고객중심의 과학기술정보 서비스를 위한 FAST 검색엔진 커스터마이징

FAST Search Engine Customizing for S&T Information Service

한희준, 이태석, 김선태, 예용희, 이상기, 여일연
한국과학기술정보연구원 지식정보센터 서비스개발팀

Han Hee-Jun, Yi Tae-Seok, Kim Sun-Tae, Yae Yong-Hee, Lee Sang-Gi, Yeo Il-Yoen
Service System Development Team, Korea Institute of Science and Technology Information

요약

다양한 인터넷 기술이 개발 및 발전됨에 따라 정보 제공자는 사용자에게 보다 효율적이고 고객중심의 서비스를 제공하기 위해 노력하고 있다. 특히 방대한 양의 정보에 대하여 고객이 원하는 정보를 정확하고 쉽게 제공하기 위해서는 검색기능의 효율성이 필수이다. 한국과학기술정보연구원(KISTI)에서는 국가과학기술포털서비스 성능향상을 위하여 FAST(Fast Search & Transfer ASA) 검색엔진을 도입하였다. 하지만 무엇보다도 서비스 환경에 적합하게 검색엔진의 하드웨어 및 소프트웨어적 성능을 최적화하는 것이 중요하다. 본 논문에서는 국가과학기술정보의 효율적 서비스를 위한 FAST 검색엔진 설계 및 최적화 기법에 대해 논한다.

Abstract

According to develop the web technology, the data providers are trying to offer the efficient service for customers. Specially it is necessary to improve efficiency of the search function to help user access easily useful information their want. KISTI has introduced and customized the FAST search engine to improve search performance of the national science and technology information portal service system. But the design work for hardware and software implementation of search engine is important above all. In this paper, we discuss about the design and custormizing skill of FAST engine for the KISTI S&T information search service.

I. 서 론

1. 검색서비스 요구사항

인터넷을 통한 정보유통 기술이 고도화되면서 정보제공자는 서비스 차별화 및 고객 중심의 편리한 정보이용 환경을 제공할 필요성이 증대되고 있다. 정보 서비스 측면에서 고객 요구사항 및 정보 소비환경 변화를 적시에 반영하기 위한 사용자 인터페이스의 유연성이 필요하며, 초보자부터 전문가에 이르는 다양한 소비 계층의 요구사항을 모두 충족시켜야 한다. 또한 정보 서비스의 기본 기술인 검색엔진은 고성능을 지원하고, 대용량의 정보를 효율적으로 처리할 수 있도록 설계 및 운영되어야 할 뿐만 아니라 지능형 검색을 지원하여야 한다.

한국과학기술정보연구원에서는 국가과학기술정보(논문, 특허, 연구보고서, 분석동향정보, 규격, 과학기술인력 등)약 1억건에 대해 검색서비스를 제공하고 있으며, 안정된 검색 성능과 검색결과 분류 및 그룹핑, 문서 클러스터링 등 고성능의 검색기능을 갖춘 KISTI 과학기술정보 통합검색 플랫폼 구축을 위해 FAST 검색엔진을 도입, 커스터마이징하였다. 본론에서는 검색엔진 설계 및 최적화 기법과 관련된 커스터마이징에 대해 논하고 결론에서는 고성능의 검색기능 결과와 향후 계획에 대해 논한다.

II. 본 론

1. 검색엔진 구성

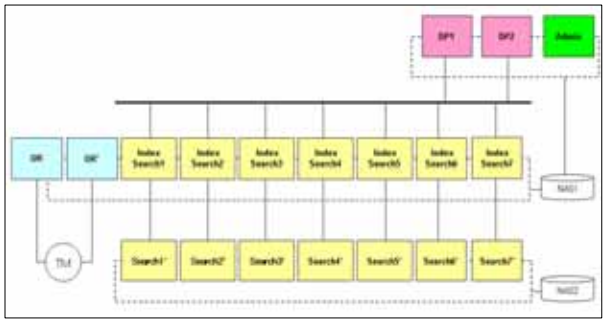
FAST 검색엔진의 하드웨어는 Admin 서버, Content Distrubute(CD) 서버, Document Process(DP) 서버, Query and Result(QR) 서버, Index 서버, Search 서버, 크게 6개 기능을 수행하는 서버로 구성된다. 각 서버의 기능은 표 1과 같다.

[표 1] FAST 엔진의 서버 기능

구분	서버 기능
Admin Server	검색엔진의 전체 프로세스를 관리하는 서버로 색인 및 검색 과정에서 발생하는 모든 이벤트를 모니터링
CD Server	JDBC와 같은 커넥터에서 수집한 데이터를 하나 이상의 DP에 분배, JDBC 는 DBMS에 저장된 소스데이터를 수집하는 프로세스이며 일반적으로 데이터의 수집 속도가 데이터 처리 및 색인 속도보다 월등히 빠르므로 CD 및 DP 서버에서 동작

DP Server	CD로부터 전송받은 데이터를 처리(언어식별, 개체명 인식, 형태소 분석 등)하여 색인 전단계 파일인 FIXML 생성
QR Server	이용자 질의 파싱/변환 후 검색요청 및 검색결과 수신 후 통합 및 정렬
Index Server	FIXML을 이용하여 실제 binary 색인을 생성
Search Server	QR 서버의 요청을 받아 검색 질의어를 받아 검색 수행하고 결과셋을 반환

1억여건의 과학기술정보 검색시스템 구성을 위하여 총 19대의 Linux 서버를 이용하여 FAST 엔진 설계를 위한 서버를 구성하였다. Admin 서버는 전체 프로세스의 동작을 실시간 모니터링하므로, 다른 서버들에 비해 CPU 작업이 많으므로, 별도의 1개 독립 서버로 구성한다. JDBC Connector 와 CP, DP 기능을 수행하는 프로세스는 색인 및 검색 이벤트에 비해 그 수행주기가 짧고 I/O process 가 적게 발생하는 기능이므로, 총 2대의 서버를 구성한다. 검색서비스 이중화를 위하여 검색 row 를 2개로 설계하여, 각 row 당 각각 7개의 검색서버를 할당하고, 각 row 는 동일한 binary 색인정보를 가지게 된다. 색인 프로세스는 row1 에 포함된 7개의 검색서버와 동시에 탑재되어 동작하며, row1에서 생성된 색인정보는 row2 검색서버에 복사된다. 검색쿼리를 분석하여 검색서버에 결과를 요청하고 Application 에 검색결과를 사용자가 원하는 형태로 반환하는 역할을 수행하는 QR 서버는 2대로 구성하여 이중화한다. 검색엔진 서버 구성은 그림 1과 같다.



▶▶ 그림 1. 검색서버 구성도

2. 프로세스 설계 및 최적화

데이터 색인은 먼저 오라클 데이터베이스의 원천데이터를 접근하여 Content Distributor 에게 전달하는 JDBC connector 가 담당한다. 하나 이상 동시에 동작 가능하며, 수집된 데이터는 CD에 의해서 Data Processing 단계로 넘어가게 된다. DP 과정에서는 데이터 특성에 따라 각각 다른 언어처리, 형태소분석, 정제, 사전처리 등을 거칠 수 있으며, 이런 각각의 데이터 처리 과정을 하나의 Pipeline 이라고 표현한다. 하나의 Pipeline 은 여러개의 Stage 단계를 거쳐 이루어진다. 표 2는 한글과 논문이 혼합된 형태의 레코드셋을 처리하는 일련의 Stage 들로 이루어진 Pipeline 의 예이다. 이 과정을 거쳐 데이터와 데이터 처리과정이 기술된 FIXML이 생성되고, Index Dispatcher 에 의해 각 7개의 색인서버에 FIXML 이 분배되어 색인이 생성된다.

[표 2] 논문데이터 처리 Pipeline

Stage 명	기능
DocInit	문서의 속성정보를 초기화
CopyUri	문서의 고유 ID 생성위해 Unique uri 를 예약된 속성값에 복사
FastXMLReaderData	데이터 값들을 XML 형태로 파싱
Tokenizer	공백처리, 조사분리, 복합어 분리 등 tokenization
Lemmatizer	각 Token에 대해 원형복원 등 형태소 분석 수행
AttributeAssignerLangEN	기본 처리언어를 영어로 부여
EngLemmatizer	영문제목, 영문요약 필드의 lemmatization
AttributeAssignerLangKO	기본 처리언어를 한글로 부여
Vectorizer	docvector(문서 핵심어) 생성
DateTimeNormalizer	날짜 및 시간표시 형식 정규화
DateTimeSelector	대표 날짜 및 시간 선택기록
MapperTransformer	숫자형태의 데이터를 내부형태 변환
RankTuner	랭크 튜닝 정보 추가
FIXMLGenerator	FIXML 생성
RTSOutput	Indexer 에 색인 수행을 위한 출력

검색서버(Search Server)는 각각 6개의 파티션으로 구성된다. 각각의 파티션은 최대 허용 Attribute Vector Memory(소팅, 그룹핑을 위한 색인정보)를 1.2GB 로 설정하여 최대 300만건의 문서에 대한 색인을 가진다. 서비스 대상이 되는 과학기술정보의 8종류이며, 이는 세부적으로 30개의 collection (JDBC connector 에 의해 크롤링되어 관리되는 단위)으로 구성된다. 검색 Application 의 요구에 의해 한 개 이상의 collection 을 머지하여 하나이상의 view 를 생성하여 사용할 수 있다.

[표 3] 과학기술정보 서비스 단위별 색인정보

DB종류	FAST Collection	View name
논문	논문(article)	newarticlesppublished
	페이퍼(paper)	
	코리아사이언스(ksarticle)	newksarticlesppublished
	저널(journal)	newjournalsppublished
	프로시딩(proceeding)	
	NDSL저자(author)	newauthorsppublished
	DDC코드 주제(subject)	newsubjectsppublished
	FSTA(fsta)	newfstasppublished
	INSP(inspec)	newinspecsppublished
	OAI논문(oai)	newoaisppublished
	OA저널(oajo)	newoajosppublished
	OAI저자(oaiauthor)	newoaiauthorsppublished
	학위논문(degreearticle)	newdegreearticlesppublished
특허	한국특허(kpatent)	newpatentsppublished
	일본특허(jpatent)	
	미국등록특허(upatentreg)	
	미국공개특허(upatentopen)	
	유럽국제특허(epatent)	
	한국디자인(kpatentdesign)	newdesignsppublished

연구보고서	국내보고서(koresearch)	newresearchsppublished
	해외보고서(foresearch)	
분석동향	분석리프트(analysis)	newanalysissppublished
	동향지식지(tld)	
	글로벌동향브리핑(trend)	
표준규격정보(standard)		newstandardsppublished
사실정보(fact)		newfactsppublished
인력정보(hum)		newhumansppublished
통합검색용 view(전체 collection포함)		newtotalsppublished

3. 검색기능 고도화

3.1 Docvector 개선

indexer는 한 문서에 대해 자주 출현되는 단어 및 어구 또는 중요단어를 추출하여 문서를 대표하는 docvector 를 생성한다. docvector 추출은 색인단계에서 수행되는데, 이 때 불용어 사전을 정의하여 docvector에서 제외시킬 단어 및 어구를 적용하였다. 각 문서를 대표하는 단어 리스트인 docvector는 유사문서를 검색할 때 쿼리로 사용되는데, 실제 문서에 자주 출현하는 대표단어라 할지라도 유사성 측정에 관련성이 없는 단어들은 제외시킬 필요가 있다. 하지만 docvector 에서는 제외되더라도 실제 검색에서는 유용한 단어가 될 수 있기 때문에 실제 binary 색인 생성시에는 불용어 사전을 적용하지 않는다. docvector 성능 개선을 위해 적용한 불용어의 예는 논문정보에서 자주 출현되나 중요 키워드가 아닌 '연구', '동향', '논문', '현황', '이용', '구현', '효율적' 등이다. 아래는 docvecot의 예이다.

- 문서제목 : '다양한 디스플레이 장치를 위한 xy 색도도상에서의 색역 사상 및 확장 기법'
- docvector : [디스플레이 장치, 1][display devices, 0.774597][색역 사상, 0.774597][xy 색도, 0.774597][색도도상, 0.707107][gamut mapping, 0.632456][chromaticity diagram, 0.632456][xy chromaticity, 0.632456][display device, 0.547723][extension method, 0.547723][확장, 0.489898][chromatic adaptation, 0.447214][순응 모델, 0.447214][색 순응, 0.447214][각 디스플레이, 0.447214]

사용자에게 보다 유용한 정보를 제공하기 위해 유사문서검색 기능을 구현하였다. 이는 docvector를 이용한 재검색을 통해 해당 결과문서와 유사성이 있는 문서리스트를 제공한다. 즉 한 문서의 상세보기를 수행할 때 해당 문서의 docvector 리스트의 상위 10개 단어를 이용하여 제목, 요약, 키워드 필드에 재검색을 요청하여 유사성이 높은 10개 문서 리스트를 동시에 제공한다.

3.2 Anti-Phrase 적용

사용자의 검색 쿼리를 제한하지는 않으나, 검색성능및 속도에 영향을 미칠 수 있으며, 중요한 의미를 가지지 않는 단어들은 Anti-Phrasing 기능을 적용하여 쿼리에서 제외시킨다. 예를 들어 제목필드에 대한 쿼리가 resolution of image 일 때 엔진에 결과를 요청하는 쿼리는 AND(TI:resulotion, TI:image) 형식이 되어 검색속도 및 성능을 향상시킬 수 있다. 적용 단어는 NDSL 시스템에서 분석된 출현회수가 많은 상위 30개 불용어를 적용하였다.

[표 4] Anti-Phrase 적용 단어리스트

of, and, in, the, for, on, to, with, by, from, an, at, as, during, between, after, or, under, into, how, that, up, what, over, can, among, against, about

3.3 Navigator 를 이용한 검색결과 Refine

색인단계에서 Navigator 로 사용할 필드를 미리 지정하면, 해당 필드별로 다양한 문서 브라우징 및 그룹핑이 가능하다. 검색결과 간략리스트를 표현할 때 검색결과를 포함하고 있는 분류체계나 메타데이터별로 군집하여 건수를 표수함으로써 2차 검색을 위한 유용한 옵션을 제공한다. 날짜와 같은 숫자 데이터에 대해서는 구간을 설정하거나, 비율별로 구간을 가변적으로 조절하여 나타낼 수 있다. 그림 2에서 보는 바와 같이 논문 색인단계에서 저널/프로시딩명, 저자명, 발행년도, 언어, 주제분야 필드를 Navigator 항목으로 지정하여 결과셋에 대해 군집화해 Refine 기능을 제공한다.

▶▶ 그림 2. 논문 Refine 기능화면

Ⅲ. 결 론

본 논문에서는 대용량의 과학기술정보를 효율적으로 색인

및 서비스하고 사용자에게 다양한 검색기능을 제공하기 위하여 색인을 위한 하드웨어, 소프트웨어적인 설계와 검색성능을 향상시키기 위한 search view 생성방법, Anti-Phrase 적용방법, 유사문서검색방법, Navigator 이용 방법 등, FAST 검색엔진을 최적화하는 방법에 대해 논하였다.

향후에는 검색어 로그 및 통계를 활용하여 응용프로그램에서 활용하는 방법과 유사문서 검색 및 군집화 품질 향상을 위하여 소스데이터 분석 및 docvector 추출 개선이 필요하며, 문자열 필드에 대한 정렬 기능 및 네비게이터의 빠른 제공을 위해 요구되는 서버별 메모리 자원 활용 분석을 통해 검색 응답속도의 안정성을 보장하는 연구가 필요하다.

▌참 고 문 헌▐

[1] FAST Installation Guide
[2] FAST Operation Guide
[3] FAST Search Fron End(SFE) Developer's Guide
[4] Application Integration Guide
[5] Indexing Database Content and XML Guide