스템의 협력이 없이는 운영이 불가능하다는 한계를 가지고 있다. 그러나 다양한 알고리즘으로 생성된 랭킹과 서로 다른 문서집합을 통합하여 양질의 단일 결과를 생성하는 것은 기술적으로 중요한 문제이며 아직 해결해야 할 과제로 남아 있다.

2.5 자동 분류 및 필터링

텍스트나 멀티미디어 객체를 이미 정의된 범주로 분류하는 기술은 다양한 분야에서 그 응용을 찾을 수 있다. 예를 들어 전자 메일을 스팸 메일과 그렇지 않은 메일로 분류하는 기능을 들 수 있는데, 이 응용은 동적으로 유입되는 문서를 사용자가 원하는 것만 선별해 주는 좀 더 일반적인 필터링 문제와 동일하다. 자동 분류의 경우 범주의 개수가 매우 많을 수도 있고 웹 포탈의 디렉터리 서비스에서 사용되는 것 같이 범주가 계층적으로 분류되어 있는 경우도 있다. 지난 10년간 많은 연구가 진행되어 상용화가 가능한 기술이 되었다[15].

자동 분류 기술은 대부분 지도학습(supervised learning) 방법을 사용한다. 즉 학습단계에서 이미 범주가 결정된 정보 항목들로부터 각 범주 별 특성을 표현하는 자질(feature)을 자동적으로 추출한 후 새로운 항목이 어느 범주에 속하는 가를 결정하는 기계학습 방법을 사용한다. 이 방법은 사용자가 직접 각 범주 별 규칙을 생성해야 하는 어려움을 덜어 준다. 가장 많이 연구되고 또 현재 사용되고 있는 방법으로 단순 베이지언 (Naïve Bayesian) 분류기, k-NN, 결정트리(Decision Tree), 지지벡터기계(Support Vector Machine) 모델, 벡터 기반 유사도 계산 방법 등이 있는데 이들에 대한 비교평가에서 지지벡터 모델이 일반적인 우위를 차지하고 있는 것으로 알려져 있다[16]. 그러나 학습 데이터의 종류나 규모 그리고 범주의 특성에 따라 각각 장단점이 있어 지속적인 연구가 요구된다.

자동 분류 기술을 실제 문제에 적용하는데 있어서의 최대 걸림돌은 범주가 이미 결정된 학습 데이터를 수집하는 것이다. 즉 상당한 양의 정보 항목에 대한 범주를

수작업으로 할당해야만 분류기가 작동될 수 있다. 이 문제를 극복하기 위해 클러스터링 방법 등을 이용해 범주가 정해지지 않은 데이터를 사용하는 방법, 의사피드백(pseudo-feedback) 방법, 동시학습(co-training) 등의 기술이 개발되고 있다.

분류의 대상으로 웹 문서가 중요한 자리를 차지하면서 새로운 자동 분류 기술이 개발되고 있다. 웹 문서가 갖는 특징은 하이퍼링크와 같은 일반 문서에 존재하지 않는 추가적인 정보가 존재한다는 것이다. 분류에 있어 주어진 페이지와 하이퍼링크로 연결된 주변 페이지를 활용하는 기술[17]은 하이퍼텍스트에 존재하는 새로운 정보를 이용한 예이며 웹 문서를 특정 페이지와 그와 연결된 페이지를 통틀어 하나의 문서로 간주하는 새로운 관점을 보여 준다. 한편 웹 페이지의 경우 매우 뚜렷한 목적을 가지고 생성이 되는데 이러한 목적을 범주로 하는 장르 범주 기반 분류 기술도 개발되어[18], 문서의 내용뿐만 아니라 새로운 기준에 의한 문서 분류의 가능성을 보여주고 있다. 정보접근의 경로를 다변화 시킴으로써 궁극적으로는 검색의 효과를 높일 수 있다는 면에서 다양한 차원에서의 분류 기술이 개발되어야 할 것이다.

2.6 주제 탐지 및 추적

주제 탐지/추적(TDT: Topic Detection and Tracking)은 주어진 사건과 연관된 기사를 추적하고 새로운 사건에 대한 기사를 탐지하는 목적으로 최근에 정보검색의 새로운 응용 분야로 부상하였는데[19], 기술적으로 보면 자동 분류 및 클러스터링의 한 지류라고 할 수 있다. 사건 기사의 특성은 일반 문서와 비교하여 특정 시간 및 장소가 명시되어 있는데 이러한 특성을 자동 분류 및 클러스터링 기법에 적용시켜 추적 기술이 개발되었다.

추적 기술이 일반 자동 분류나 필터링 기술과 다른 점은 크게 두 가지이다. 첫째, 두 기사의 내용이 유사하더라도 동일 사건에 대한 보도가 아닐 수 있다는 것과, 둘째, 추적을 위해 사용되는 단서 기사의 수가 매우 적어서 (시작 시 한 건) 학습 기반 방법을 적용하기 어렵다는 것이다. 전자의 경우 각 기사에 존재하는 시간과 장소 정보를 사용하여 이 문제를 극복하는 기술이 개발되고 있고, 후자의 경우 단서 문서와 과거 기사 전체와 비교하여 차별화되는 자질을 추출하여 사용하는 방법을 쓰고 있다.

최초 기사 탐지(First Story Detection)는 과거에 보도되지 않은 새로운 사건에 대한 기사를 탐지하는 문제이다. 이 기술은 새로운 사건의 발발을 최대한 신속히 탐지해야 하는 보안이나 경제 분야에 매우 유용하게 사용될 수 있다. 그러나 이 기술은 과거 모든 사건과 비교하여 새로운 사건이라는 것을 판단해야 하고, 때로는 이미 보도된 사건일지라도 특정 사용자에게 새로운 사건인 경우도 있어, 고난도의 기술을 요구한다.

2.7 자동 요약

자동 요약은 주어진 문서 혹은 문서집합으로부터 핵심 내용을 정리하여 사용자에게 제시함으로써 정보의 과부하를 방지하고 보다 효율적인 정보 습득을 가능하게 한다. 자동 요약 기술의 목표는 주어진 문서의 내용이 무엇인지 판단하는 정도의 요약, 주어진 문서에 포함된 내용을 최대한 충실히 담아 내는 요약, 주어진 문서의 내용에 대한 평가의 생성 등 다양하게 설정될 수 있고, 이 목표에 따라 적절한 기반 기술이 적용되어야 한다[20].

현재 상용화 수준에 와 있는 기술은 대부분 핵심 키워드 추출이나 핵심 문장을 추출하여 제시하는 방법으로, 핵심 문장을 예고하는 문장성분을 활용하는 방법[21, 22], 핵심 주제어 혹은 주제 문장을 찾아낸 후 이와 연결된 문장을 추출하는 방법[23] 등이 사용되고 있다. 추출된 문장의 자연스러운 연결을 위해 대용어 처리 등 자연언어처리 기술의 접목이 필요하다.

문장 추출 수준의 자동 요약에는 한계가 있으므로 보다 심층적인 문장 분석을 통해 인간이 작성하는 수준의 요약을 생성하는 기술의 개발이 필요하다. 개별 문장이 가지고 있는 의미를 표현하고 통합하여 문서 전체의 의미를 지식표현 언어로 표현한 후 핵심 내용이 요약된 문장을 생성하는 단계를 거쳐야 하는데[24], 특정 영역에 한정된 응용이 가능하나 자연언어처리 기술의 발전이 선행되어야 하고 보다 정보검색 분야와 자연언어처리 분야의 협력이 필요한 분야이다.

2.8 질의응답

현재 연구 개발이 이루어지고 있는 질의응답 시스템은 자연언어로 기술된 질문을 받아 맥락 정보가 포함된 사실(fact)을 답으로 제공하는 것이 주류를 이룬다. 이는 과거 인공지능 분야에서 다루어 왔던 시스템과는 달리 지식기반 추론과정 없이 주어진 텍스트 자원으로부터 필요한 답을 식별하고 필요하면 여러 정보 자원으로부터의 답을 통합하는 과정을 거친다. 예를 들어 "백두산의 높이는?" 라는 질문이 들어 왔을 때, 그 답을 텍스트에서 추출하거나 그 답이 들어 있는 일정 크기의 텍스트를 제공한다.

일반적인 접근 방법은 자연언어처리 기술을 적용하여 질의의 유형을 분류하고, 질의에 포함된 키워드를 추출한 후, 그 키워드를 사용하여 정보 자원으로부터 답을 포함하고 있을 만한 문서 혹은 문단을 먼저 검색한다.

분류된 질의 유형에 따라 정답의 형태를 결정하고 이런 형태의 정답이 존재할 만한 문장 및 문서를 패턴매칭에 의해 찾아나간다. 여기서 문제의 유형과 답안의 유형에 대한 시소러스나 지식베이스를 사용하게 되는데 이들의 완성도와 품질이 답안의 정확도를 좌우한다. 또한 질의 및 문장의 부분에 대한 자연언어처리 기술이 시스템 성능에 많은 영향을 미친다.

2.9 교차언어 검색

교차언어 검색(Cross Language IR)은 질의를 표현한 언어와 문서에 사용된 언어가 다른 경우의 검색을 지칭한다. 흔히 자국어로 질의를 표현하여 외국어로 쓰여진 문서를 찾거나 여러 가지 언어로 기술된 문서를 찾는 경우 사용되는데, 후자의 경우는 특히 다중언어 검색(Multilingual IR)이라 한다.

교차언어 검색에서 흔히 사용되는 방법은 사용자 질의를 문서에 사용된 언어(목적 언어)로 변환을 하는 것이다. 이를 위해 사용되는 기술은 대역어 사전을 통해 원시 질의어에 나타난 단어 혹은 구를 목적언어로 번역을 하는 것이 가장 보편화된 방법인데, 원시 단어 및 구의 의미상의 애매성을 분별하여 적절한 번역어를 찾는 기술과 번역된 단어 및 구에 가중치를 계산하는 기술이 개발되었다[25,26].

사전이 존재하지 않거나 사전이 불충분한 경우, 병행 코퍼스(parallel corpus)나 비교 코퍼스(compatible corpus)를 사용하여 코퍼스에서 대역어를 찾아내거나 직접적인 대역어는 아니더라도 원시 질의를 대변하는 질의를 생성하는 기술도 개발되었다. 여러 가지 통계량을 사용하거나 기계번역 분야에서 개발된 문장 배열(alignment) 방법을 적용하기도 하고, 잠재의미색인(latent semantic indexing)에서 사용된 방법을 적용하여 원시 언어 및 목적 언어가 공존하는 공간을 구성하여 원시 질의 단어와 의미적으로 가까운 목적 질의 단어를 선택하는 방법[26]도 개발되었다.

교차언어 검색 기술의 완성도는 동일한 질의에 대하여 단일언어 검색 결과의 성능과 비교하여 몇 퍼센트까지 도달하는가의 척도를 사용하는데, 그 동안 많이 연구된 언어 쌍에 대해서는 단일언어 검색 성능과 유사한 결과를 보이고 있다. 다만 새로운 언어 쌍에 대한 기술 개발이 필요하며, 오히려 교차언어 검색에 사용되는 자원을 이용하여 단일언어 검색의 효과를 향상시키는 방향으로 기술 개발이 진행되고 있다. 다중언어 검색의 경우는 다양한 언어로 쓰여진 문서가 각각 검색되었을 때 이들을 효과적으로 통합하는 문제가 생기는데 이에 대한 지속적인 연구가 필요하다.

교차언어 검색은 스위스, 캐나다, 싱가폴 등 다국어를 사용하는 국가나 다국적 기업 등에서 필요성이 많이 대두되었고, 또 인터넷에 존재하는 문서의 절반 이상이 영어가 아닌 언어로 작성되어 있어 이를 활용하기 위한 기술로 발전이 되었다. 비록 연구 관점에서 단일언어 수준의 검색 결과를 생성할 수 있다고는 하나, 이 기술이 실제로 사용되기 위해서는 사용자와의 상호작용에 다양한 장치가 필요하게 된다. 예를 들어 검색된 결과 문서의 내용을 자국어로 요약하여 제시함으로써 완전 번역 과정을 거칠 문서를 선별하게 해 주거나, 질의 생성시 혹은 검색 과정에서 생기는 피드백을 자국어로 지원하는 장치 등은 교차검색 시스템의 실용화에 필수적일 것이다.

2.10 분산 검색

원천(source) 데이터가 분산되어 있는 상황에서의 검색은 데이터가 중앙에 집중되어 있는 경우에 고려할 필요가 없는 다른 문제를 안고 있다. 분산 검색 상황을 크게 보면 두 가지로 대별할 수 있다. 웹에서의 메타 검색과 같이 동일 데이터를 대상으로 분산 검색한 후 결과를 통합하는 경우가 있는데 이를 데이터 퓨젼(fusion)이라 하고, 분산 저장소에 서로 다른 데이터가 있어 사용자 질의 처리시 검색 데이터 소스의 선택, 질의 변환, 검색 결과 통합을 수행해야 하는 경우를 컬렉션(collection) 퓨젼이라 한다.

데이터 퓨젼에서는 일반적으로 각각의 소스에서 검색된 결과에 적합성 점수가 존재하므로 각 소스로부터의 적합성 점수 범위를 정규화(normalize)하여 타 소스로부터의 검색 결과와 비교가 가능하게 한 후 랭킹을 통합하는 방법을 사용한다. 컬렉션 퓨전에 있어서의 결과 통합은 데이터 원천이 다르기 때문에 각 컬렉션마다 통계량과 검색 방법이 상이하다. 따라서 결과 통합에 있어서 점수 범위를 정규화하는 방법보다는 각 컬렉션이 질의에 적절한 문서를 포함할 확률 값을 계산하여 그 확률에 따라 편향된 라운드 로빈 형식을 사용하기도 한다[27].

컬렉션 퓨전의 경우 질의를 보낼 컬렉션을 선정해야 하고 각 컬렉션에서 검색이 된 후 결과를 통합해야 한다. 일반적으로 각 검색기에서 허용하는 컬렉션 관련 정보가 제한되어 있기 때문에 외부에서 관찰 가능한 정보를 사용하여야 하는데, 각 용어가 각 컬렉션에서 갖는 문헌 빈도수와 각 용어의 컬렉션 빈도수 즉 특정 용어가 출현하는 컬렉션의 수를 활용하여 검색 대상 컬렉션을 랭킹하는 방법[28]이 있다.

분산 검색 결과의 정확성 측면 이외에 실용 시스템에서 해결되어야 할 문제는 검색 속도이다. 다양한 하부 검색 시스템마다의 속도가 근본적으로 다르기도 하고 질

의에 따라 혹은 시간대에 따라 사용자가 느끼는 검색 속도가 변하기 때문에 통합되어야 할 검색기의 개수가 많은 경우 통합 검색기의 속도는 항상 가장 느린 검색기보다 늦어지게 된다. 이를 해결하는 방법으로 시간제한을 둘 수 있으나 이 경우 특정 검색기의 결과를 수용 못하는 결과를 초래해 같은 질의에 대해 일정하지 못한 결과를 사용자에게 제공하게 된다.

이외에 엔지니어링 문제가 존재하는데, 예를 들면 각 검색기의 질의 인터페이스나 검색 결과 인터페이스가 변하는 경우 통합기도 수정되어야 한다. 이러한 상황에서 수작업을 최소화하고 동시에 새로운 검색기를 통합할 경우의 수작업도 최소화하는 방법으로 프로토콜에 기반한 분산 통합기 기술도 개발되었다[29].

2.11 멀티미디어 검색

정보검색 기술의 발전이 대부분 텍스트 검색에 집중되었으나, 최근에 멀티미디어 컨텐츠의 생성 및 활용이 급증하면서 멀티미디어 색인, 검색, 가공 기술에 대한 관심이 높아지고 있다. 색인/검색 대상이 되는 멀티미디어 객체는 소리(예: 음악, 음성), 이미지(예, 사진, 클립아트, 스캐닝된 문서), 비디오(예: 디지털 TV, 보안 카메라 출력, DVD) 등이 있는데 종류에 따라 이들로부터 색인을 생성하고 검색하는 기술의 종류와 수준이 매우 다양하다.

멀티미디어 검색을 위한 접근 방법은 크게 네 가지 경우로 나누어 볼 수 있다.

1) 비디오 자료의 캡션과 같이 멀티미디어 객체에 연관된 텍스트가 이미 존재할 경우 사용하는 접근 방법
2) 음성 인식이나 문자 인식 등의 기술을 이용하여 객체의 일부를 텍스트로 변환할 수 있는 경우의 접근 방법
3) 멀티미디어 객체에 대한 메타 데이터가 수작업으로 생성되어 있는 경우의 접근 방법
4) 자동적으로 멀티미디어 객체에 대한 자질을 추출할 수 있는 경우의 접근 방법

1)의 경우 텍스트를 기반으로 색인 생성이 가능하므로 텍스트 기반 정보검색 기술을 그대로 적용할 수 있으나, 다른 자질의 추출이 가능할 때 다중 모달리티(modality) 하에서의 통합검색 기술의 개발이 효과적이다. 2)의 경우 인식 결과에 오류가 항상 존재하므로 텍스트의 경우보다 색인의 정확도가 떨어질 수 밖에 없다. 따라서 이런 오류에 강건한 검색 모델의 개발이 중요하다. 3)의 경우 메타 데이터에 의한 검색은 정형 데이터 및 비정형 데이터 검색과 동일하므로 기존의 기술을 적용할 수 있지만, 메타 데이터의 생성 시 모든 사용자의 정보 요구를 만족시키는 메타 데이터 스키마를 구축할 수 없다는 한계를 갖는다. 4)의 경우 자질의 자동 추출은 멀티미디어 종류를 불문하고 고난도의 기술을 요구한다. 예를 들어 이미지로부터의 자질 추출은 색상 히스토그램(color histogram), 객체의 윤곽(shape), 질감(texture) 정도가 가능하므로[30] 주어진 이미지와 유사한 이미지를 검색하는데 있어 이들 자질밖에 사용할 수가 없다. 즉 이미지가 가지고 있는 객체간의 관계나 다양한 관점에서의 의미 등을 추출하는 기술은 아직 매우 미흡한 실정이다. 자동으로 추출된 자질은 오류를 포함하고 있거나 검색 목적에 충분히 부합되지 않는 경우가 많으므로 이 방법은 특정 도메인이나 응용에 국한시켜 사용되는 것이 대부분이다.

3. 전반적 발전방향

3.1 기술 발전

위에서 정보검색의 세부기술 별 현황을 기술하면서 현재 개발되고 있거나 가까운 미래에 해결되어야 할 기술적인 문제는 각각 언급 하였으므로, 여기서는 정보검색 분야를 총체적으로 볼 때의 중장기적 발전방향을 기술한다.

3.1.1 의미기반 상세정보 검색

현재까지 정보검색의 발전은 단어에 대한 통계적인 모델을 기반으로 이루어졌다. 텍스트, 문장, 문장 구성 성분이 가지고 있는 의미가 명시적으로 분석, 표현되어 사용되기 보다는 통계적인 방법으로 동일한 효과를 얻어내는 성과를 이루었다. 미래에는 자연언어처리 기술이 보다 적극적으로 적용되어 양질의 색인어를 추출하는 방법에만 그치지 않고, 텍스트 및 문장의 구성 요소를 분석하여 문서의 내용을 단어 집합보다 풍부한 의미를 내포하는 형식으로 표현함으로써 보다 정교한 검색이 수행될 수 있도록 하여야 한다. 위에서 언급한 자동 분류, 질의응답 혹은 정보 추출과 같은 텍스트 마이닝 기술이 동원되어 접목되는 경우 상세 정보검색의 실현을 보다 앞당길 수 있을 것이다.

3.1.2 시맨틱 웹 서비스 기술과의 접목

웹 서비스는 웹 상에서 다양한 기능을 가진 서비스를 표준적인 방법으로 정의하고 존재를 등록하게 함으로써 이들 서비스를 호출/연계하여 필요한 기능을 구현할 수 있도록 한다. 시맨틱 웹은 웹 상의 정보를 의미에 기반하여 표준적인 방법으로 기술하도록 하고 이들 간의 의미적 호환성을 제공하여 에이전트 프로그램들이 웹 정보를 활용하여 사용자 태스크를 수행할 수 있도록 하는 환경이다. 시맨틱 웹 서비스는 이들을 통합한 환경으로 서비스에 관한 온톨로지와 개념 온톨로지를 사용하여 의미

기반 웹 서비스를 가능하게 한다. 따라서 다양한 정보검색 기능을 웹 서비스로 정의하고 이들간의 의미적 정보 교환이 이루어지게 함으로써 고차원의 사용자 정보 요구를 만족시킬 수 있다. 이러한 환경이 구축되기 위해서는 보다 실용적인 텍스트 마이닝 기술이 개발되어 현재 웹 문서에 대한 의미기반 메타 데이터가 생성이 되어야 하고, 온톨로지가 구축이 되어야 하며, 이들을 활용하는 검색 기술이 개발되어야 한다. 이러한 환경에서는 분산 통합검색 문제가 다루는 검색 시스템들간의 호환성 문제가 자연스럽게 해결될 것이다.

3.1.3 다중언어 정보 서비스

웹과 글로벌 환경의 발전에 따라 다양한 언어로 기술된 정보의 습득이 어느 때보다 중요한 역할을 할 것이다. 단순한 교차언어 검색 기술을 초월하여 교차언어 질의응답, 교차언어 요약, 교차언어 정보추출 등의 고급 서비스 기술이 개발되어 보다 상세하고 압축된 정보를 얻을 수 있게 함으로써 다중언어 정보 서비스의 활용성을 높일 수 있다. 예를 들어 검색된 모든 문서를 번역 서비스에 맡기는 것 보다 추출된 상세정보만 번역하는 것이 보다 현실적인 대안이 될 것이다.

3.1.4 사용자 혹은 태스크의 맥락정보 활용 기술

위에서 언급한 바와 같이 간단한 사용자 질의만을 가지고 사용자의 정보 요구에 충실한 양질의 정보를 제공하는 것은 매우 어려울 수 밖에 없다. 따라서 사용자의 속성, 선호도, 환경, 태스크 등을 모델링하여 질의 처리에 반영함으로써 보다 적절한 정보를 찾아 줄 수 있다.

3.1.5 멀티미디어 검색 기술

멀티미디어 데이터의 양이 급증하고 그 활용도가 높아지면서 멀티미디어 데이터의 메타 데이터를 생성하고 주석을 자동적으로 붙이는 작업에 대한 필요성이 날로 높아지고 있다. 이러한 기술의 발전과 더불어 추출된 자질이나 메타 데이터를 활용한 검색 모델이 개발되어야 하고 사용자 태스크 맥락에서의 검색 기술이 적용되어야 할 것이다.

3.1.6 무선 유비쿼터스 환경에서의 정보검색 서비스

무선 장치의 개발이 급진전되면서 피어간(peer-to-peer) 네트워크가 동적으로 자유자재로 형성이 되는 환경이 도래하고 있다. 이러한 환경에서는 새로운 정보검색 응용이 창출될 것이며 이에 대한 기술 개발이 요구된다. 예를 들어 여행 중 현지에서 가장 가까운 식당을 찾거나 현지 일기예보를 알아보기 위해 PDA로 질문하고 답을 받을 뿐만 아니라 특정 유적지에 도착했을 때 관련된 정보를 즉석에서 받아 볼 수 있는 서비스는 정보검색 관점에서 새로운 도전을 요구한다.

3.2 산업 전망

정보검색 및 저장, 관리 기술의 대상 시장은 웹 페이지 검색 및 번역, 전자 도서관 자료 검색 등의 서비스 요구 사항으로부터 도출된 멀티미디어 정보검색 엔진, 국제 표준 기반 문서관리 시스템, 상품 정보 검색, 디지털 방송 자료 검색 등의 직접 시장과 이를 활용하여 형성되는 디지털 방송, 군사 정보 시스템, 교육 정보화 등 응용 시장으로 나눌 수 있고 이러한 환경은 그림 1과 같다[31].

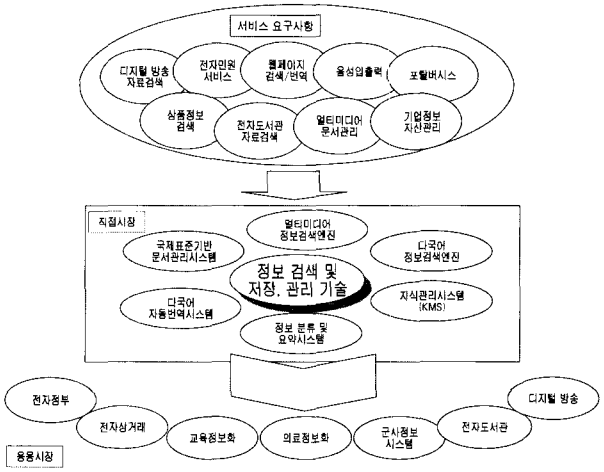

그림 1 정보검색 시장

응용분야는 사회 전 분야에서 정보화가 이루어짐에 따라 정보검색 관리 시스템에 대한 필요성이 제기되고 있어 매우 다양하다고 할 수 있다. 교육정보, 군사정보, 디지털 도서관, 디지털 방송, 유전자 분야, 의료 정보, 전자상거래, CRM 등의 분야를 들 수 있다.

IDC보고서[32]에 의하면 세계의 정보검색 시장은 2002년에 8억6천만불의 규모로 조사되었으며, 2005년에는 263억불의 규모로 성장하고 연평균 49.7%의 높은 성장률로 시장이 확대되어 2012년에는 세계 시장규모 443억불의 큰 시장을 형성할 것으로 예상된다.

4. 결 론

정보검색 분야는 웹의 보편화 및 컨텐츠 산업의 발전과 더불어 기술적인 면이나 산업적인 면에서 국내외적으로 급성장하였다. 기술 관점에서 전통적인 검색분야 뿐만 아니라 텍스트 마이닝의 대부분이 정보검색의 뿌리를 가지고 있다. 향후에도 독립된 분야로 지속적인 기술 개발이 이루어지겠지만, 학문 간의 간격이 좁아지고 융합 기술에 대한 중요성이 어느때 보다 강조되고 있는 현 추세를 볼 때, 자연언어처리, 데이터베이스, 인공지능, 멀티미디어 분야와의 합작 기술 개발이 활발하게 진행될 것으로 보인다.

참고문헌

[1] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill, New York, 1968.

[2] J. Allan et al. "Challenges in Information Retrieval and Language Modeling." http://ciir.cs.umass.edu/irchallenges/presentations/irchallenges428.pdf

[3] H. Turtle & B. W. Croft, "Evaluation of an inference network-based retrieval model," *ACM Transactions on Information Systems*, 9 (3), pp. 187-222, 1991.

[4] Robertson, S.E. and Walker S., "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232-241, 1994.

[5] J. Ponte & W. B. Croft (1998). "A language modeling approach to information retrieval," in Proc. of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275-281.

[6] R. Baeza-Yates & B. Ribeiro-Neto (1999). Modern Information Retrieval. Addison-Wesley.

[7] Salton, G., Fox, E,, & Wu, H., "Extended Boolean information retrieval," *Communications of ACM*, 26 (12), pp. 1022-1036, 1983.

[8] S. H. Myaeng, S. H. & C. Khoo (1993). "On Uncertainty Handling in Plausible Reasoning with Conceptual Graphs" in Conceptual Structures : Theory and Implementation. H. D. Pfeiffer & T. E. Nagle, Springer-Verlag, 1993.

[9] C. J. van Rijsbergen (1986). "A non-classical logic for Information Retrieval," The Computer Journal, 29: 481-485, 1986.

[10] S. Deerwester et al. (1990). "Indexing by latent semantic analysis," Journal of the American Society for Information Science, 41 (6), pp. 391-407.

[11] J. J. Rocchio (1971). "Relevance feedback in information retrieval," in The SMART Retrieval System - Experiments in Automatic Document Processing (ed: G. Salton), Prentice Hall Inc., Englewood Cliffs, NJ.

[12] J. Xu & W. B. Croft (1996). "Query expansion using local and global document analysis," in Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 4-11.

[13] S. H. Myaeng & R. R. Korfhage, R. R. (1990). "Integration of User Profiles: Models and Experiments in Information Retrieval." Information Processing and Management, Vol. 26, No. 6, pp. 719-738.

[14] O. Zamir & O. Etzioni (1998). "Web Document Clustering: A Feasibility Demonstration," Proc. of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, 1998, pp. 46-54.

[15] F. Sebastiani et al. (2002). Proc. of the Workshop for Operational Text Classification Systems, held at the 25th ACM SIGIR Conference on Research and Development in Information Retrieval, August, Tampere, Finland.

[16] Y. Yang (1999). "A re-examination of text categorization methods," Proc. of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, August, Berkeley, CA, pp. 42-49.

[17] H. J. Oh, S. H. Myaeng, & M. H. Lee (2000). "A practical hypertext categorization method usiing links and incrementally available class information," in Proc. of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 264-271.

[18] Y. B. Lee & S. H. Myaeng (2002). "Text Genre Classification with Genre-Revealing and Subject-Revealing Features," Proc. of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, August, pp.

145-150.
[19] J. Allan (ed.) (2002). Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers.
[20] K. Sparck Jones (1999). "Automatic summarizing: factors and directions," in Advances in Automatic Text Summarization (eds: Mani & Maybury), the MIT Press.
[21] Kupiec, J., Pedersen, J., and Chen, F., A Trainable Document Summarizer, Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, 1995.
[22] S. H. Myaeng & D. H. Jang (1999). "Development and evaluation of a statistically based document summarization system," in Advances in Automatic Text Summarization (eds: Mani & Maybury), the MIT Press.
[23] R. Barzilay & M. Elhadad, "Using Lexical Chains for Text Summarization," in Advances in Automatic Text Summarization (eds: Mani & Maybury), 1999.
[24] K. McKeown et al. (1999). "Generating concise natural language summaries," in Advances in Automatic Text Summarization (eds: Mani & Maybury), 1999.
[25] G. Grefenstette (1998). "The problem of cross-language information retrieval," in Cross-Language Information Retrieval (ed: G. Grefenstette), Kluwer Academic Publishers.
[26] G. Grefenstette (ed.) (1998). Cross-Language Information Retrieval, Kluwer Academic Publishers.
[27] E. Voorhees et al. (1995). "Learning collection fusion strategies," Proc. of ACM SIGIR, Seattle, WA.
[28] J. P. Callan et al. (1995). "Searching distributed collections with inference networks," in Proc. of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21-29.
[29] S. H. Myaeng, C. Jeong (2001), "A Protocol-Based Architecture for Federated Searching in Digital Libraries," Proceedings of th 4th International Conference of Asian Digital Libraries, p116-p124.
[30] M. Flickner et al. (1995), "Query by image and video content: the QBIC system," IEEE Computer, 28(9).
[31] 과학기술부 (2002). 국가과학기술지도.
[32] IDC, Search and Retrieval Technologies Market Forecast, 2001. 9.

맹 성 현

1983 미국 캘리포니아 주립대학 학사
1987 미국 Southern Methodist University(SMU) 석사 및 박사
미국 Temple University 조교
Syracuse University 종신교수
충남대학교 교수 역임
현재 한국정보통신대학교 공학부 교수
2002 ACM SIGIR Conference Program Committee Chair
ACM Transactions on Asian Language Processing 편집부위원장
Information Processing & Management
Journal of Natural Language Processing
Journal of Computer Processing of Oriental Languages 편집위원 등으로 활동
Home page : http://ir.cnu.ac.kr
관심분야 : 정보검색, 텍스트마이닝, 디지털 도서관, 시맨틱 웹 등
E-mail : myaeng@icu.ac.kr