# 고성능 인터넷 검색엔진

한국정보처리전문가협회장상

## 1. SOFTWARE명

### 고성능 인터넷 검색엔진

## 2. 제작자


### 대구대학교 정보통신공학부 컴퓨터응용기술연구실

지 도 교 수 : 이용두, 김희철
개발 참여 학생 : 정원교, 김성훈, 김무권, 최철규
노재운, 전영진, 이영탁

- 주소: 경북 경산시 진량면 내리리 15번지
대구대학교 공과대학 정보통신공학부
- 전화: 053-850-6611(6620), 팩스: (053)-850-6619


인터넷이 급속히 발전함에 따라 정보량이 방대해지고, 그 상태가 불안정한 인터넷상의 문서들을 효율적으로 정보 검색자들에게 제공하기 위해 본 연구실에서는 확장성, 관리의 효율성, 성능면에서 우수한 고성능 검색 시스템을 개발하게 되었다. 본 검색엔진은 단어의 추가와 삭제가 용이하며 검색 속도가 빠른 특성을 가지는 다차원 이진 트리(Multi-dimensional Binary Tree) 방식을 개발하고, 이를 기반으로한 전자사전을 구축하여 다이나믹 인덱싱(Dynamic Indexing)이 가능하며, 검색 처리에 있어서 기존의 방식보다 속도면에서 5배 이상 빠른 특징을 갖는다.

## 3. SOFTWARE 전체 요약 설명

### 3.1. 시스템 개요

검색엔진 시스템은 크게 로봇 모듈, 전자사전 모듈, 색인 모듈, 그리고 사용자 인터페이스 모듈로 구성되어 있으며 그 구성도는 (그림 1)과 같다.

(그림 1)의 시스템을 간단히 살펴보면, 로봇 에이전트는 인터넷상의 서버들의 IP 주소를 담은 사용자 호스트 리스트 파일로부터 대상 호스트를 입

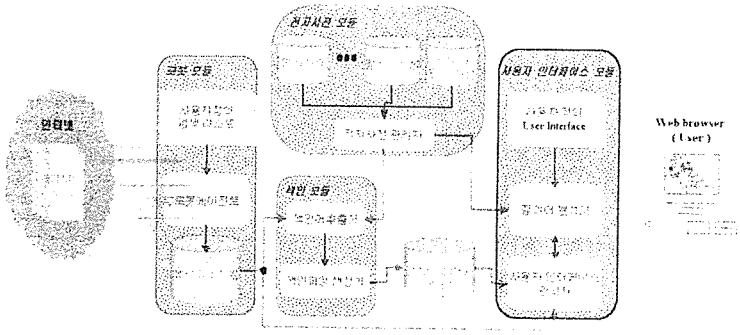



그림 1. 검색엔진의 구성도

받은 후 이들 호스트들에 등록되어 있는 HTML 및 문서자료들을 자동 수집을 하여 데이터베이스를 구축한다. 그리고 나서 이 수집된 자료 데이터베이스에서 색인어 추출과 문서번호를 생성시키게 된다. 이 과정에서 약 200 백만 단어로 구성된 색인어 사전과 사용자가 필요에 따라 등록해 놓은 사용자 정의 색인어사전을 담은 파일의 내용을 읽어 구축한 전자사전을 참고로 색인어를 추출한다. 추출된 색인어에 대하여 가중치가 계산되고 색인파일 생성기를 통하여 색인데이터베이스가 작성된다. 이러한 과정을 마치면 사용자 인터페이스 모듈에서 검색서비스가 가능하게 된다.

사용자 인터페이스 모듈은 사용자의 질의에 대하여 질의분석과정에서 키워드를 추출한다. 이 키워드에 대하여 내부 색인데이터로부터 문서정보를 얻어내고 그 문서정보를 이용하여 데이터베이스에서 해당 문서를 추출하며, 이들은 사용자 화면에 주어진 인터페이스 형식에 준하여 출력된다. 이때 특정한 사용자 인터페이스의 양식을 사용할 수 있도록 시스템 관리자가 제공할 인터페이스 형식을 설정할 수 있다.

## 3.2. 각 모듈의 기능

- **로봇 모듈 구성**

  일차적으로 자료를 수집하고, 가공할 수 있도록 기초자료를 수집하여 인덱싱 모듈에 제공해 주는 모듈이다. 자체적으로 로봇의 검색범위에 대한 데이터 베이스를 보유하고 있기 때문에, 주기적으로 업데이트된 문서에 대해서만 문서를 수집한다. 로봇모듈의 기본 구성은 (그림 2)와 같다.

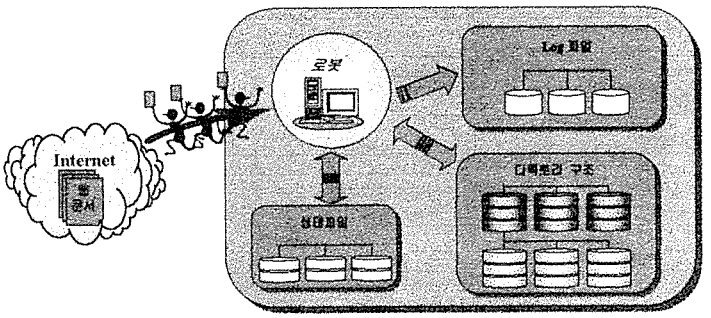



그림 2. 로봇 모듈의 구성

- 로봇 모듈은 업데이트된 문서에 대해서만 인덱싱을 수행할 수 있는 다이나믹 인덱싱에 적합한 형태로 수집된 정보를 저장한다.
- 필요에 따라 관리자는 로봇의 수행 지역을 제한, 또는 추가하여 검색대상을 일차적으로 제한할 수 있다.
- 관리자가 정의한 데이터베이스에 대하여 인덱싱이 가능한 형태로 자료를 수집한다.

## ● 인덱싱 모듈 구성

색인어 추출기는 로봇이 모아온 자료를 (그림 3)과 같은 구조로 색인 작업을 수행하게된다. 색인의 추출은 주어진 문장 속에서 색인어로 사용될 수 있는 단어들만 추출해내야 한다. 이러한 추출방법으로 한글절단방법(Stemming)을 사용하며, 절단 방법을 결정하기 위해 한글 용어 사전, 한글 불용어 사전, 한글 어미, 한글 조사 사전 등을 이

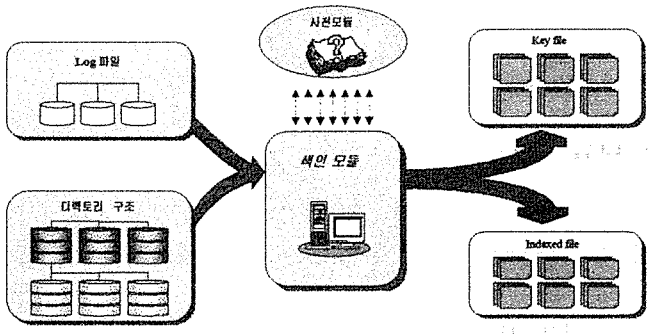



그림 3. 색인 파일의 생성구조

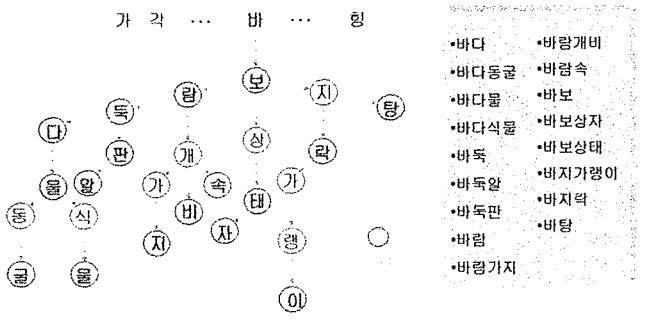



그림 4. 다차원 이진 트리 구조

용한다.

□ 색인 방법

색인어 추출기를 통해 추출된 색인어에 대하여 사전에 등록된 절대 번호를 부여하고, 같은 절대 번호를 가지는 리스트를 작성한다.

□ 동적 색인(Dynamic Indexing)

NEW, UPDATE, DELETE된 문서에 대하여 모든 문서에 대한 색인을 수행하지 않고 NEW, UPDATE, DELETE된 문서에 대해서만 색인을 수행한다.

## ● 사전 모듈 구성

인덱싱 과정은 검색에 필요한 색인어를 추출하는 과정으로, 정확한 색인어의 추출은 검색 시스템의 생명이다. 사전 모듈은 자체 개발한 다차원 이진 트리 구조(그림 4 참조)를 가지고 있으며, 색인과정에서 필요한 색인어에 대한 여러가지 정보를 제공해준다.

본 검색 시스템은 정확한 색인어 추출을 위하여 다음과 같은 여러 종류의 색인에 필요한 사전을 가지고 있으며, 사전모듈은 사용자 질의 처리에 있어서도 사용자의 질의를 정확히 분석하는 기능을 수행한다.

□ 한글 용어 사전
□ 한글 불용어 사전
□ 한글 어미·조사 사전
□ 영어 불용어 사전

- 사용자 사전
- 신조어 사전

## ● 사용자 인터페이스 모듈 구성

사용자 인터페이스 모듈은 앞 절에서 구축한 검색시스템을 이용하여 사용자가 입력한 질의들을 검색엔진의 구조에 맞게 분석하고 다시 조합한다. 특히 검색어들 가운데 색인어 추출이나 한글절단(Stemming) 과정을 필요로 하는 단어들은 절단을 해주고 검색에 필요한 연산자나 괄호 등에 대한 처리를 해주게 된다.

- 질의어 분석기

  사용자가 입력한 질의어를 검색이 가능한 형태로 가공한다. 주어진 문장에 대하여 이 문장들을 모두 처음에 색인하였을 경우와 동일한 색인어로 검색을 수행하면 검색 효율이 높아지게 된다. "대구에 있는 대학교들"이라는 검색문장을 입력하였을 경우 색인할 경우와 마찬가지로 "대구+대학교"라는 검색어로 변환하여 검색을 하면 더 정확한 검색을 할 수 있다. 그러므로 질의어 분석기는 검색 문장으로부터 색인어를 추출한다.

- 사용자 인터페이스 모듈

  사용자가 입력한 질의어에 대한 검색 결과를 여러 가지 검색 옵션에 맞게 가공하여 사용자에게 제공하는 기능을 수행한다. 사용자 인터페이스 모듈은 주어진 사용자의 질의어에 대하여 AND, OR 등의 검색연산자 분석을 통해 검색 옵션을 결정하며 괄호 등을 이용한 복합 검색의 경우 Stack을 이용한 후위연산자 형태로 변환하여 검색을 할 수 있도록 전처리 과정을 수행한다. 이러한 과정을 통해 얻은 키워드에 대하여 색인 데이터베이스로부터 해당하는 문서를 찾아 그 내용을 검색 데이터베이스에서 검색하여 사용자에게 출력한다(그림 5 참조).

- 사용자 인터페이스

  사용자 인터페이스는 널리 사용되고 있는 웹 인터페이스를 이용하고 웹브라우저인 넷스케이프나 인터넷 익스플로어를 사용할 수 있도록 한다. 검색 결과에서 해당하는 결과를 클릭 하였을 경우 세부 페이지 조회로 연결되는데 이 세부 페이지의 내용은 설정된

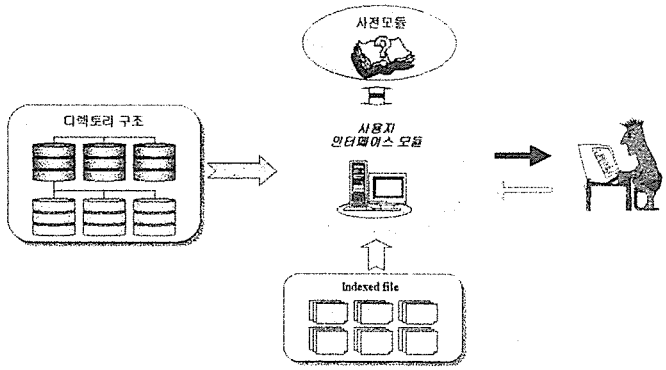



그림 5. 사용자 인터페이스 모듈

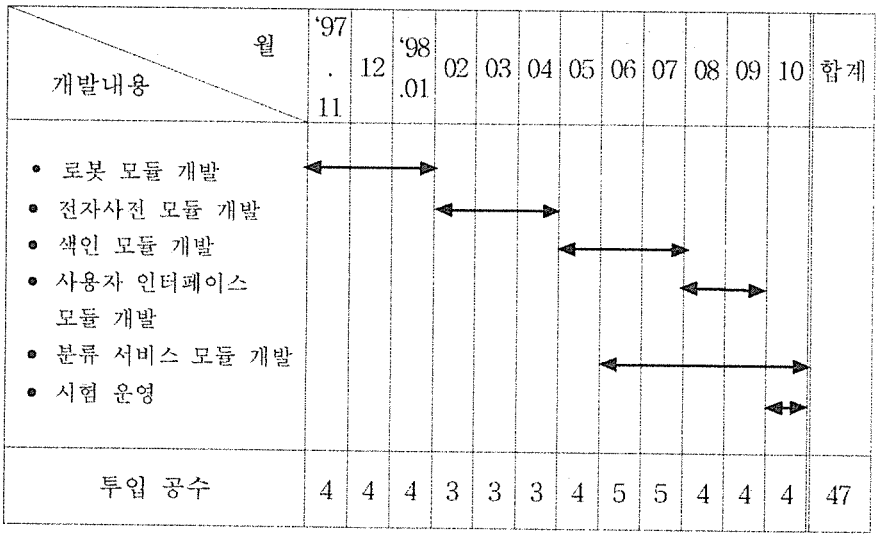
환경에 준하여 로봇에이전트가 생성한 요약파일을 보여주게 된다.

□ 검색 연산자

- 기본 연산자
  AND, OR, ANDNOT, '+' 연산자, NEAR 연산자
- 확장 연산자

  날짜검색, Domain / Host별 검색, 제목내 검색, 검색내 검색

## 4. 개발 단계별 기간 및 투입 공수



| 개발내용 \ 월 | '97.11 | 12 | '98.01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 합계 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| • 로봇 모듈 개발<br>• 전자사전 모듈 개발<br>• 색인 모듈 개발<br>• 사용자 인터페이스 모듈 개발<br>• 분류 서비스 모듈 개발<br>• 시험 운영 | | | | | | | | | | | | | |
| 투입 공수 | 4 | 4 | 4 | 3 | 3 | 3 | 4 | 5 | 5 | 4 | 4 | 4 | 47 |

## 5. 관계 프로그램 수

## 6. 사용 또는 개발 언어. TOOL

- 개발언어 : gcc 2.8.0

## 7. 사용 시스템

- SUN Enterprise 3000
  - O/S: Solaris 2.6
  - Main Memory: 1 GBytes
  - 디스크 용량: 30 GBytes
  - Web Sever: Apache 1.2.7

## 8. 직접 효과

- 효율적인 정보관리 및 이용
- 분산된 정보의 자유로운 접근을 통한 정보활용 능력강화
- 다이나믹 인덱싱 기법을 통한 최신 정보 활용 가능

## 9. 간접 효과

- 인터넷을 이용한 응용 시스템 구축의 기반 기술 확보
- 국내 지능형 에이전트 기술 선진화에 기여
- 국내 지능형 검색 에이전트 기술의 실용화를 앞당겨서 고부가 수출 전략화에 기여
- 인터넷/인트라넷 소프트웨어의 국산화를 통한 수입 대체 효과

## 10. 기타

- **현재 서비스중인 검색엔진의 메인 화면**