Search Formula-1

"순수 국산원천기술에 의한 대용량,자동분류,분산 구조 고성능 통합검색엔진 Search Formula-1"

(주)코리아와이즈넛(대표:추호석 www.wisenut.co.kr)이 개발한 고성능 통합검색엔진 Search Formula-1은 대용량 처리 기술 및 자동 분류 기술을 기반으로 정확한 검색 결과를 보여주는 고성능 분산 검색엔진으로 개발 초기부터 유연성(Flexibility)과 확장성(Extensibility)을 염두에 두고 개발되어. 고객의 다양한 환경과 요구를 만족시키기 위한 어떠한 Customization 작업이라도 최소의 비용으로 빠른 시일 안에 구현할 수 있고 다양한 플랫폼 및 솔루션과 효과적으로 통합될 수 있다.세계 최고의 대용량 처리기술 노하우 보유하고 있어 일일 최대 1억건의 색인 가능하고, RDB 데이터(Oracle, Sybase, Informix, DB2), 비정형문서(MS-Word, Excel, Powerpoint, HWP, PDF, 훈민정음, 아리랑), 웹페이지 검색등 다양한 형식의 데이터 검색이 가능하다. 또한 순수 국산 기술로 개발되어 한글처리에 뛰어나며 가장 진보된 방식의 랭킹알고리즘을 적용하여 검색의 정확성이 매우 우수하고 다양한 방식의 자동분류 기술을 통해 정보의 구조성 강화시켰고 편리하고 다양한 관리자 인터페이스를 제공한다. 코리아와이즈넛은 통합검색엔진 Search Formula-1 출시 후 코래드, 서울대학교 통합검색 시스템 구축을 비롯하여 국군 기무사령부, 필룩스, 한국수력원자력, 한국과학기술기획평가원(KISTEP), 육군 조달본부, 이타임즈인터넷, NATE.com, 제일은행 등에 통합검색엔진 공급계약을 체결하였다.
하고 있다. 문의: 02-589-6116, 011-9313-5939 담당 : 박용철 대리

Search Formula-1

1. 작품명 : Search Formula-1

2. 제작자 : (주)코리아와이즈넛

대표자 : 추호석
개발참여자 : 박재득, 서광준, 윤종완, 임성채 외 8명
주소 : (137-943) 서울시 서초구 양재동 275-5 태석빌딩18층
전화 : 02) 589-6100
팩스 : 02) 589-6109
E-mail : park@wisenut.co.kr

3. S/W 요약설명

코리아와이즈넛 Search Formula-1은 대용량 처리 기술 및 자동 분류 기술을 기반으로 정확한 검색 결과를 보여주는 고성능 분산 검색엔진으로 개발 초기부터 유연성(Flexibility)과 확장성(Extensibility)을 염두에 두고 개발되었다.

순수 국산 원천기술을 보유하여 고객의 다양한 환경과 요구를 만족시키기 위한 어떠한 Customization 작업이라도 빠른 시일 안에 구현할 수 있도록 개발되어, 다양한 플랫폼 및 솔루션과 효과적으로 통합된다.

4.개발 배경

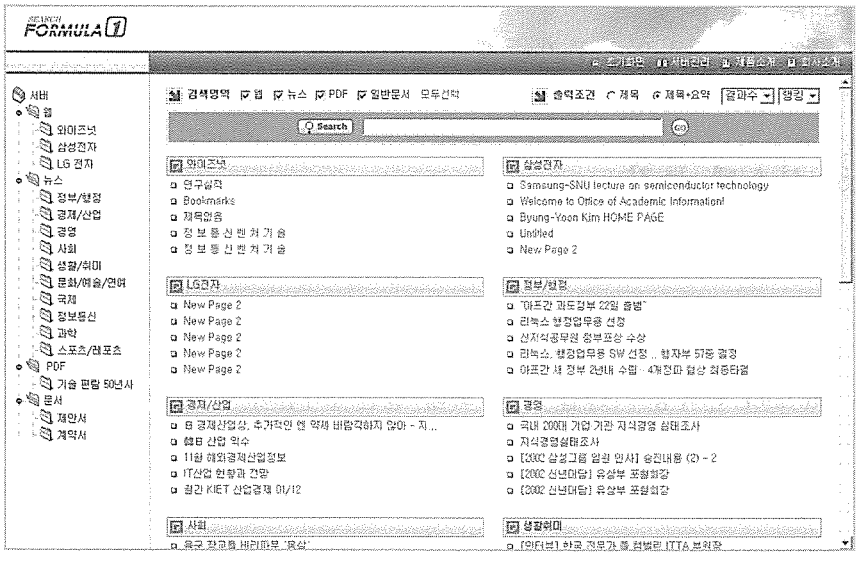
기존 외산 검색엔진의 경우는 원천기술을 보유하고 있지 않아 customization이 쉽지 않고 한글 형태소 분석에 한계가 있어 한글

검색에 있어 검색정확도가 떨어진다. 또한 일반 검색엔진은 분산처리 구조가 이루어지지 않아 대용량 처리에 한계를 나타낸다.
이러한 대용량 처리 및 뛰어난 한글 검색의 요구를 반영하고 수입대체효과를 수반하는 고성능 검색엔진를 개발하게 되었다.

5. 시스템 개요

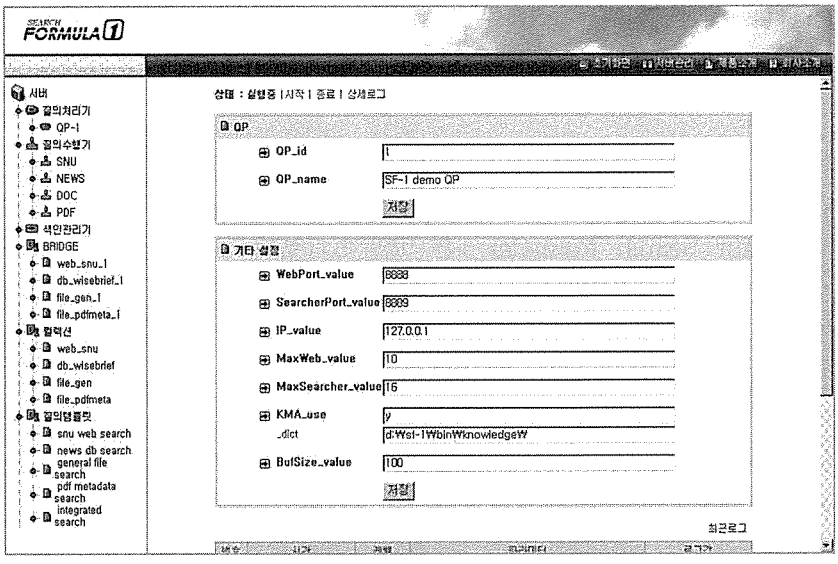
WISEnut Search Formular-1은 개발초기부터 유연성과 확장성을 염두에 두고 개발된 고성능 분산검색엔진이다. 현재 검색엔진이 사용되는 컴퓨팅환경에서 가장 큰 문제로 대두되고 있는 것은 대용량 데이터의 정확한 검색 처리이다. WISEnut Search Formula-1은 실제 대용량 검색 시스템을 구현하고 이를 기반으로 일반인에게 24x7 서비스를 안정적으로 제공한 노하우를 기반으로 WISEnut Search Formula-1을 설계하였다.

1) 사용자 메뉴 구성도



■웹문서, 실시간 뉴스, 비정형 문서, DB 검색 선택 기능
■검색결과 출력방법(제목, 제목+요약), 검색결과수, 순위에 대한 선택 기능

2) 관리자 메뉴 구성도



■질의 처리기, 질의 수행기, 색인관리기, Bridge, 컬렉션에 대한 관리/통제 기능

■간단한 설정으로 관리/통제 가능

6. 시스템 특징

15억, 1억, 5천만(대용량 분산처리 구조) : 코리아와이즈넛은 분산 구조를 바탕으로 한 대용량 처리에 대한 세계 최고의 기술 노하우 보유하고 있다. 현재까지 15억건의 데이터를 처리하는 검색엔진을 구현한 경험이 있으며, 최대 1조건까지 지원 가능하다. 또한 parallel 인덱싱 기술을 통해 최대 일간 1억건까지 색인이 가능하다. 정보 수집의 경우 Multi-threading 기법을토대로 최대 일간 5천만건의 정보 수집 속도를 제공할 수 있다.

높은 검색품질(뛰어난 한글처리, 적합성 알고리즘) : WISEnut Search Formula-1은 실리콘밸리의 뛰어난 기술환경을 기반으로 개발에 착수되었지만, 순수히 한국인에 의해 개발된 국산 검색엔진으로서 한글 처리에 뛰어나다. 한글형태소분석기 및 각종 사전을 통한 자연어검색을 제공하며, 잘못 입력된 질의어에대한 단어추천기

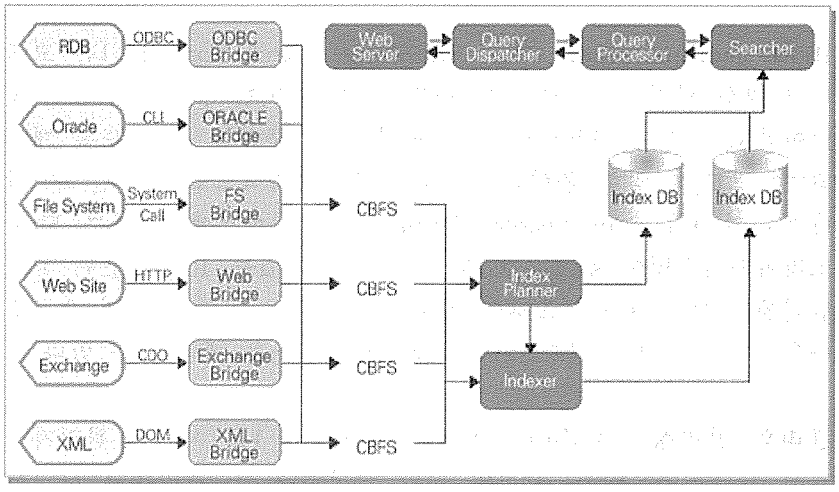
능, implicit한 질의어 확장 기능 등 사용자의 검색활동을 지원하는 각종 기능을 제공한다. 또한 복수 filed에 대한 정렬 및 text-analysis 등을 적용한 진보된 방식의 랭킹 알고리즘을 적용하므로 검색결과의 적합성이 매우 우수하다. 또한 문서 내의 표 등도 추출해내는 비정형문서 filter를 사용하여 더욱 정교한 검색이 가능하다.

100% Customization 가능 : WISEnut Search Formula-1은 고객의 다양한 환경과 요구를 만족시키기 위한어떠한 Customization 작업이라도 최소의 비용으로 빠른 시일 안에 구현할 수 있도록 개발되었으므로, 다양한 플랫폼 및 솔루션과 효과적으로 통합될 수 있다.게다가 WISEnut Search Formula-1은 외산검색엔진과는 달리, 코리아와이즈넛이 source code를 보유하고 있는 100% 순수 국산 검색엔진이다. 따라서 고객이 원하는 경우에는 어떠한 고객 요구도 수용하여 완전 맞춤형으로100% customization을 제공할 수 있다.

관리의 편리성 : WISEnut Search Formula-1은 시스템 관리자를 위해 로그뷰어, XML 방식의 통합 설정파일 등 편리하고 다양한 검색엔진관리 인터페이스를 제공한다. 또한 Query Language, API 등을 제공하여 타 응용프로그램과 손쉽게 연동할 수 있도록 지원한다.

7.시스템 구성

Search Formula-1은 수십억 건의 데이터를 처리할 수 있는 구조를 구현한 기술력을 바탕으로 대용량 데이터 처리에 적합한 분산처리 구조를 설계하였다.



1) 질의처리의 3-tiered 분산구조

WISE Query Dispatcher(이하 Query Dispatcher)와 복수개의 WISE Query Processor(이하 Query Processor)및 WISE Searcher(이하 Searcher)가 3-tiered 병렬 구조로 상호 작동하는 구조이다. 각 서버는 독립적인 서버로서 autonomous하게 작동한다.

Query Dispatcher는 Client 응용프로그램과 Query Processor들 사이에서 작동하며 양쪽과 communication한다. 사용자가 질의를 하면, Query Dispatcher는 Query Processor들의 load를 조절하여 적당한 Query Processor에게 질의 처리를 지시한다. Query Processor는 사용자 질의에 대한 전처리(pre-processing)를 담당하여 한글형태소분석기 및 각종 사전 등을 적용하여 질의어를 parsing한 후, 데이터 컬렉션에 따라 적절한 다수의 Searcher들에게 해당 query를 분배하여 보낸다. Query Processor로부터 검색 요청을 받은 Searcher들은 병렬적으로 동시에 해당 질의를 처리하여 결과를 Query Processor로 보낸다.

인덱스 DB는 Searcher마다 분할되어 관리되며, Searcher들은 해당 인덱스 DB 안에서 문서들을 검색한다. Query Processor에서는 복수의 Searcher가 각각 보내온 검색 결과를 취합하여 전체적인 ranking 및 정렬 등 후처리를 한 후 Query Dispatcher에게 전달하며, Query Dispatcher는 이 결과를 클라이언트 프로그램에 전달한다. 이러한 분산구조 때문에 검색 솔루션의 사용자 수가 증가하더라도 기 구축된 시스템에 대한 영향을 최소화하면서 쉽고 저렴한 비용으로 시스템을 추가할 수 있다.

2) 정보수집/저장의 3-tiered 분산구조

정보 수집/저장 부분도 또한 Index Planner와 복수개의 Bridge 및 Indexer로 구성되어 있다. Bridge들은 다양한 정보원들에 특화된 문서수집기(crawler)이다. Index Planner는 Indexer들을 구동시키고 indexing 프로세스를 조절하는 모듈이다. Bridge들은 복수의 서버 또는 하나의 서버에서 구동할 수 있으며, Indexer 또한 하나의 서버에서 하나의 프로세스로 작동할 수도 있고 대용량 데이터를 처리하기 위해 다수의 서버에서 동시에 실행될 수도 있다.

이러한 분산 병렬처리 구조를 통해 빠르고 유연하게 문서들을 인덱싱할 수 있다. 때문에 다양한 정보원으로부터 방대한 정보들이 수집되고 처리되더라도 성능이나 검색결과의 품질에 영향을 받지 않으며, 조직내 정보들이 증가하거나 새로운 정보원이 추가되더라도 쉽고 저렴한 비용으로 시스템을 추가할 수 있다. 대용량의 복잡한 정보들이 다양한 정보원에 걸쳐 있는 기업일수록 이러한 강점이 반드시 필요하다.

8.주요기능

1) 색인 및 검색 가능한 데이터 유형

- DB 정보 : DB gateway 제공
- Web 정보 : multi-threading 기술을 이용한 고성능 web crawler 제공
- 비정형 문서 : MS Office군, PDF, HWP 등
- Web Server, File Agent를 이용한 개인 데이터 색인 기능

2) 색인 기능

■real-time indexing 및 parallel indexing 지원

■대용량 색인 DB를 재 색인 작업 없이 여러 개의 독립된 색인 DB로 분할하여 검색 및 필드 추가 가능

■대소문자 구별 색인, 한자 한글 변환 색인, 특수문자 색인, 사용자 지정필드 색인 지원

■사용자 사전을 고려한 customization된 색인 기능 지원

3) 검색 기능

■불린 검색, 절단검색,exact phrase 검색(구검색), nearby 검색 지원(ISO Z39.58 사용자 질의규약 준수)

■결과 내 재검색, 자연어 검색 지원

■통합 검색 시 target collection 선택 가능

4) 검색 결과에 대한 다양한 view & action 지원 및 자동분류 기능

■날짜별 혹은 relevancy별 정렬 기능

■summary 요청 및 하일라이트 기능 on/off 지원

■통합검색 시 결과에 해당하는 collection 이름 표시

■다양한 format의 비정형문서 보기 기능

■검색결과 저장 및 메일 송부 기능

5) Security

■분산 구조상에서의 security 강화를 위한 자체 프로토콜 제공

■Collection별, 문서별 사용자 접근 권한 조정

9. 개발단계별 기간 및 투입인원수

개발단계	개발시간	인원	비고
시스템 계획 및 자료분석	2001.02 ~ 2000.02	3	- 프로젝트 구상 - 구축목표 설정 - 구축내용 결정 - 자료 수집
요구사항 분석 및 설계	2001.02 ~ 2001.05	9	- 자료 분석 - 동종 서비스 분석 - 프로세스 설계 - 데이터 베이스 설계 - 접근/관리영역 구축
시스템 개발 및 디자인	2001.04 ~ 2001.08	10	- 데이터 베이스 구축 - 기능구축/디자인 - 접근/관리영역 구축 - 추가 기능 구축 - 시스템 최종 테스트
운영 및 유지보수	2001.08 ~ 2001.08	5	- 시스템 및 운영 지침서 작성 - 기타 매뉴얼 작성 - 시스템 설정 테스트 - 시스템 보완사항 보수
계	총 7개월	27	

10. 사용 시스템과 개

	최저사항	추천사항
Process	Pentium III 1GHz, 1CPU	Pentium III 1GHz, 2CPU
OS	Windows 2000, Unix(Sun Solaris, HP-UX, IBM AIX), Linux 등	
Memory	512MB	1G
HDD	10GB (Crwaling + Indexing)	33GB(Crawling) 33GB*2 (Indexing)
개발언어	C++, ASP, JSP, PHP	