사례 발표

텍스트마이닝 기반 고정밀 검색시스템

안 태 성* 서 형 국** 이 경 일***

목 차

1. 서 론

1.1 정보의 홍수와 검색 시스템

지난 10년 동안 인터넷의 대중화 덕분으로 World Wide Web과 e-mail은 이미 정보 전달의 일반적인 수단으로 자리를 잡았다. 인터넷과 이에 기반한 e-Business는 기존 산업의 전부분에 걸쳐 효율성과 생산성 증대를 위한 전략적인 도구로 그 중요성이 지속적으로 증대되고 있으며, 지식 노동자들은 업무 시간의 대부분을 문서로 대표되는 정보와 지식을 생산하고 검색하는데 보내고 있다. 새로운 기업정보 자료들이 끊임없이 등록되고, 지난 자료들이 수정, 갱신되는 등 전 세계에 있는 수 많은 기업에서 다양한 지식 자산(Knowledge Asset)들이 지속적으로 생성, 재활용되고 있다. 그러나 이렇게 기업이 생성, 저장, 재사용하는 정보 중 20% 만이 활용성이 높은 정형 데이터로 구성되어 있고, 나머지 80%는 워드프로세서, e-mail, 프리젠테이션, 스프레드시트, PDF와 같은 복합문서와 인터넷 페이지 등의 비정형 텍스트 형태로 구성되어 있다[1].

정형데이터의 SQL 검색으로부터 시작한 검색은 비정형 데이터를 위한 검색으로 발전하게 되고 다양한 검색 agent를 이용한 웹 검색으로 발전하게 된다. 그러나 이러한 검색 시스템도 새로운 e-Business 환경을 충분히 뒷받침하지 못하는 한계에 다다르게 되었다. 검색 엔진들이 너무나 많은 정보를 검색해 주기 시작하면서 검색의 문제는 원하지 않는 정보들 사이에서 유용한 정보를 찾는 것으로 변화하였다. 기존 시스템은 검색을 수행하면 결과 화면에 검색어가 포함된 문장만 짤막하게 표시되는 것이 대부분으로 사용자가 이 짧은 문장을 읽고 자신에게 필요한 문서인지를 판단하는 것은 매우 어렵고, 문서를 매번 다운로드하여 전문을 읽고 확인하는 것 또한 거의 불가능하거나 비효율적인 일임에 틀림없다. 이와 같은 정보 검색 환경에서 유용한 정보를 효과적으로 찾기 위해서 비정형 데이터인 문서로부터 유용한 정보를 추출하고 가공하는 기술의 필요성이 대두되게 되었다.

1.2 텍스트마이닝 기술

대량의 정보를 효과적으로 다룰 수 있는 방법에

* 모비코앤시스메타(주) IT Solution 팀장
** 모비코앤시스메타(주) IT Solution팀 주임연구원
*** 모비코앤시스메타(주) 부사장

대한 연구는 이미 활발히 진행되고 있다. DB에 저장된 자료와 같이 정형화 된 데이터로부터 정보를 추출, 가공하는 data mining은 이미 실용성을 갖추고 많은 분야에서 널리 활용되고 있다. 그러나 디지털 정보의 대부분은 일반 text 형태의 비정형 데이터이거나 mark up language형태의 반정형 데이터, 혹은 워드프로세서, e-mail, 프리젠테이션, 스프레드시트, PDF와 같은 복합 문서로 생산되고 있다. Text Mining은 이러한 비/반정형 데이터에 대하여 자연언어처리(Natural Language Processing)기술과 문서 처리 기술을 적용하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 문서 요약(summarization), 문서 분류(classification), 문서 군집(clustering) , 특성 추출(feature extraction) 등이 text mining의 핵심 연구 분야며 그 응용 분야는 매우 다양하다[2,3]. Data mining 관점에서 문서로부터 구조화된 정보를 추출하여 database화 시키거나 규칙을 찾아내는 것은 가장 일반적인 응용이며, 사용자가 Web 상에서 문서를 찾는 것을 도와주거나 사용자 profile의 생성 및 분석, 문서에 쓰인 자연 언어 식별, 대량 DB에서 문서의 분류 및 군집화, 문서 분류 정보를 이용한 문서 재해석, 신문/논문/보고서 요약, 문서 번역, 시계열(time series) 정보의 획득을 통한 시장 및 위험도 분석, 문서 색인, 문서 여과(filtering) 및 추천(re -commendation), 대표적 키워드나 토픽(topic)의 추출, 질의 응답 시스템, 대규모 문서에서의 탐색 등이 가장 대표적인 응용 분야라 할 수 있다[4].

2. 기존 검색시스템의 문제점 및 텍스트마이닝 기술의 필요성

〈표 1〉은 상용화 되어 있는 기존 검색엔진들의 문제점을 (1)원문 분석 성능, (2)사용성 및 기능성, (3)관리성 부문으로 구별하여 보여주고 있다.

〈표 1〉 기존 검색 엔진의 문제점

항목	문제점 목록
분석성능	• 형태소 분석 실패, 정보 간과 문제 • 단순 패턴 매칭에 의한 과도 검색 문제 • 검색 정확성 및 랭킹 문제 • 문서 분석 실패, 인덱싱 불가 오류 문제 • 다국어 정보 검색의 문제 • 비정형 지식 자산의 검색 문제
사용성	• 검색 결과, 정보량 과다의 문제 • 검색 속도 및 확장성의 문제 • 검색 결과의 표현 및 접근성의 문제
관리성	• 검색 인덱스 갱신, DB 동기화 문제 • 이기종 DB, 검색엔진의 통합 검색

2.1 형태소 분석 실패, 정보 간과 문제

질의한 키워드(단어)가 검색 대상 문서에는 분명히 존재하는데, 검색 엔진이 해당 문서를 찾아내지를 못하는 문제가 종종 발생한다. 특히 외산 검색 솔루션의 경우, 한국어 처리 부분이 취약하기 때문에 이러한 문제가 많이 발생하고 있다.

이 문제는 검색 엔진이 검색 대상 문서에 대해 간단한 형태소 분석을 수행한 후 추출된 형태소들을 통째로 인덱싱하는 방식을 사용하기 때문에 발생을 하는데, 형태소 분석기에 등록되어 있지 않은 신조어 및 고유명사에 대해서는 인덱싱을 수행하지 못하고, 사용자 검색 시 정보 간과 문제를 발생시키게 된다[5]. 특히, 철자 및 띄어쓰기 오류가 있는 문서와 복합명사 검색에 있어서 낮은 성능을 보이며, 꾸준한 신규어(미등록어) 등록 및 형태소 분석기 성능 향상 등의 지속적 노력을 필요로 한다.

이 문제를 해결하기 위해 띄어쓰기 및 철자 오류 보정기능을 가진 고정밀 한국어(다국어) 형태소 분석기[6] 사용이 필요하며, 인덱싱 방식에 있어서도 키워드 방식과 n-gram 방식을 하이브리드로 적용해야만 한다.

2.2 단순 패턴 매칭에 의한 과도 검색 문제

사용자가 검색 요청한 키워드가 아닌 동음이의어 혹은 단어 중간의 일부분을 과도하게 분석한 과도 검색 및 오검색 결과를 출력하는 문제다. (예 : "생선" 질의 → 대학생선교회 검색) 이 문제는 n-gram 형식의 역파일 기법을 사용한 검색엔진들이 보여주는 일반적인 오류 형태로, 형태소 분석 혹은 단어의 의미 분석을 수행하지 않고 어절들을 기계적으로 의미 없는 2개 혹은 3개의 글자 단위로 분할하여 인덱싱함으로 발생하게 된다. 이 경우에도 복합명사 등의 사용자 띄어쓰기 표현에 따라 오히려 검색 문서 누락 문제가 발생하기도 한다.

문제의 해결을 위해서는 이미 거론된 첫 번째 문제 해결 방식과 같이 고정밀 형태소 분석기가 사용된 키워드 방식과 n-gram 방식의 인덱싱을 하이브리드로 적용할 필요가 있으며, 한 단계 더 나아가 텍스트마이닝에 기반한 핵심 키워드 검색 및 유사 문서 검색 기술 등을 적용함으로 이 문제를 효과적으로 해결하고 사용자 편익을 증진시킬 수 있다.

2.3 검색 정확성 및 랭킹 문제

검색된 결과 상위에 중요 정보가 위치해 있지 않고, 오히려 중요도가 떨어지는 문서들이 상위에 출력되어, 검색 결과에서 필요한 정보 찾기를 포기하게 되거나, 정말 많은 사용자 노력을 필요로 하게 되는 문제를 많이 경험하게 된다.

이러한 문제는 정교하지 못한 검색 랭킹 알고리즘이 사용되었거나, 사용자 관점에서 검색 결과 순위가 조정되지 못했기 때문에 발생한다. 사용자 혹은 시스템 구축자가 랭킹에 대한 독창적 모델을 가지고 있다고 하더라도, 검색 엔진이 랭킹 모델을 수정하거나 변경하기 어려운 구조로 되어 있거나, 랭킹 순위 조정을 위한 변수가 제한되어 있어서 검색 대상 도메인, DB에 따라 적절한 대응을 할 수 없는 경우가 흔하다.

이 문제 해결을 위해 검색엔진은 다양한 랭킹 조정 변수 적용이 가능해야 하며, 검색 분야 및 고객별 유연한 랭킹 정책을 적용할 수 있게 개발되어야 한다. 기존의 랭킹 시스템에 텍스트 마이닝 기술 적용을 통해, 검색 정확도가 매우 높은 "알짜 검색" 기능을 구현함으로도 해결 가능하다.

2.4 검색 정확성 및 랭킹 문제

정보 관리 및 검색 시스템(특히 EDMS, KMS 및 자료관 시스템 등)에서 일부 형식 혹은 일부 버전의 복합문서(첨부문서)를 등록하지 못하거나 검색을 수행하지 못하는 문제가 종종 발생한다. 이는 문서로부터 검색 대상 텍스트를 추출해 내는 문서 필터의 성능이 떨어지거나, 다양한 형식의 문서를 처리하지 못하기 때문에 발생하게 되는데, 대부분의 검색엔진 회사들이 문서 필터를 외부 필터 회사로부터 구매해 사용하고 있기 때문에 고객의 요구에 적절히 대응하고 있지 못하는 문제를 가지고 있다.

이 문제 해결을 위해서는 고품질 문서 필터를 개발, 확보하여 다양한 버전과 문제들에 대해 효과적인 대응이 가능하여야 하며, 단순한 텍스트 추출 수준이 아닌, 텍스트, 표, 그래프, 삽입 이미지 등의 문서 개체 분석이 가능하고, 공통 문서 포맷을 이용해 이들을 유기적으로 통합할 수 있는 통합 문서 분석 시스템 개발이 필요하다.

2.5 검색 결과, 정보량 과다의 문제

검색 엔진에 등록된 정보가 많아서 웬만한 검색 질의어에 수천건 이상의 검색 결과를 보여 줌으로, 사용자가 검색 결과를 모두 검토하기 어려워 정보 검색을 포기하게 되는 경우가 있다. 경우에 따라서는 중복 문서 등록의 문제를 포함하기도 한다.

이 문제는 기존의 검색엔진들이 사용자 질의 키워드가 문서에 포함되어 있기만 하면 그 키워드가

문서에서 어떤 역할을 하는가, 어떤 의미를 가지고 있는가와 상관없이 무조건 검색 결과로 출력하기 때문에 발생을 하는데, 다수 키워드로 질의된 경우에도 각 키워드의 관련성(연관성, 거리)이 무시되고 단순히 검색 대상 문서에 키워드들의 포함 여부만으로 검색을 하기 때문에 발생하게 된다.

이 문제의 해결을 위해 최근에 많은 연구가 진행되었는데, 특히 고품질 랭킹 시스템의 개발과 함께 사용자 질의 키워드가 정보의 핵심 단어인 경우만 출력하는 "알짜 검색" 기술의 개발, 검색 엔진에 인덱싱 시 유사 문서 확인을 통한 동일 내용 문서의 중복 등록 방지 기술 적용, 클러스터링에 기반한 유사 문서 검색, 예제기반 검색(QBE), 텍사노미(Taxonomy)의 적용을 통해 검색 결과 수를 효과적으로 줄이거나 원하는 지식 정보에 빠르게 접근하도록 하는 기술들이 제공되고 있다.

2.6 검색 인덱스 갱신, DB 동기화 문제

일부 검색엔진들은 DB 혹은 File 서버에 저장된 정보가 신규 등록, 갱신, 삭제될 때 이를 검색 엔진에 반영하는 절차가 복잡하거나 자동화되어 있지 않아 전체를 다시 인덱싱해야 하는 문제를 가지고 있다. 이 경우 검색엔진이 DB와 유기적으로 실시간 연동되어 있지 못해 DB 무결성에 심각한 영향을 주는 경우가 있다.

이 문제는 검색 엔진의 인덱스 구조가 증분 인덱싱을 지원하고 있지 못해서 인덱스 파일을 갱신하고자 할 때 전체를 full 인덱싱 해야만 하거나, 인덱싱 속도가 느려서 실시간 인덱싱이 불가능하여 정기적으로 일괄 인덱싱을 수행해야 하는 경우에 발생한다. 이 경우 DBMS와 검색 엔진이 유기적으로 결합되어 있지 않아 수작업으로 DB에서 텍스트를 export하고, 다시 검색엔진에 등록을 해야 하는 관리상의 문제가 발생하게 된다.

최근의 검색 엔진들은 고속의 실시간 증분 인덱싱(incremental indexing)을 지원하고 있으며, 인덱스 에이전트(index agent)를 통하여 자동 인덱싱 및 유기적 연동성을 보장하고 있다.

2.7 이기종 DB, 검색엔진의 통합 검색

최근 들어서는 여러 조직에서 개별적으로 구축된 DB 혹은 리포지토리를 통합하여 검색하고 관리하기 위한 연구가 많이 진행되고 있다. 일반적인 메타검색 방법을 통해서는 이질적인 DB 혹은 file system에 저장된 방대한 정보들을 통합 검색하기 불가능하거나, 그 속도가 너무 느려서 사용이 매우 어렵게 된다. 이 문제는 다양한 정보원으로부터 변경, 갱신된 정보를 자동으로 통합, 중앙의 검색엔진에 적용하는 기술이 도입되지 않았거나, 검색엔진의 폐쇄적 구조로 상호 연동되기 어려운 근본적인 문제로부터 기인한다. 또한 메타 검색 시스템의 구조적 문제로 각 검색엔진에 개별 질의한 결과를 각각 분석하여 공통 포맷으로 통합하는데 많은 시간이 소모되고, 이 경우 연결된 이기종 검색 엔진의 구성과 스키마에 민감하여, 해당 검색 시스템이 변경, 개선되면 오히려 심각한 오류를 보이게 된다.

이러한 문제를 해결하기 위해서는 각 검색 엔진들이 상호 연동을 위한 표준을 설정, 지원하여야 하며, 이것이 어려울 때는 기존의 각 검색엔진을 수정하거나 단순 연동하는 메타검색 방식 보다는 index agent가 개별의 검색엔진 단에 쉽게 plug-in 되어 정보를 수집, 정리하여 분산 처리하는 분산 인덱싱, 분산 검색하는 기술의 적용이 필요하다.

3. 구축 사이트 분석 및 전략

3.1 구축 사이트 분석

본 논문에서는 국내 최고의 법률 정보 전문 포탈인 로앤비(LawnB)에 텍스트마이닝 기반의 새로운 검색 엔진을 도입하여, 기존 검색 서비스가 가

진 문제를 어떻게 효과적으로 해결했는지를 고찰하고자 한다.

로앤비는 법률 관련 총 100만건 이상의 국내 최대 고품질 DB를 확보하고 있으며, <표 2>와 같이 총 40개의 서로 다른 형태의 복잡한 schema 구성의 DB(혹은 리포지토리)를 통합 검색해야하는 높은 목표를 가지고 있다.

로앤비에서 사용하던 기존의 검색 엔진은 보유하고 있는 총 정보량의 30% 밖에 인덱싱을 하고 있지 못했으며, 그마저도 위에서 논의된 "정보 간과" 문제를 심각하게 발생하고 있었다. 또한, 방대한 DB를 통합 검색하기 때문에, 웬만한 사용자 질의 키워드에 대해서는 수천건 이상의 과다한 검색 결과를 출력해, 사용자가 실제로 필요한 정보를 찾아내는데 많은 어려움이 있었다. 기존 검색 엔진은 증분 인덱싱 및 각 DB와의 유기적 연동을 지원하지 못해 관리 상의 여러 어려움도 존재하였다.

<표 2> 로앤비의 보유 DB 내역

DB 명	데이터수
판례	12만건 이상
법령	15만건 이상
최근제개정법령	3000건
법령안입법예고등록	2000건
조약	1500건
법령용어	5000건
주석서	25000건
문헌(평석,논문)	35000 건
법률용어	1600 건
법조인명록	10000 건
법률서식	1100 건
생활법률사례	1500 건
기업법무사례	1600 건
웹법률정보	1500 건
법률뉴스	23000 건
뉴스속의판례	2500 건
단행본	4500 건
논문집	5500 건
개별문헌	2200 건
정기간행물	10000 건
학위논문	50000 건
학술정보	11만건
최신법령해설	40 건
계약서실무	100 건
정보통신법률연구	50 건
판례속보	1000 건
삼일TexBrief	100 건
칼럼	150 건
법률상담사례	1500 건
기업법무메뉴얼	30 건
파워리포트	60 건
고시계	15000 건
E-Book	100 건
연혁법령	50관건
섹션, 케어클럽, 일반 게시판 및 파일 등	수천 건

본 구축 사례에서는 다음과 같은 10가지 핵심 목표를 설정하였다.

(1) 전체 정보의 통합 검색 구현
(2) 고품질 전문 검색 성능 구현
(3) PDF 형태의 고급 문서 전문 검색 구현
(4) 문서 자동 열기, 자동 위치 이동 및 하일라이팅 기능 구현
(5) 고정밀 알짜 검색 구현(*)
(6) 핵심 키워드 및 특성 추출 기능 구현(*)
(7) 판례 자동요약 기능 구현(*)
(8) 유사 판례 검색 기능 구현(*)
(9) 자동 증분 인덱싱 및 통합 관리 기능 구현
(10) DRM 연동을 통한 콘텐츠 보호 기능 구현

이 중에 특히, 5,6,7,8번 목표는 텍스트마이닝에 기반한 고정밀 검색 시스템을 구현하는 것으로, 국내에서 텍스트마이닝 연동 검색 시스템으로는 최대 규모가 될 것으로 추정된다.

3.2 제공 시스템 기능

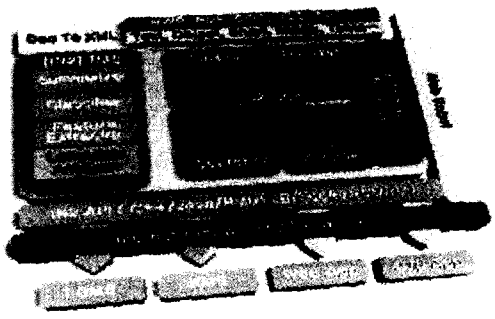
본 구축 과제의 성공적 수행을 위해서 모비코앤시스메타(주)의 차세대 검색 엔진인 [IN2] DOR[7]과 텍스트마이닝 시스템인 [IN2] TMS가 사용되었다.

[IN2] 시스템은 (그림 1)과 같은 기본 엔진 구성을 가지고 있으며, XML 및 JAVA로 개발되어 다양한 환경에서 원활히 작동되며, 높은 확장성을 가지고 있다.

[IN2] 시스템은 다음과 같은 기본 기능을 제공

하고 있다.

- 한국어, 영어, 일본어 등 다국어 고정밀 언어 분석(형태소 분석 등) 모듈 포함
- 각 언어별 형태소 사전, 교차어 사전, 시소러스 사전 등을 포함
- 키워드 인덱싱과 n-gram 인덱싱의 장점만을 결합한 하이브리드 인덱싱 기능을 지원
- 다양하고 복잡한 검색 연산식 지원 (무제한 검색식)
- 대용량 검색과 대규모 동시 사용자를 위한 고성능 분산 인덱싱, 분산 검색 기능 지원
- 문서 내에 포함된, 표, 차트, 그림 등의 비정형 지식 정보를 분석, 인덱싱, 검색할 수 있는 기능을 제공
- 다양한 검색식 지원과 Auto-Navigation & Highlighting, 검색 문장 미리보기 기능 제공
- 핵심 키워드 추출 및 고유명사(이름/지명 등), 사건, 전화번호, e-mail 주소 등의 특징 추출 기능 구현
- 생성 요약, 추출 요약을 모두 지원하는 자동 요약 시스템 포함
- 문서 자동 분류 및 군집 기능 제공
- 텍스트마이닝에 기반한 알짜검색 지원
- 검색 결과로부터의 유사문서 검색 기능 지원
- 자연어 질의, 특히 자연언어 예제기반 질의(QBE) 기능 지원

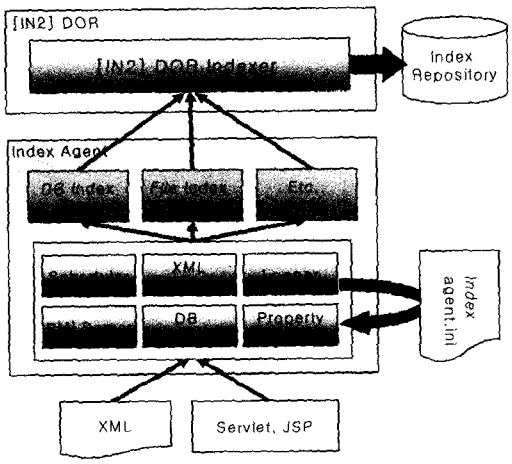


(그림 1) 차세대 검색 시스템 [IN2]의 구성

3.3 구축 방법

3.3.1 통합 검색 구현

40 종의 상이한 DB에 대한 통합 검색과 DB와 검색엔진의 유기적 연동을 위해, 인덱스 에이전트를 구성하였다.



(그림 2) 인덱스 에이전트

인덱스 에이전트는 DB 및 파일의 검색 정보를 인덱스에 추가, 갱신, 삭제에 관한 업무를 자동으로 수행하는 모듈로서, DB 접속 정보와 SQL문, 색인 타입 등의 정보를 구조화된 XML의 형태로 관리하며, 인덱스 에이전트가 이를 분석하여 해당하는 작업을 수행한다. 인덱싱을 수행하는 시기는 명령을 내리는 즉시 혹은 특정 시간에 예약 수행할 수 있으며, 동작 주기를 지정하여 주기적 인덱싱도 가능하도록 설계되어 있다.

3.3.2 고품질 검색 성능 구현

고품질 전문 검색 기능을 달성하기 위해, "띄어쓰기 비종속 한국어 형태소 분석기"를 사용한 키워드 인덱싱 박식과 n-gram 방식을 결합한 하이브리드 인덱싱 방법을 사용했다. 중요도순, 날짜순 등 다양한 기준으로 검색 결과가 정렬될 수 있도록 기

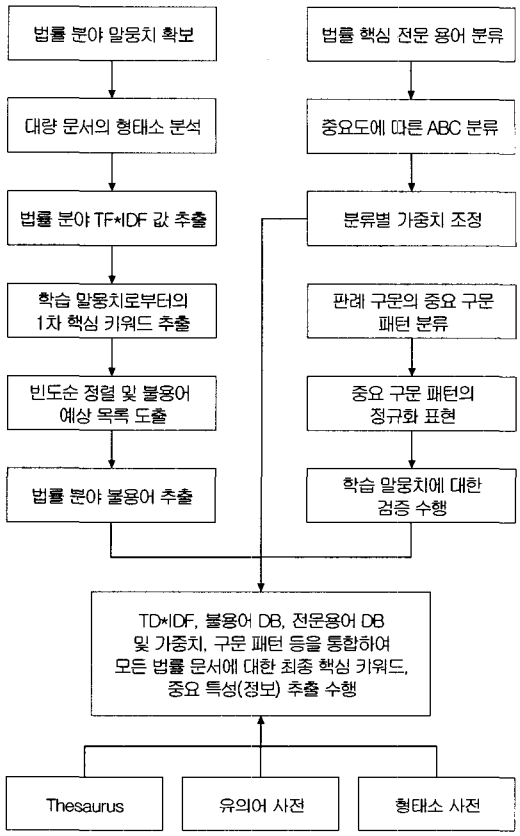
능개선 하였으며, 전문 검색 결과를 미리 볼 수 있는 미리 보기 기능을 포함하도록 하였다. 본 시스템 구축에서는 DB와 연동되어 DB 내의 text를 인덱싱하는 것뿐만 아니라, 수만건의 PDF 문서에 대한 인덱싱도 병행이 되었다. PDF 문서에 대한 인덱싱은 단순 텍스트 필터링뿐만 아니라, 검색 결과 page로의 자동이동과 하일라이팅 기능 구현을 위해서, 분석된 형태소들의 위치정보가 검색엔진에 같이 인덱싱 되었다. 실제 사용자가 검색을 하였을 때는 Web browser 상의 Active-X 콘트롤에 의해 PDF 문서가 자동 download되어 열리고 해당 위치로 자동 이동을 하게 된다.

3.3.3 핵심 키워드 및 특성 추출 기능 구현

핵심 키워드 및 특성 추출은 텍스트마이닝에 기반한 고정밀 알짜 검색, 판례 자동요약 기능, 유사 판례 검색 등의 기능 구현을 위해 가장 중요한 부문이다. (그림 3)과 같이 키워드와 특성 추출을 위해 통계적 기법과, 구문 패턴 매칭을 통한 규칙 기반 특성 추출, 사전에 의한 중요도 가중치 부여를 통합 평가하는 방식으로 핵심 키워드와 특성 추출을 수행하였다.

3.3.4 텍스트마이닝에 기반한 고정밀 검색

본 과제에서는 사용자가 필요로 하는 핵심적인 내용만 소수정예로 출력되는 "알짜 검색" 기능, 검색된 판례를 "자동으로 요약"하는 기능, 검색 및 검토된 관심 판례와 유사한 판례만 재 검색해 주는 "유사 문서 검색" 기능이 구현되었다. "알짜 검색"은 사용자가 질의한 질의어가 포함된 모든 문서를 사용자에게 제시하는 것이 아닌, 사용자 질의어가 "핵심 키워드"혹은 중요한 개념 및 특징으로 사용되는 문서에 대해서만 검색 결과로 출력하는 기능으로서, 통상 수천건의 검색 결과에서 수십건 미만의 핵심적인 중요 검색 결과만 사용자에게 제시, 빠른 정보 획득과 의사 결정을 도울 수 있도록 하며, 정보가 부족하였을 때 유사문서 검색 및



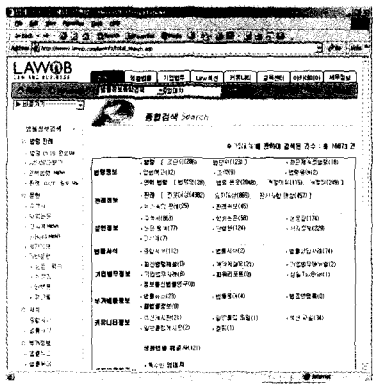
(그림 3) 핵심 키워드 및 특성 추출

전문 검색 등으로 추가 확장할 수 있도록 설계되었다. "자동요약"기능은 기존의 중요문 추출 요약뿐만이 아니라, 추출된 특성과 핵심 키워드 및 개념과 다단계 template set에 기반한 생성요약을 구현하였다. 구현된 자동요약 기능은 사용자가 검색된 결과에서 빠르게 자신에게 필요한 정보인지 아닌지를 판별할 수 있도록 돕는다. 마지막으로 문서 클러스터링 기술에 기반한 "유사문서 검색"기능은 사용자가 찾아낸 관심 법률 정보에 대해 가장 유사한 문서들만 키워드가 아닌 문서 전체로 클러스터링하여 재검색하는 기능을 제공한다. 문서 클러스터링은 기 추출된 핵심키워드와 특성 벡터들을 인덱싱하고 특성 벡터의 유사도를 고속으로 연산하는 방식으로 구현되었다. 빠른 검색 속도 구현을 위해, [IN2]DOR 엔진의 하부에서 시스템 통합을 수행 하였다.

4. 시스템 구축 내용 및 결과

아래 (그림 4)는 텍스트마이닝 기반의 고정밀 검색 시스템이 도입된 LawnB 포탈 사이트를 보여주고 있다. 상단의 네비게이션 부분에 통합 검색을 위한 UI를 제공하고 있으며, 왼쪽의 메뉴바에 각 부문별 검색을 위한 link를 걸어 두었다.

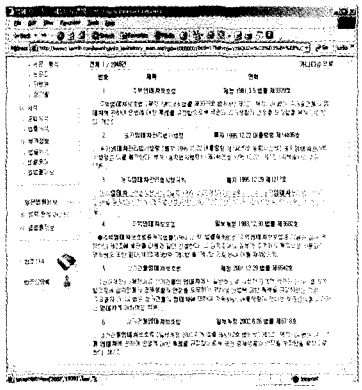
(그림 4) 시스템 구축 화면



(그림 5) 통합 검색 화면

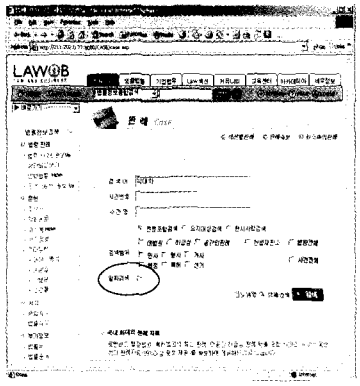
위 (그림 5)는 통합 검색 결과 화면이다. 앞에서 설명한 것과 같이 본 사례에서는 총 40개의 DB와 리파지토리로부터 100만건 이상의 정보를 통합 검색하도록 구현되어 있다. 인덱스 에이전트의 적용을 통해 각 정보원과 검색 엔진이 유기적으로 동기화되었으며, P4 2Ghz, 2CPU 환경에서 1초 미만의 검색 속도와 200명 이상의 동시 접속 성능이 달성되었다.

아래 (그림 6)은 통합 검색 중 하나를 선택했을 때, 해당 DB 중 사용자 질의 결과를 출력하는 화면을 보이고 있다.



(그림 6) DB 상세 검색 화면

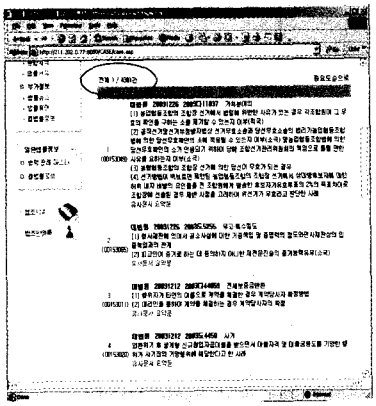
아래 (그림 7)에서 (그림 11)까지는 텍스트마이닝에 기반한 고정밀 검색 결과를 보여주고 있다. (그림 7)은 판례DB 검색 부분에서 "임대차"의 키워드로 질의를 하고 있는 화면이다.



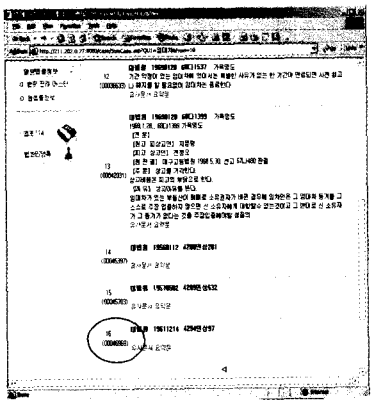
(그림 7) 판례 부분 검색 화면

아래 (그림 8)은 일반 전문 검색을 수행된 결과로 총 4381건이 검색되었음을 알 수 있다. 이와 비

교해서 (그림 7)의 검색 옵션에서 "알짜 검색"을 선택하여 검색을 했을 경우 (그림 9)과 같이 총 16건의 핵심적인 결과만 출력을 하여 사용자에게 필요한 알짜 정보만을 우선 검토할 수 있는 방법을 제공하고 있다.

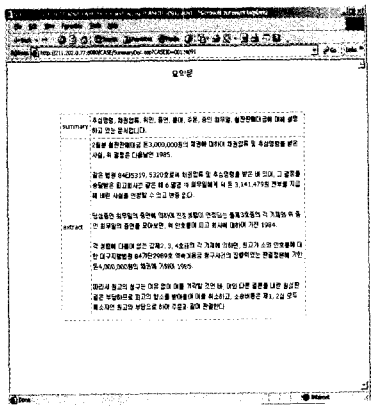


(그림 8) 일반 전문 검색 결과

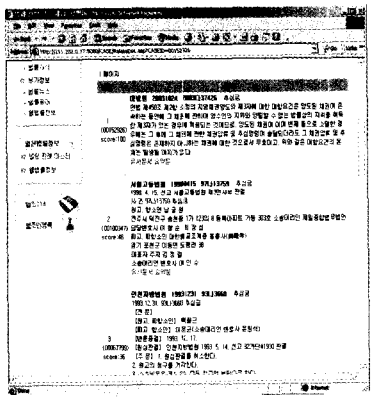


(그림 9) 텍스트마이닝을 통한 알짜검색

아래 (그림 10)은 검색된 판례에 대해 자동 요약한 결과를 보이고 있다. 상단에는 추출된 키워드와 특성 중심의 생성 요약을 보이고 있으며, 하단에는 본문 전체에서 가장 중요하다고 판단되는 문장을 추출 나열한 추출요약 결과를 보이고 있다. 사용자는 자동 요약문을 검토함으로써 판례 본문 전체를 읽어보지 않더라도 자신과 관련 있는 내용인지를 빠르게 판단할 수 있다.



(그림 10) 자동 요약 결과 화면



(그림 11) 유사문서 검색 결과 화면

(그림 11)은 유사 문서 검색 결과를 보이고 있다. 유사 문서 검색은 문서 클러스터링 기술에 기반하고 있는데, (그림 8) 혹은 (그림 9)의 1차 검색이 완료된 후에 자신이 관심 있는 내용의 문서를 기준으로 가장 유사한 문서들을 추가로 검색을 할 때 매우 유용하게 사용될 수 있다. 유사 문서 검색의 경우, 사용자 키워드에 의한 검색이 아닌, 문서 전체를 기반해 질의를 하게 된다. (그림 11)의 최상단에는 질의한 문서 자신(100% 매칭)이 위치하고, 그 아래 46% 유사성을 보이는 문서를 검색하여 보여주고 있다.

5. 결론 및 향후 과제

본 구축 사례를 통해, 그 동안 연구 단계에 머물러 있었던 텍스트마이닝 기술을 대규모 정보 검색 부문에 적용, 원활한 상용 서비스가 가능함을 검증하였다. 텍스트마이닝 기술은 기존 검색 시스템의 여러 문제점과 한계들을 효과적으로 극복하고 사용자 편익을 증가시켜 높은 서비스 만족감을 제공할 수 있을 뿐만 아니라, 정보 검색 절차 및 비용을 획기적으로 줄여, 시스템 도입에 따른 높은 ROI 달성에 기여할 것으로 기대된다.

앞으로는 기존의 통계 및 규칙 기반의 텍스트마이닝 기술에 온톨로지 기반의 지식 검색 기능을 부여하여, 한 단계 발전된 정보 추론을 수행하는 연구가 진행되어야 할 것으로 생각된다.

참고문헌

[1] 노나카 이쿠지로, Michael Polanyi, Delphi Group, 1998.

[2] Fabrizio Sebstiani, "Machine Learning in Automated Text Categorization", ACM Computing Surverys, Vol.34, No.1, 2002.

[3] Koller D, Sahami M, "Hierarchically classifying documents using very few words", Proceedings of the Fourteenth International Conference on Machine Learning(ICML 97), 1997.

[4] 언어정보산업협의회, "언어정보산업 동향", 2002년도 Annual Report, 2002.

[5]이준호, 김광현, 김지승, "다양한 한글 문서 색인 방법들에 대한 평가", 제5회 한국 과학기술정보인프라 워크샵 학술발표 논문집, 2002

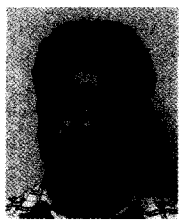
[6] 이경일, 안태성, "띄어쓰기 비종속 품사태깅 시스템 개발", 제 15회 한글 및 한국어 정보처리 학술대회, 2003.

[7] 안태성, 임중수, 김명훈, 안우람, 이경일, "복합 문서 검색 시스템 - [IN2]DOR", 제 15회 한글 및 한국어 정보처리 학술대회, 2003.

저자약력

안 태 성

1995년 인하대학교 전자재료공학과(공학사)
1995년~1997년 (주) 정소프트 주임연구원
1997년~1998년 (주) 유니소프트 선임 연구원
1999년~2000년 Learnout & Hauspie S/W Engineer (美)
2000년~2002년 (주) 시스메타 기술이사
2003년~현재 모비코앤시스메타(주) IT Solution 팀장
관심분야 : 자연언어처리, 검색, 유비쿼터스
이 메 일 : albert@sysmeta.com



서 형 국

1997년 한국과학기술원 전산학과(공학사)
2001년~현재 모비코앤시스메타(주) IT Solution팀 주임연구원
관심분야 : 자연언어처리, 소프트웨어엔지니어링
이 메 일 : apollos@mobico.com

이 경 일

1996년 인하대학원 전자재료공학과(공학석사)
1997년~1999년 LG중앙연구소 연구원
1999년~2001년 현대전자연구소 연구원
2001년~2003년 (주)시스메타 대표이사
2003년~현재 모비코&시스메타(주) 부사장
관심분야 : NLP, Information Retrieval, Text mining, Machine Translation
이 메 일 : tony@sysmeta.com