대용량 분석 시스템을 이용한 교통 연구 검색 방법론에 관한 연구

배진아*, 윤 청**
*충남대학교 컴퓨터공학과 석사과정
**충남대학교 컴퓨터공학과 교수
e-mail: 1503kpop@naver.com*, cyoun@cnu.ac.kr**

A Study on Traffic Research Retrieval Method using Large Capacity Analysis System

Jin-Ah Bae*, Cheong Youn**
Dept of Computer Science, Chung-Nam University

요 약

지난 몇 년간 우리는 소셜 검색에 몰두하여 연관검색 및 소비자의 만족을 위해 빅데이터 분석을 하였다. 최근에는 빅데이터 분석이라는 흐름에 맞춰 기업 및 기관별 본연의 정보를 통합하여 효율적인 검색을 할 수 있도록 하는 솔루션을 대거 도입하고 있다. 또한 기업 및 기관에서 가지고 있는 정보는 기존 비정형 데이터로 방대하여 기존의 방법이나 도구로 수집 및 저장·분석이 어려운 실정이다. 이에 공공기관 및 민간기업 등에서는 키워드 중심의 다양한 검색엔진을 개발하거나 도입하고 있으며, 정보 분류의 확대, 메타데이터의 활용, 태그정보의 제공, 개인 맞춤형 서비스 등 고객의 만족도를 제고하기 위한 다양한 방법을 시도하고 있다.

본 연구에서는 기관의 교통 연구와 관련한 일련의 작업 중 행정문서, 연구정보, 유관기관 게시물 등의 통합 빅데이터를 가지고 검색시스템을 구현하였다. 이와 더불어 사용자 사전 및 동의어 사전을 통한 검색 키워드를 데이터베이스에 저장하여 검색 효율성을 제고하는 방안을 제시한다.

1. 서론

최근에는 빅데이터 분석이라는 흐름에 맞춰 기업 및 기관별 본연의 정보를 통합하여 효율적인 검색을 할 수 있도록 하는 솔루션을 대거 도입하고 있다. 또한 기업 및 기관에서 가지고 있는 정보는 기존 비정형 데이터보다 방대하여 가존의 방법이나 도구로 수집 및 저장·분석이 어려운 실정이다. 더불어 검색 사이트에서 원하는 정보를 신속·정확하게 검색하는 것이 무엇보다 중요하게 되었다.

이에 키워드 중심의 다양한 검색엔진을 도입하거나 정보 분류의 확대, 메타데이터의 활용, 태그정보의 제공, 개인 맞춤형 서비스 등 고객의 만족도를 제고하기 위한 다양한 방법을 시도하고 있다.

그럼에도 불구하고 사회 환경의 변화와 기술발달의 속도가 더욱 빨라져 인터넷에 유통되는 정보의 양이 급격하게 증가하고 있고, 키워드 검색은 사용자의 검색 의도와는 관계없이 검색 결과가 너무 많거나, 전혀 발견할 수 없는 경우가 발생하기도 하고, 사용자가 원하지 않는 정보를 제공하는 등 인터넷 정보에 대한 검색의 효율성 측면에서 많은 단점을 가지고 있다.

이러한 문제점을 해결하기 위해 공공기관과 민간 포털 사이트를 중심으로 컴퓨터가 사용자의 의도를 정확하게 파악하여 검색 결과를 서비스 하는 기술이 적용되고 있다. 또한, 사용자의 검색 패턴을 분석하고 통계적 기법 등을 활용하여 정보에 우선 순위를 부여하는 제공 서비스가 늘어나고 있으며, 사용자 질의어와 관련된 연관어를 제공하는 사례가 증가하는 등 검색 기술을 활용한 검색 서비스가 증가하고 있다. 특화된 공공기관의 경우 전문용어가 새롭게 생겨나고 생활분야 및 활동 범위에 따라 다양한 교통용어들이 산재되어 있기에 원하는 정보를 얻기가 어려웠기 때문이다.

본 연구에서는 기관의 교통 연구와 관련한 일련의 작업 중 행정문서, 연구정보, 유관기관 게시물 등의 통합 빅데이터를 가지고 검색시스템을 구현하였다.

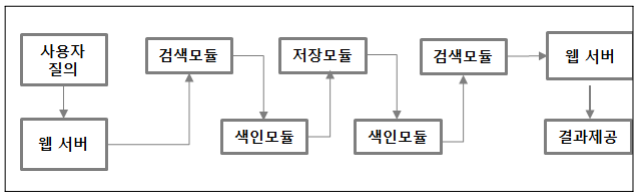
이와 더불어 사용자 사전 및 동의어 사전을 통한 검색 키워드를 데이터베이스에 저장하여 검색 효율성을 제고하는 방안을 제시한다.

2. 대용량 분석시스템 설계

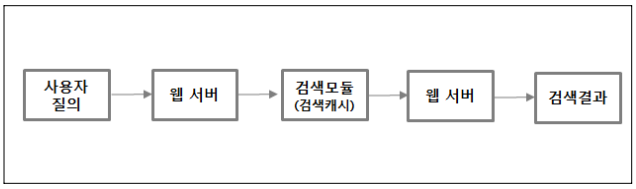
2.1 검색시스템 설계

본 연구에서는 교통 연구 목적의 마리너4 프로그램으로 검색사이트를 구축하여 교통 관련 연구 용어를 검색해보았다. 먼저 일반검색과정은 [그림1] 사용자의 질의가 웹서버를 통해 검색/색인/저장모듈을 거쳐 웹서버에서 검색결과를 제공한다. 반면 본 연구에서 구현한 대용량 분석 시스템 검색과정을 보면 [그림2] 메모리캐시기능이 탑재되어있어서, 사용자 질의와 그에 따른 결과를 미리 메모리 영역인 검색캐시에 저장하였다가 추후 검색 시 결과를 바로 제공 한다.

기존 검색방법과 차이점은 단계를 최소화하여 서버의 부하를 줄이고, 해당 검색어에 대한 색인결과를 사용자에게 빠르게 제공할 수 있다. 검색 결과 내용이 많거나 검색 복잡도로 인하여 검색연산이 많은 경우에 효율적으로 대응이 가능하다.



[그림1] 일반검색과정



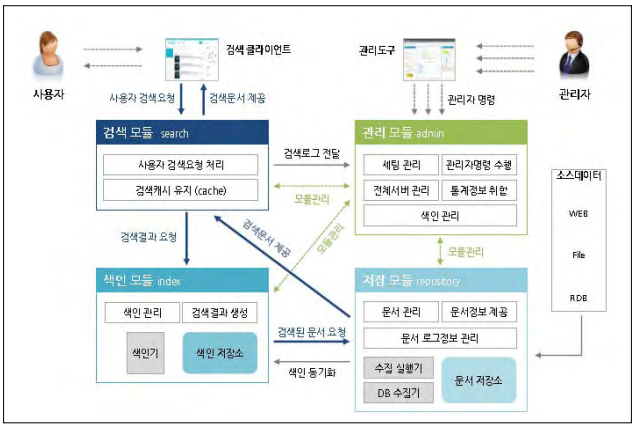
[그림2] 대용량 분석 시스템 검색과정

빅데이터를 위한 검색 성능과 최적화된 검색 결과를 제공하여 사용자들이 입력한 검색어와 사용자가 선택한 문서의 상관관계 데이터를 바탕으로 검색어별 관심도를 분석하여 검색 정확도에 활용하는 랭킹 기법을 제공하며, 사용자, 검색어 카테고리 등 여러 속성별 프로파일에 따른 차별화된 검색 결과를 제공한다.

2.2 시스템 구성

본 검색 시스템의 구성은 다음과 같다.

- 운영체제: Red Hat Linux 6
- 웹 서버: Apache-Tomcat 8
- 사용언어: java(jdk version 1.7)
- 데이터베이스: Oracle 11g



[그림3] 교통 연구 검색시스템 개발 프레임워크

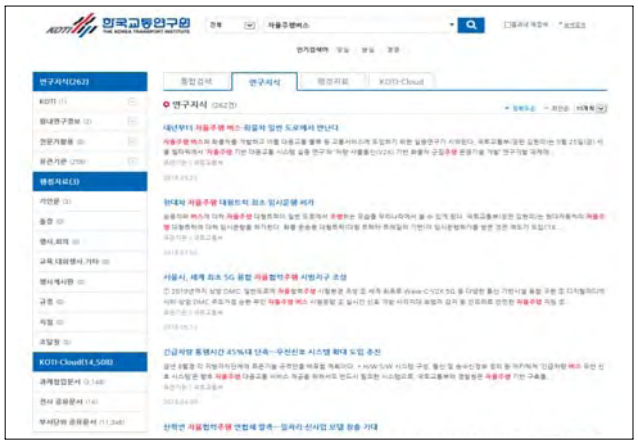
[그림3]은 본 연구에서 구현한 대용량 분석 시스템 개발 프레임 구조이다. 검색시스템은 검색모듈, 관리모듈, 색인모듈, 저장모듈로 나뉘고 있다. 사용자가 검색요청을 하면 검색모듈과 색인모듈을 통하여 검색 문서를 제공 한다. 그리고 검색 키워드 결과 로그를 관리모듈에 저장하여 축적 하여 관리한다. 이에 관리자는 관리모듈을 통하여 사용자들이 어떤 검색을 하는지 주제별, 키워드별, 분류를 하여 관리 통계를 수집할 수 있다.

2.3 검색엔진 구현

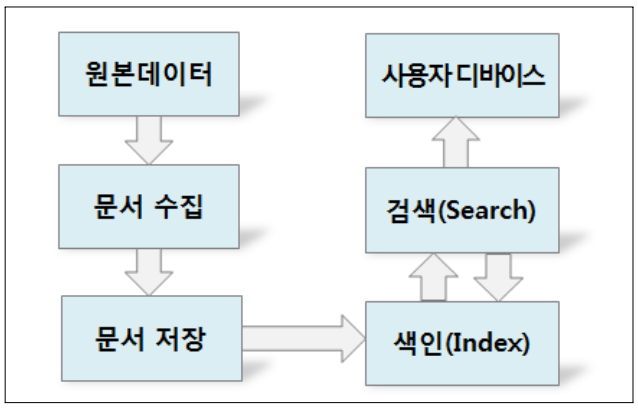
[그림4]는 대용량 분석시스템을 이용한 교통연구 검색사이트의 예시 화면이다. 검색 시스템의 카테고리는 연구지식, 행정자료, 클라우드 자료로 세가지로 나뉜다. 연구지식에는 연구제안서, 연구수행계획서와 같은 연구에 관련된 데이터를 볼 수 있고, 행정자료는 연구에 필요한 제반사항을 이행하기 위한 행정처리, 즉 회의개최 및 자문회의, 출장계획/결과 와 같은 비용 지급처리에 대한 데이터들이다.

클라우드 카테고리는 연구를 수행하는 과정에서 발생한 정제 되지 않은 중간단계의 file 또는 최종 완료된 보고서 file 등이 속한다. 이 카테고리의 데이터는 내부 인트라넷에서 각 시스템에 저장된 데이터를 RDBMS를 사용하여 추출하고 사용자에게 검색결과 정보 제공한다. [그림5]와 같이 검색흐름도를 보면 각 시스템의 분산된 데이터를 수집하고자 전문문서수집기능이 서버에서 주기적으로 실행된다. 문서수집은 데이터양이 방대할수록 수집 소요시간이 걸리므로 주로 새벽시간대에 일1회 수행되도록 설정되어있다. 따라서 동의어 사전 또는 유의어사전 등록을 통한 사용자사전을 구축한 다음 결과값을 비교해보려면 전문 수집이 완료 된 이후 검색결과비교를

하여야 한다. 전문 수집이 수행되어야 사용자사전에 따른 수집한 정보를 색인하여 사용자 디바이스에 검색 결과로 표출된다.



[그림4] 교통 연구 검색 서비스 화면 예시



[그림5] 검색 시스템 흐름도

2.4 사용자 사전 검색방법론 및 랭킹기법

사용자가 원하는 검색결과를 효율적으로 추출하기 위해서 사용자사전과 동의어사전, 유의어사전 등의 기능이 있다. 더불어 검색키워드에 가중치를 두어서 랭킹기법을 이용할 수 도 있다. [그림6]은 사용자사전 등록 화면의 예시이다. 그림과 같이 복합명사와 사용자 사전 편집을 하여 언어 처리의 정확도를 높이고, 사용자가 원하는 검색 결과를 도출하여 검색 품질을 향상 시킨다.

사용자 또는 관리자가 의도한 특성에 맞게 지식사전을 편집하고 반영하여 관리자나 사용자가 원하는 최적화된 검색결과를 제시한다.

[그림6] 사용자 사전 등록 화면 예시

2.5 동의어, 유의어 검색방법론

사용자가 검색어로 입력한 단어가 다른 키워드로 검색되었을 경우, 이를 보완하기 위해 유의어사전, 동의어사전 등록으로 언어를 확장하여 사용자에게 제공한다.

동의어사전은 사용자가 교통수단을 키워드로 검색 하였을 때 버스, 기차, 택시, 자동차, 자전거, 트램 등을 동의어로 등록 해놓으면 교통수단을 검색한 결과 값은 동의어를 포함하여 확장된 검색결과를 추출해준다.

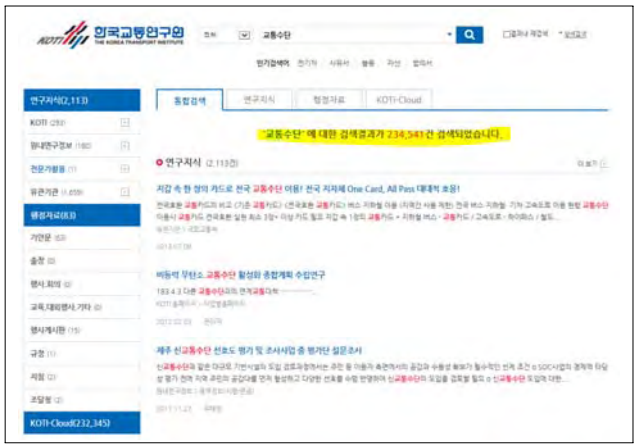
[그림7]을 보면 교통수단으로 검색하였을 때 결과 건수가 89,586건으로 값이 나왔다. 단순히 교통수단 키워드가 본문 또는 첨부파일에 포함된 결과자료만을 표시한 것으로 보인다.

이에 동의어사전에 교통수단은 버스, 기차, 택시, 자동차, 자전거, 트램을 등록하였다.

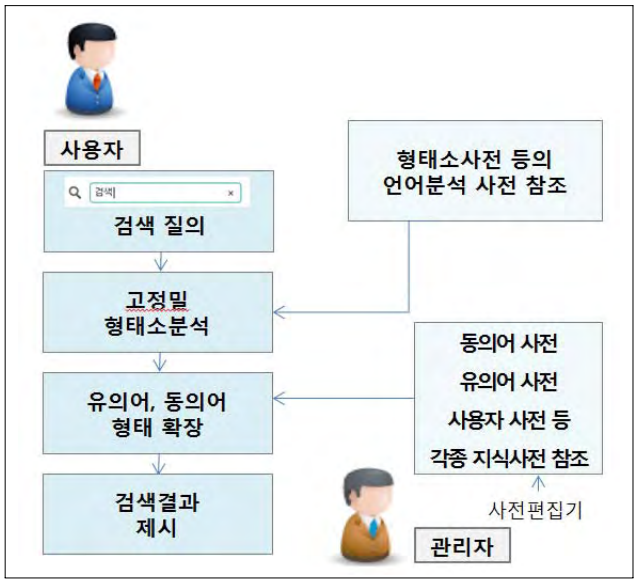
[그림8]을 보면 검색결과 건수가 234,541건으로 확장된 결과 값을 얻을 수 있을 것이다.

이 결과치를 효과적으로 보기 위해서는 카테고리 분류가 편리하게 되어있어야 할 것이다. 해당 검색 사이트에서는 연구관련 지식과 행정지식을 나누어 카테고리 분류를 하였다.

[그림7] 동의어사전 적용 전 검색결과



[그림8] 동의어사전 적용 후 검색결과



[그림9] 동의어/유의어 검색 프로세스

3. 결론

본 연구에서 제시한 교통 연구 정보 검색시스템 방법론은 교통이라는 특수목적을 가진 교통 분야 키워드를 통해 비슷한 성향의 사용자들의 관심도를 기반으로 한 검색결과 랭킹기법을 제공한다. 검색어에 따른 프로파일 별 클릭 통계를 검색 정확도에 반영하여 사용자의 검색 만족도를 점진적으로 향상시킨다.

이를 통해 기존 키워드 검색 방식에서 동의어, 유의어사전을 통한 연관 검색어 등록하여 새로운 교통 분야에 지식이 없더라도 연관키워드로 검색을 확장시켜 나갈 수 있도록 하여 정확성(Precision)을 향상시킬 수 있다.

또한 사용자가 원하는 정보를 쉽게 검색할 수 있도록 유도하여 동의어 및 사용자사전으로 관련검색어를 보다 효율적으로 연관되어 검색될 수 있도록 하였다.

검색의 연관성에 대하여 사용률이 높아질수록 사용자 사전의 DB가 축적되어 보다 나은 검색결과를 얻게 될 것으로 기대된다. 향후 시멘틱 검색 서비스의 고도화된 알고리즘을 반영한 검색시스템이 필요할 것이다.

이에 빅데이터에 기반한 교통 정보의 다양한 정보의 통합검색기능 시스템을 시멘틱 분석 알고리즘으로 확장시켜 나가는 연구가 필요할 것으로 보인다.

참고문헌

[1] 김건우, 안석준, 문혜정, “빅데이터 분석을 통한 정치 문화 경제 뉴스의 미디어 프레임 연구” (2007)

[2] 김학래, 김홍기. 2007. 시멘틱 웹/온톨로지 기술을 이용한 개인용 전자문서 검색 시스템. [한국전자거래학회지], 12(1): 135-149

[3] 허선영, 김은경. 2008. 시멘틱 검색엔진 설계 및 구현. 2008 한국컴퓨터종합학술대회 논문집, 35(1): 331-335.

[4] Jung, S. T., “A Study on Methodology of Semantic Search for Law Information using Life Term”, Thesis for Master Degree at Yonsei University, 2011.

[5] Chang, I. H., “Developing and Evaluating an Ontology-based Legal Retrieval System”, Journal of KLISS, Vol.45, No.2(2011), 345~366.

[6] Lame, G., “Using NLP Techniques to Identify Legal Ontology Components : Concepts and Relations”, Artificial Intelligence and Law, Vol.12, No.4(2005), 169~184.

[7] Atteveldt, W. H., Semantic network analysis: Techniques for Extracting, Representing, and Querying Media Content. BookSurge Publishing, Charleston, SC, 2008.

[8]Albertoni, Ricardo, Alessio Bertone, and Monica De Martino. 2004. “Semantic Web and Information Visualization.” 1st Italian Semantic Web Workshop, Ancona, Italy.

[9] Bangyong, Liang, Tang Jie, and Li Juanzi.2005. “Association Search in Semantic Web: Search + Inference.” WWW 2005 Conference, Chiba, Japan.

[10] Gruber, T. 2003. “It Is What It Does: The Pragmtics of Ontology, invited talk at Sharing the Knowledge.” International CIDOC CRM Symposium, Washington, DC., USA.