웹기반 한글정보검색시스템의 구현

An Implementation of Web-Based Korean Language Information Retrieval System

홍기채(G.C. Hong) 정보유통연구팀 선임기술원
정현수(H.S. Chung) 정보유통연구팀 책임연구원, 팀장

최근 인터넷상에는 매일 방대한 양의 정보가 창출되어 유포되고 있으며, 수많은 정보 제공 사이트들이 늘고 있다. 이용자들은 필요한 정보를 찾고 활용하기 위해 야후(Yahoo), 알타비스타(AltaVista) 등 국외 검색엔진(search engine)들과 심마니, 미스 다찾니 등 국내 검색엔진 등 인터넷상에 운용되고 있는 이들 시스템들을 이용하고 있지만, 대부분의 시스템들은 자체 정보 제공보다는 로봇 에이전트를 이용하여 인터넷 사이트에 등록되어 있는 다양한 분야의 홈페이지 정보들을 수집/분석하여 관련 사이트를 연결해주는 방식의 메타 검색엔진들로서 불필요한 정보들까지 제공함에 따라 이용자들이 필요로 하는 정보를 찾기에는 너무 많은 노력과 시간을 소모하게 되는 문제점을 안고 있다. 이에 본 고에서는 형태소 분석 및 시소러스 사전을 이용하여 검색의 정확성 및 재현율 향상을 고려하고, 주제어 중심의 불리언 검색뿐만 아니라 하이퍼텍스트 기반의 주제어 카탈로그 검색, 각기 다른 사이트의 검색엔진들로부터 질의한 결과를 통합하여 제공하는 지능형 통합검색, 이용자 프로파일에 근거하여 최신 업데이트된 정보를 주기적으로 제공해주는 맞춤정보서비스(Selective Dissemination of Information Service: SDI) 등을 통합한 인터넷 기반의 한글 정보검색시스템의 구현에 대한 내용을 기술하고자 한다.

I. 서 론

PSTN(Packet Switched Telephone Network), PSDN(Packet Switched Data Network) 등 공중망 서비스 환경에서의 정보검색시스템이 주로 텍스트 기반의 자체 보유의 정보를 제공하는 반면에 현재 보편화되어 있는 인터넷을 이용한 웹(World Wide Web: WWW) 기반의 정보검색시스템은 로봇 에이전트를 이용하여 다른 웹사이트의 정보까지 제공해주는 확장성 및 정보소재를 정확하게 알 필요가 없는 투명성을 제공해주고 있다. 이에 웹 기반의 정보검색시스템은 다양한 기능, 즉 관련된 사이트를 방문하여 정보를 수집해오는 로봇 에이전트, 이용자 프로파일에 근거하여 최신 업데이트된 정보를 주기적으로 E-mail 또는 팩스나 웹브라우저 등을 통하여 제공하는 맞춤정보서비스(SDI), 분야별로 정보를 재분류하여 메뉴에서 한 단계씩 정보를 찾아 내려가는 방법을 제공하는 디렉토리 서비스, 이미지 및 원문정보 등 대용량의 정보저장을 위한 효율적인 데이터베이스 구축, 다양한 원문정보의 색인을 위한 각 문서포맷에 대한 필터링 기술, 축적된 정보의 색인 및 검색기능을 제공하는 검색 엔진, 통신환경 및 다수의 접속자 수를 고려한 분산 환경 하의 웹서버 구성 등이 요구되어 진다. 이러한 웹 기반의 정보검색시스템은 단순히 검색성능 향상만을 고려하였던 텍스트 기반의 검색시스템과는 다르게 구성되어야 함을 알 수 있다. 따라서 본 고에서는 이러한 다양한 기능들을 고려한 웹 기반의 한글 정보검색 시스템의

구현에 대하여 살펴보고자 한다.

II. 정보검색시스템

정보검색이란 정보를 분석·가공하여 축적해 놓은 데이터베이스 및 색인정보로부터, 이용자의 정보요구에 적합한 정보를 탐색하여 찾아내는 일련의 과정을 의미하는 것으로, '93년에 웹이 탄생하면서 인터넷 정보검색이란 새로운 분야가 대두되었다. 이것은 웹의 특성인 하이퍼텍스트와 하이퍼링크를 제공하고 윈도우 개념의 유저인터페이스가 단순하여 인터넷으로의 접속이나 넷스케이프사의 네비게이터 또는 마이크로소프사의 인터넷 익스플로러와 같은 웹브라우저 등의 사용법을 알고 인터넷에 접속할 수 있는 환경만 갖추어지면 일반인들도 쉽게 접할 수 있다.

인터넷은 수많은 컴퓨터가 거미줄 같은 네트워크로 연결되어 있고 이들 컴퓨터는 각각 자신들의 주소를 갖고 있는데 자신이 원하는 정보가 들어있는 컴퓨터의 주소에 직접 접속을 해야만 된다. 어느 컴퓨터에 무슨 정보가 들어 있는지 모르는 이상 수많은 컴퓨터들에 직접 연결을 시도해서 일일이 알아내야 한다. 그러나 일일이 모든 컴퓨터를 찾아 다닌다는 것은 불가능한 일이다. 아무리 많은 정보가 있다고 해도 정보를 찾는데 많은 노력과 시간이 걸린다면 대부분의 사람들은 지쳐서 포기하고 말 것이다. 바로 이런 일을 대신 해주는 도구를 검색엔진이라고 한다. 이와 같이 인터넷의 가장 큰 특징인 정보의 양이 방대하고 다양함에 따라 이용자들이 특정정보를 쉽게 찾아 제공할 수 있도록 하는 도구가 필요하게 된다. 정보를 수집·분석·가공하여 데이터베이스로 축적하고 축적된 정보를 쉽게 찾을 수 있도록 색인(indexing)하고, 색인된 정보를 바탕으로 시스템의 이용자가 정보에 대한 요구를 할 때 검색엔진을 이용하여 적합한 정보를 검색하여 제공하는 시스템을 정보검색시스템(information retrieval system)이라고 한다.

정보검색시스템은 특성 또는 방법에 따라 여러 형태로 분류될 수 있으나, 정보를 수집하고 제공하는 서비스 형태에 따라 분류하면 로봇 검색엔진과 주제별 카탈로그 검색엔진, 그리고 키워드형 검색엔진으로 분류할 수 있다.

1. 로봇 검색엔진(Robot Search Engine)

웹 에이전트(agent)라는 로봇(robot)을 이용해 자동으로 웹 페이지와 관련된 링크를 가져와 추가하는 방식으로 가장 보편화된 방식이다. 로봇은 인터넷 사이트에 산재해 있는 자료를 정리해 검색자료를 구성, 추가하고 사용자가 검색을 의뢰하면 검색한 결과를 되돌려 주는 방식이다. 하지만 어디까지나 로봇이 선별하는 만큼 정확한 검색결과를 기대하기는 어렵다.

웹 에이전트는 스파이더(spider), 로봇, 크롤러(crawler), 웜(worm)이라고 불리는 자동화된 로봇이라는 소프트웨어를 인터넷상에 있는 각각의 사이트에 보내 한 URL(Unified Resource Locator)에서 다른 URL로 돌아다니며 일반에게 공개된 웹의 모든 사이트를 방문하고 주소를 기록하여 자세한 웹의 정보를 수집한다. 그 다음엔 검색엔진이 검색된 제목이나 모든 텍스트 내용을 가지고 데이터베이스를 만들고 색인을 하는 것이다. 대표적인 로봇 검색엔진으로는 알타비스타나 웹크롤러(WebCrawler) 등이 있으며, 국내에는 정보탐정 등이 이에 속한다.

2. 주제별 카탈로그 검색엔진

주제별 카탈로그 검색엔진(subject catalogue search engine)은 사람이 일일이 관련 사이트를 방문해 보고 이들 정보에 대한 분류를 통해 데이터베이스화 및 색인하여 다단계 메뉴형태로 정보를 제공하는 것이다. 예를 들면 인터넷에 있는 정보를 사회, 문화, 예술, 스포츠, 정치 등 주제에 따라 대분류, 중분류, 소분류 등의 메뉴로 구성할 수가 있다. 주제별 카탈로그는 해당 주제에 해당하는 각종정보를 목록으로 제공하기 때문에 디렉토리 서비스, 주제별 검색엔진, 메뉴검색, Subject-oriented searching 등

으로 부르기도 한다. 장점은 찾고자 하는 것에 대하여 특정한 주제어, 키워드, 중심어 등을 알기 힘들어도 대분류 정도만 알 수 있다면 정보를 찾을 수 있다는 것이고, 단점은 원하는 정보를 얻기까지 여러 단계를 거쳐야 하므로 중간에 분류를 잘못 선택하면 찾는 것과 거리가 먼 내용만 찾을 수도 있다.

주제별 카탈로그 검색엔진의 대표적인 시스템은 분류 검색 엔진의 시조라고 알려져 있는 야후와 국내 사이트를 전문으로 안내하는 코-시크(Kor-Seek)와 한메소프트에서 운영하고 있는 ZOOM이 있다.

3. 키워드형 검색엔진

키워드형 검색엔진(keyword search engine, word-oriented searching)은 특정한 키워드를 입력하여 그에 해당하는 정보를 찾아내는 방식의 검색엔진으로, 어떤 내용을 갖추고 있는지에 따라 키워드형, Front-End 형, 지능형 검색엔진으로 나눌 수 있다.

키워드형 검색엔진은 알타비스타, 리이코스, 웹크롤러 등과 같이 로봇이나 스파이더를 이용하여 인터넷 자원을 데이터베이스로 구축해 놓은 검색엔진으로, 검색어 입력상자에 키워드를 입력한 다음 검색버튼을 눌러 원하는 정보를 찾도록 한다. 장점은 몇 개의 키워드, 즉 검색어를 통하여 원하는 정보를 신속하게 찾을 수 있으며, 단점은 색인이 정확하지 않은 때나 자료를 분류해 놓은 후 해당 내용이 변경되었을 때에는 원하는 정보를 찾을 수 없는 경우도 발생한다.

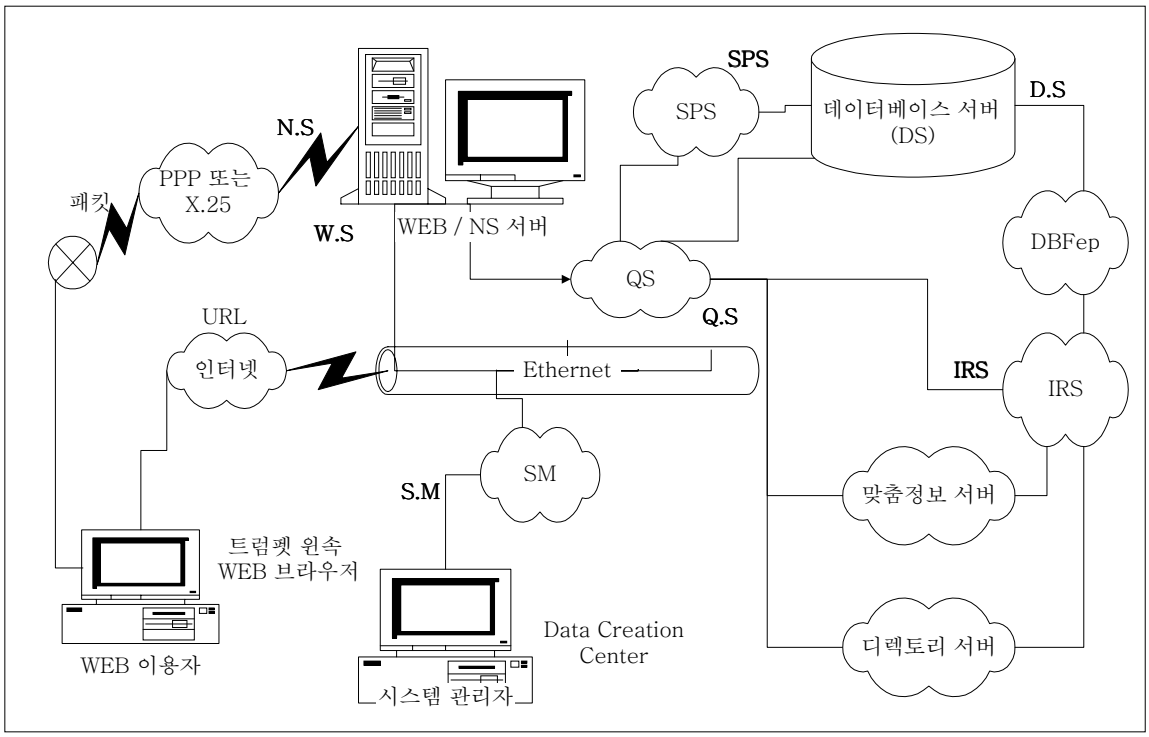
Front-End 형 검색엔진은 여러 개의 일반 키워드형 검색엔진을 한 곳에 모아놓은 것으로, 각각의 검색엔진에 일일이 접속하지 않고도 한 화면에서 이들을 이용할 수 있다. Front-End 형 검색엔진은 메타 검색엔진이라고도 하며, 자신은 정보를 가지고 있지는 않지만 찾고자 하는 정보를 각각의 검색엔진들에게 의뢰한 후 그 결과를 종합하여 이용자에게 제공해준다. 장점은 검색자료의 특징이 다양한 각각의 검색엔진을 옮겨 다니면서 검색할 필요없이 한 화면 안에서 각각의 검색엔진을 이용하는 것처럼 종합된 결과를 얻을 수가 있고, 단점은 자신의 고유 데이터베이스를 갖지 않을 뿐만 아니라 각각의 검색엔진에서 사용할 수 있는 여러 가지 검색옵션을 모두 지원해주는 것이 아니므로 복잡한 질의나 세부적인 내용까지 정교한 검색을 하기는 어렵다는 것이다. Front-End 형 검색엔진은 메타크롤러(MetaCrawler)나 새비서치(SavySearch)가 있으며, 국내에는 미스 다찾니가 대표적이다.

지능형 검색엔진은 자기자신은 데이터베이스를 가지지 않는다는 점에서 Front-End 형 검색엔진과 동일하다. 그러나 Front-End 형 검색엔진은 단순히 여러 개의 검색엔진을 정리/분류하여 한 곳에 모아놓은 반면 지능형 검색엔진은 로봇 에이전트를 활용하여 다른 검색엔진들을 참조한 후 정보를 직접 찾아주고 그 결과까지 보여준다.

즉, 사용자는 한 번만 키워드를 입력하고 검색버튼을 누르면 지능형 검색엔진이 라이코스, 알타비스타, 웹크롤러 등에 검색을 의뢰하여 정보를 찾아 제공해주는 것으로, 통합검색엔진 또는 Search the search engine 등으로 불리기도 한다. 장점은 한 번의 키워드 입력만으로 다양한 검색엔진을 참조하여 검색을 하므로 간편한 정보 찾기와 다양한 검색엔진에서의 정보를 얻을 수 있으며, 단점은 여러 개의 검색엔진을 참조하게 되므로 검색속도가 느릴 때도 있으며, 수 개의 검색엔진에서 찾은 결과가 제공되기 때문에 원하는 정보를 가려내기가 어려울 수도 있다.

III. 시스템 구성

웹 기반의 한글정보검색시스템은 (그림 1)과 같이 웹 서비스를 지원할 수 있도록 NCSA 또는 CERN 등의 웹 서버(WEB Server: WS)를 설치하고, 모뎀 또는 LAN(Local Area Network) 환경을 이용하고 있는 다계층의 이용자들을 고려한 네트워크 서버(Network Server: NS)를 구성하여야 한다. 또한 웹 브라우저와 웹 서버의 인터페이스 역할을 위한 여러 기능을 처리하는 CGI 및 HTML 문서 제공 등을 위



(그림 1) 정보검색시스템의 기능별 구성요소

해 질의 서버(Query Server: QS) 구성, 그리고 효율적인 정보의 저장 및 관리를 위한 데이터베이스 서버(Database Server: DS) 등도 구성해야 한다. 특히 이용자들이 손쉽게 특정 정보를 찾을 수 있도록 정보검색 서버(Information Retrieval Server: IRS)를 개발 또는 기존 제품을 이용하여 정보를 색인하고 검색할 수 있는 환경 구축은 웹 기반 한글정보검색 시스템의 핵심요소라고 할 수 있다. 이외에도 이용자들에게 수시로 변경되는 최신의 관련정보를 직접 E-mail 등을 통해 제공해 주는 맞춤정보서버, 주제별 카탈로그화된 정보를 디렉토리 개념으로 제공하는 주제별 카탈로그 서버 등의 개발이 필요하다. 마지막으로 시스템 운영 측면의 이용통계 현황 등을 목록화할 수 있는 통계정보서버(Statistics Processing Server: SPS), 데이터베이스 및 이용자 관리를 위한 시스템관리 서버(System Management Server: SMS) 기능 등이 추가로 필요하며, 각각의 기능에 대해 살펴보면 다음과 같다.

1. 네트워크 서버

네트워크 서버는 이용자들이 시스템에 접속할 수 있도록 하는 통신 지원 서버로서, PSTN을 통해 PSDN에 접속할 수 있는 X.25, PPP(Point to Point Protocol), Ethernet TCP/IP(Transmission Control Protocol/Internet Protocol) 환경을 구성하여 모뎀 또는 LAN을 이용하는 불특정 다계층의 이용자들을 지원할 수 있어야 하며, 이들 각각은 관련 장비 및 소프트웨어를 별도로 설치하여 운영된다. X.25는 CCITT에 의해 제안되어 졌으며, ISO의 처음 3 layer 즉, physical, data link, network layer를 정의하고 있고, LAN이나 이에 대응하는 WAN(Wide Area Network)을 통한 통신을 위해 사용되어 진다. X.25는 보통 PSDN에 연결되고, 이러한 네트워크는 X.25 패킷의 원거리 통신에 대한 신뢰성을 제공한다. PSDN은 보통 회원(membership)과 연결비용에 더하여 X.25의 각 kbyte에 대해 비용을 청구한다. 따라서 국내의 경우는 Hinet-P 또는 천리안, 나우누

리 등의 전용선을 추가 설치하여 회선비용을 지불하여야 하며, 시스템 측면에서는 인터넷 통신환경을 접속할 수 있는 Trumpet winsock 등 윈도우용 소켓 프로그램을 개발하여 지원해주어야 한다. PPP는 직렬접속을 통하여 데이터를 전송하는 프로토콜로 TCP/IP 네트워크에 전송되어 지는 데이터를 직렬라인(모뎀 + PSDN)으로 변경하여 전송해주는 일종의 약속으로 X.25를 이용하는 통신망 환경보다는 이용자 측면에서 훨씬 편리하게 인터넷과 접속할 수 있다.

2. 웹 서버

웹 서버라 하는 것은 하이퍼텍스트 문서의 송수신을 위한 HTTP 프로토콜을 이해하고 이에 따라 요청받은 동작을 수행하는 하나의 실행 프로그램으로, 이러한 요청을 인식하기 위해서는 서버 프로세스가 항상 살아있어야 하며 이것이 HTTPD(Hypertext Transfer Protocol Daemon) 이다. 서버를 설치한다는 것은 서버 프로그램인 httpd 프로세스를 실행시키는 일이다. 서버 프로그램이 실행될 때는 서버의 서비스 환경에 대해 설정해 놓은 파일들을 참조하므로 적절하게 환경을 설정하여야 한다.

웹 서버는 운영하는 시스템의 플랫폼 즉, Unix, 매킨토시, 노벨 넷웨어 서버, 마이크로소프트 윈도우, 윈도우즈 NT 서버 등에 따라 분류할 수가 있으며, Apache, CERN httpd, NCSA httpd, EIT httpd, GN Gopher/http, Plexus Perl Server, WebWorks Enterprise Server, Netsite Communication Server and Netsite Commercial Server 등이 있다. 국내에서는 NCSA HTTPD 서버를 가장 많이 이용하고 있다.

- HTTPD

Httpd는 이용자가 웹 브라우저(NETSCAPE, Internet Explorer 등)를 이용하여 HTML(Hyper-Text Markup Language) 도큐먼트 또는 CGI(Common Gateway Interface)를 요구하면, 이를 분석하여 HTML 도큐먼트 풀로부터 필요한 HTML 문서나 CGI 프로그램을 실행하여 이용자에게 송신하는 기능과 HTTPD가 이미 가지고 있는 CGI 서비스(MIME, ftp, gopher 등)를 실행한다.

- CGI

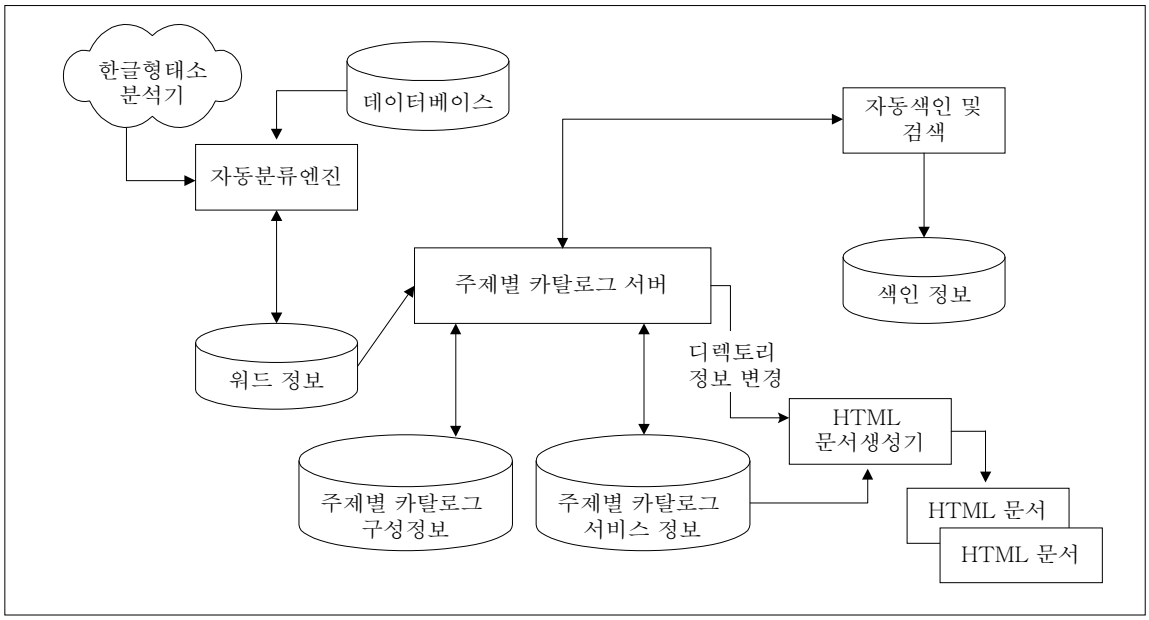
CGI는 HTML 문법을 이용하여 작성된 입력 필드의 값을 웹 서버로부터 받아서 이에 적합한 질의를 작성하여 데이터베이스 서버로 서비스를 요구하며, 데이터베이스 서버로부터 받은 결과를 HTML 문서로 만들어서 웹 서버에게 전달한다. 또한 서비스 및 이용자 통계관리를 위하여 필요한 로그를 남긴다.

3. 질의 서버

질의 서버는 웹 서버로부터 클라이언트의 요청인 HTML 문서 또는 CGI를 처리하여 웹 서버에 전달하는 기능으로, 예를 들면, URL, CGI, HTML을 지정하거나 전달하는 부분 등이 모두 포함된다. CGI는 웹 서버가 다른 프로그램을 실행시키고 이의 출력을 웹 브라우저에게 전송할 텍스트나 그래픽, 오디오에 포함시키는 역할을 하는 서버에서 처리되는 프로그램이다. 서버와 CGI 프로그램은 웹의 능력을 향상시키고 사용자의 요구를 만족시키기 위해 표준 인터페이스를 제공하며, CGI 프로그램은 개발자가 매우 다양한 도구를 사용할 수 있도록 해준다. CGI 프로그램은 양식을 처리하거나 데이터베이스의 한 레코드를 검색할 때, 전자우편을 보낼 때, 움직이는 페이지 카운터를 구성할 때 등 수많은 요구사항을 처리한다. CGI가 없다면 웹 서버는 정적인 문서들과 다른 페이지나 서버에 대한 링크만 제공하게 될 것이다. CGI가 있음으로 웹이 살아 움직이게 되는데, 사용자와 상호 작용하고, 이를 통해 정보를 얻을 수 있게 되는 것이다.

4. 데이터베이스 서버

데이터베이스 서버는 웹 서버로부터 정보 요청을 받아서 해당 질의를 생성한 후 데이터베이스 시스템으로부터 정보를 가져와 웹 서버에게 넘겨주는 기능을 가지고 있다. 데이터베이스 시스템은 일반적으로 Sybase, Oracle, Informix 등 관계형 데이터베이스



(그림 2) 주제별 카탈로그 서버의 구성도

시스템 또는 Object store 등 객체지향 데이터베이스 시스템을 모두 이용할 수 있으며, CGI 프로그램과 데이터베이스 시스템과의 인터페이스는 보통 일반적인 프로그램 등과 거의 유사하게 처리된다.

5. 정보검색 서버

정보검색 서버는 정보검색을 하기 위한 명사, 동의어, 불용어, 조사 사전 등을 가지고 있으며, 실제로 웹 서버로부터 질의 요청이 왔을 때 해당 질의에 대한 인덱스를 웹 서버에게 넘겨주는 기능을 가지고 있다.

정보검색 서버가 가지고 있는 검색기능으로는 제목 등의 간략정보를 보여줌과 동시에 제목별 검색 등을 쉽게 할 수 있도록 하는 쉬운 검색, 사전식 검색을 지원하는 사전 검색, 하이퍼텍스트 개념의 검색을 지원하는 하이퍼텍스트 검색, 불리언 연산에 의한 고급 검색 등을 구성할 수가 있다.

6. 주제별 카탈로그 서버

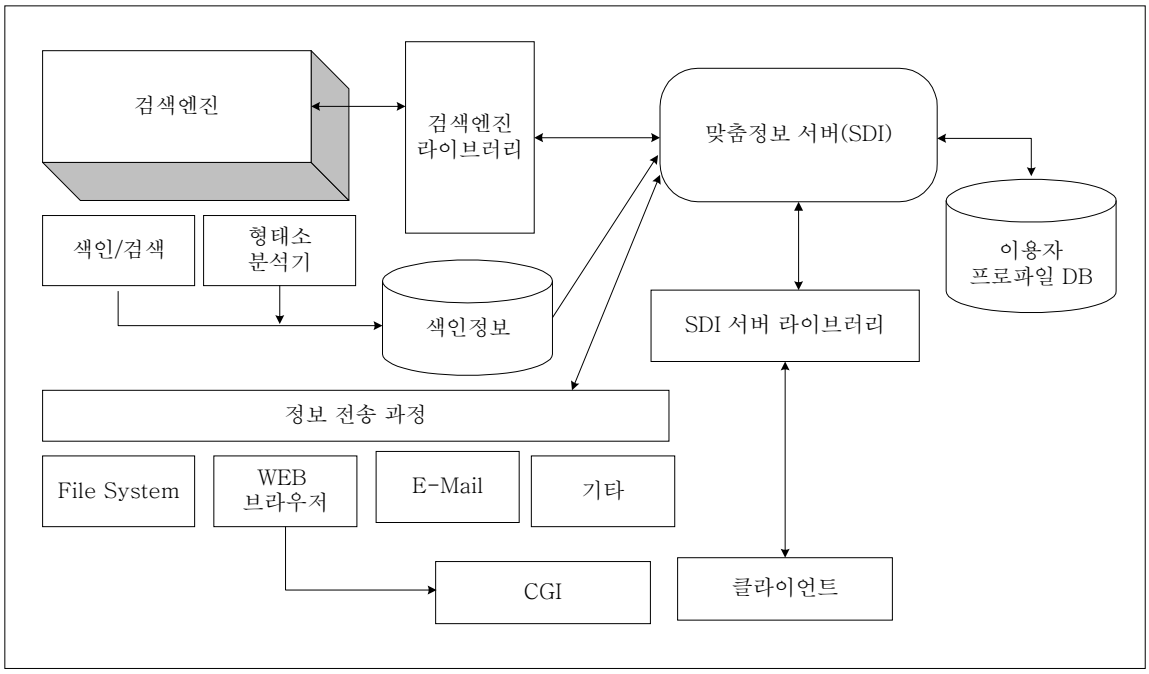
해당 주제에 해당하는 각종 정보를 목록으로 제공하는 서버로 이용자들에게 단계별로 특정 정보를 찾아가게 하여 일목요연하고 통합된 분류서비스를 제공한다. 검색엔진을 이용하여 정보들에 대해 미리 해당 주제별로 목록화하는 과정이 필요하다. 또한 주제별 카탈로그 구성에서 이용된 카탈로그는 대표단어, 가중치를 이용하여 문서를 자동분류하도록 하며, 자동분류 방법으로는 주로 코싸인 계수방법을 이용하고 있다(그림 2).

7. 맞춤정보 서버

맞춤정보서비스는 이용자들의 요청정보, 제공형태 등으로 구성되는 프로파일을 미리 받아 주기적 또는 실시간으로 신규 정보를 이용자들이 전자우편 또는 웹을 통해 제공해주는 서버이다(그림 3). 이는 이용자들에게는 정보를 찾는 시간을 절약하고 최신 정보를 빠짐없이 찾아주게 하며, 시스템 측면에서는 다수의 이용자가 정보에 대해서 유사 또는 중복성의 질의를 요구하는 오버헤드를 줄이는 장점이 있다.

8. 통계정보 서버

통계정보 서버는 시스템의 이용에 관한 각종 통계 정보를 처리하는 서버로, 웹 서버로부터 요구된



(그림 3) 맞춤정보 서버의 구성도

이용통계 및 현황에 대한 통계처리를 하며, 특성상 별도의 기능을 갖도록 구성해야 한다.

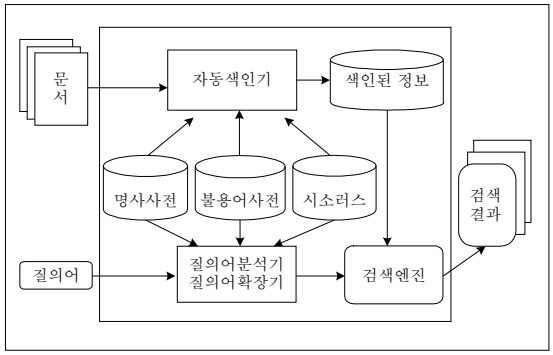
9. 시스템관리 서버

시스템관리 서버는 데이터베이스를 이용하여 각종 정보를 저장하고, 변경 등을 수행할 수 있는 정보관리센터(Information Management Center)의 기능을 갖는 DB 관리 부분 및 시스템에 등록된 이용자 정보를 변경(등록, 변경, 삭제)하는 기능을 갖는 이용자 관리 부분이다.

IV. 웹 기반 한글정보검색시스템 구성

웹 기반의 한글정보검색시스템은 인터넷 상에서 각종 정보 데이터들에 대하여 수집 및 분석을 통하여 구축된 데이터베이스로부터 이용자의 정보요구를 가장 적합한 형태의 정보로 신속, 정확하게 탐색하여 제공하는 것으로, 시스템은 크게 두 부분 즉, 자동색인기 부분과 검색엔진 부분으로 나뉘어진다.

1. 검색시스템의 구성



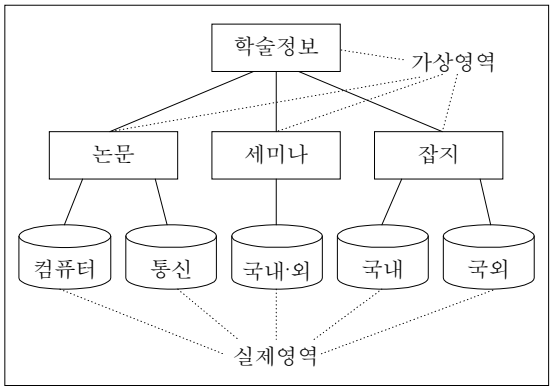
(그림 4) 검색시스템의 구성도

2. 검색시스템의 기능

검색시스템은 다음과 같은 주요 기능을 제공하도록 구성한다.

가. 전문 자동 색인과 전문 검색

검색시스템은 전문을 대상으로 형태소 분석기를 사용하여 불용어를 제거하고, 정보를 대표할 만한



(그림 5) 검색시스템의 인덱스 영역

키워드 즉, 색인어를 자동 추출하는 색인 작업을 수행할 뿐만 아니라 전문을 대상으로 검색을 실행한다.

나. 재검색 및 자연어 처리 기능

사용자의 질의어 검색 시 방대한 검색결과 안에서 다시 질의 검색하여 사용자가 원하는 정보를 좀 더 정확하게 얻을 수 있도록 하며, 사용자의 질의어 작성 시 자연어로 질의를 작성하여 검색할 수 있도록 한다.

다. 다양한 검색 알고리즘

불리언 검색은 물론 절단 검색, 시소러스 참조 검색, 제한 검색, 필드 조합 검색 등 정형 데이터와 비정형 데이터의 통합 검색이 가능하며, 정보 검색의 효율성을 높이기 위하여 탐색어 열람 기능, 시소러스 참조 기능 등도 제공해야 한다.

라. 질의어 가중치 기능

사용자의 질의어가 출현하는 단어마다 가중치를 주어서 좀더 의미있는 정보에 우선순위를 두어 검색결과의 앞부분에 나오게 한다.

마. 가중치 검색기법 제공

이용자가 입력한 질문과 저장된 정보 자료와의 유사도를 계산하여 질문에 대해 적합성으로 순위를 매겨 출력해 준다.

바. 개발환경 지원

다양한 API(Application Programming Interface) 라이브러리를 제공하여 개발환경을 지원하고, 사용자가 개발할 수 있는 기능을 제공하여 타 시스템과 통합할 수 있도록 한다.

사. 인덱스 영역의 가상적인 계층구조 지원

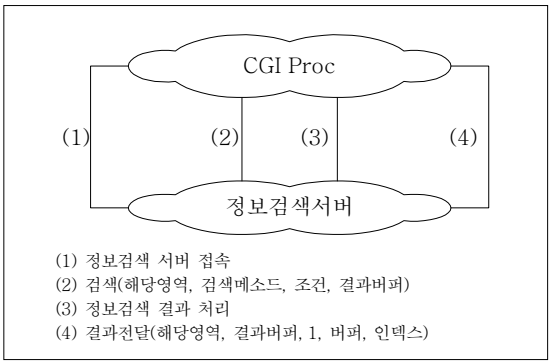
가상적인 계층구조에서 최하위 단말노드에는 실질적인 영역 DB가 있고, 그 위의 노드들은 가상적인 영역으로 실제적인 영역 DB들을 관리한다. 여기서 영역이라 함은 비슷한 구조를 가진 문서들을 모아 놓은 디렉토리와 같으며, (그림 5)에서 단말노드의 컴퓨터, 통신 등의 영역은 실제 영역 DB로 인덱스 정보를 가지고 있으며, 상위의 논문, 세미나, 잡지 및 루트의 학술정보 등은 가상영역 DB로 하위노드들을 그룹화시킨 것으로 실제적인 인덱스 정보는 갖지 않는다. 이러한 가상영역 DB를 구성하도록 하는 것은 그룹 또는 전체적인 통합검색을 지원하기 위함이다.

3. 검색시스템의 인터페이스

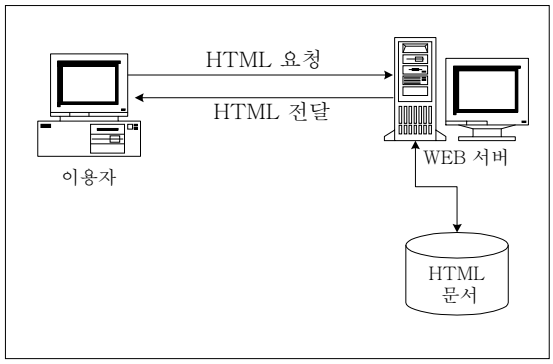
정보검색시스템은 DB 인터페이스(DB Front End Process: DBFep)를 이용하지 않고, CGI에서 바로 검색을 하도록 구성할 수 있다. 즉, 정보를 색인할 때 CGI에서 출력할 내용을 Description으로 저장해서 실제 검색 시 DB를 검색하지 않게 되므로 검색속도가 향상된다. 정보검색 서버와 CGI Proc와의 인터페이스는 (그림 6)과 같다.

4. 검색의 과정

정보검색시스템은 이용자가 주제어 입력 시, 입력된 주제어가 질의 서버를 통해 정보검색 서버에 전달된 후 검색 서버의 검색엔진은 필요한 환경을 설정하고, 입력된 주제어를 검색엔진이 처리할 수



(그림 6) 정보검색서버와 CGI Proc 의 인터페이스



(그림 7) HTML 문서 서비스 흐름도

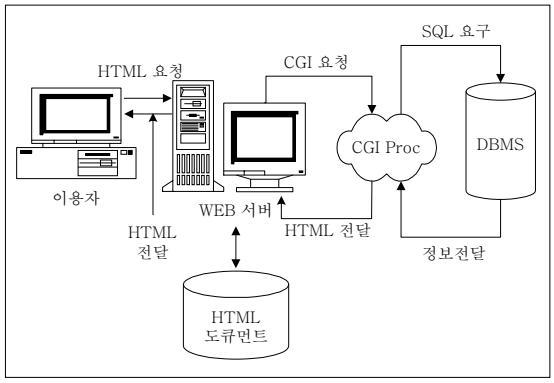
있는 질의어로 바꾼다. 이렇게 바뀌어진 질의어에 따라 검색엔진을 이용하여 검색한 다음 결과를 질의 서버에게 전달함으로써 질의 서버는 이용자에게 HTML 문서형태로 전달하게 된다. 또한 서비스 내용은 순수한 HTML 문서만을 제공하거나 데이터베이스에 질의를 보내 결과를 제공하는 검색서비스 I 및 데이터베이스와 검색엔진을 이용하여 결과를 제공하는 검색서비스 II 등의 서비스 형태로 분류할 수 있다.

가. HTML 문서 서비스

HTML 문서 서비스는 HTML 형태의 문서가 디렉토리 단위로 관리되기 때문에 검색시스템의 웹 서버가 (그림 7)과 같이 직접 HTML 문서 풀로부터 직접 액세스한다. 예를 들면 소식/안내, 질의/응답 등 비교적 간단한 내용은 정보를 데이터베이스에 올리지 않고, HTML 형태의 문서파일로 문서 풀에 저장한다.

나. 검색서비스 I

검색서비스 I의 형태는 서비스의 내용이 데이터베이스에 구축되어 있어서 이의 검색을 위해서는 SQL(Structured Query Language) 문을 이용하여야만 한다. 이때는 (그림 8)과 같이 질의어를 SQL 문으로 만들어서 데이터베이스에 처리를 요구하며, 그 결과를 요구에 따라 필요한 레코드 단위를 HTML 형태의 페이지로 만들어서 서비스할 수 있도록 한다.



(그림 8) 검색서비스 I 흐름도

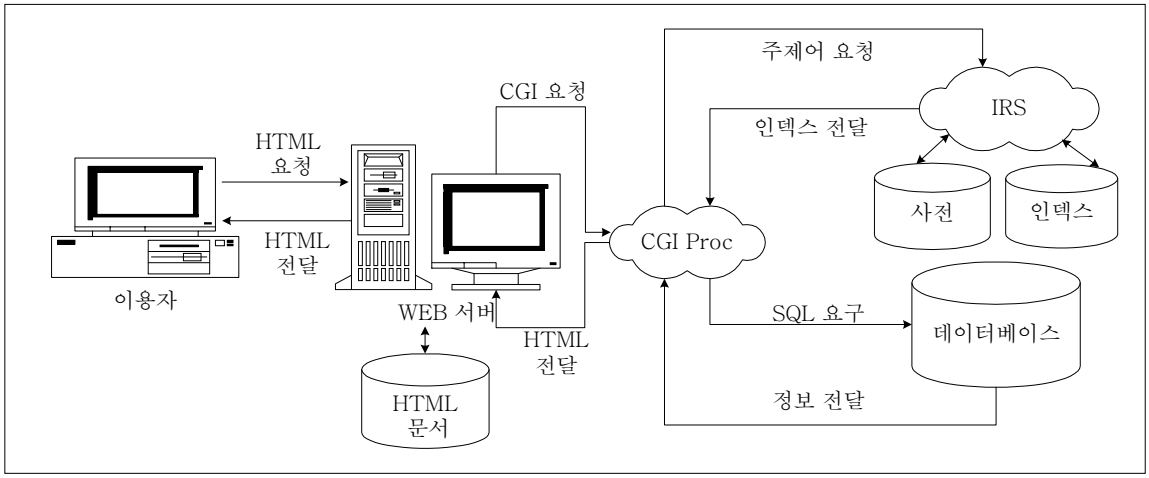
다. 검색서비스 II

검색서비스 II는 서비스의 내용이 (그림 9)와 같이 데이터베이스에 구축되어 있으며, 정보에 대한 전문(full-text) 검색을 이용하여 검색을 더욱 편리하게 한다.

5. 자동색인의 과정

정보검색을 위한 자동색인 과정을 보면, (그림 10)과 같이 한 레코드별 정보를 저장하고, 각 해당 레코드에서 단어를 읽은 다음 어근 추출, 불용어 처리, 각 단어의 정보를 저장하는 형태로 모든 레코드에 대해 반복하여 필요한 정보를 저장한다.

색인 과정을 거친 모든 정보, 즉 인덱스 파일 등은 각 DB별로 나누어서 각각의 영역(domain)에 저장되고 관리된다.



(그림 9) 검색서비스 II 흐름도

따라서 검색시스템은 이러한 색인 자료인 인덱스 파일 외에도 환경설정 파일, 영역을 관리하기 위한 영역관리 파일, 어근 추출 등을 위하여 이용되는 사전파일 등의 정보로 구성하여야 한다. 시스템은 영역관리 파일을 통해서 각 DB에 해당하는 영역에 관한 정보와 영역 자체를 관리하며, 모든 영역에 관한 정보는 영역의 이름, 영역의 ID, 영역이 사용하는 불용어 파일의 이름 등이다. 사전파일은 한글의 어근 추출을 위한 한글 명사 사전, 동사, 형용사 사전, 어미 사전, 조사 사전, 대명사 사전과 영어의 불용어 처리를 위한 영어 불용어 사전 등을 구성해야 한다.

가. 한글 명사/동사/형용사 사전

한글 명사 사전은 시스템에서 정보를 자동 인덱싱하거나 질의어를 분석할 때 이용하며, 동사 및 형용사 사전은 불필요한 인덱싱을 하지 않기 위하여, 즉 동사와 형용사 같은 경우를 문장이나 파일 내에서 찾아내기 위하여 이용된다.

나. 한글 조사/어미/부사/대명사 사전

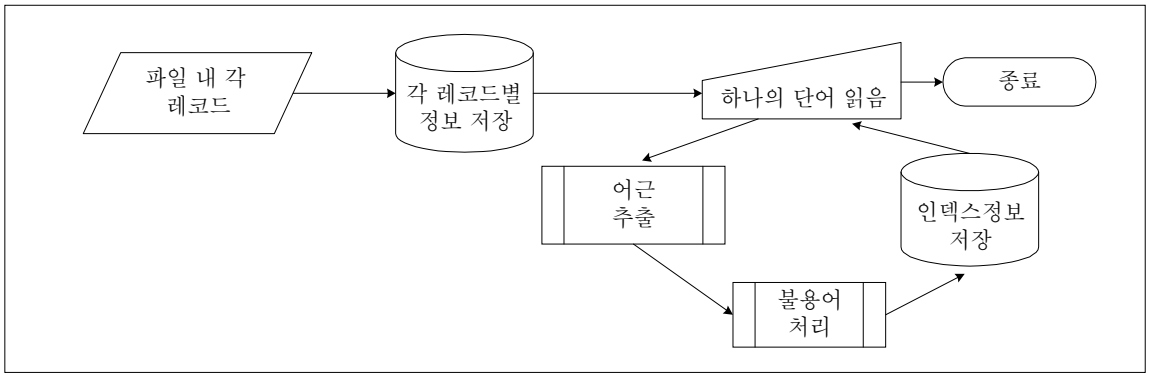
한글 조사 사전은 명사를 추출하기 위하여, 어미 사전은 동사나 형용사를 추출하기 위하여 이용되어지며, 부사 및 대명사 사전은 한글에서의 불용어 사전이라고 볼 수 있으며, 부사나 대명사는 부사와 대명사 사전을 이용하여 색인과정에서 제외된다.

다. 불용어 사전

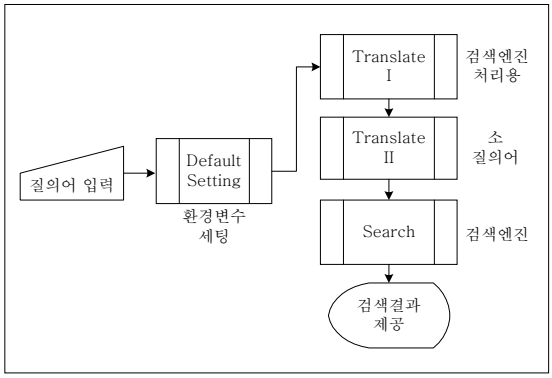
불용어란 용어가 아닌 단어를 의미하며, 시스템에서 정보를 자동 인덱싱할 때나, 질의어를 분석할 때 무시하는 단어이다. 불용어 사전은 불용어의 리스트를 가지고 있는 것으로, DB의 내용들이 서로 다르기 때문에 모든 DB별 각각의 영역마다 하나의 불용어 사전을 가지고 있다.

라. 동등어 사전

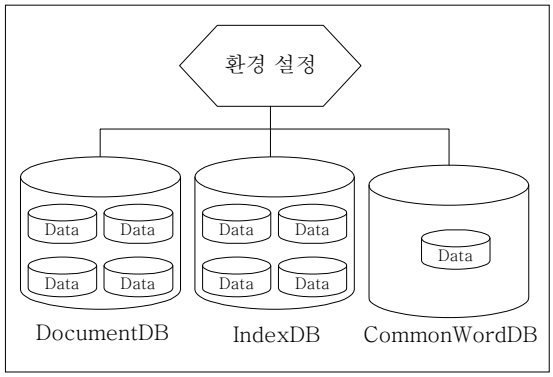
동등한 단어군에 속하는 주제어들을 모아놓은 사전으로, 검색 효율을 높이는 데 이용된다. 또한 인덱스 파일은 자동색인 과정 중에 정보를 저장하면서, 검색을 위하여 만든 파일로 시스템에서 관리하며, 데이터베이스 내에 존재하는 모든 파일들의 정보, 데이터베이스 내에 존재하는 모든 단어들의 정보, 데이터베이스 내에 존재하는 각각의 단어들에 대하여 단어들이 존재하는 파일정보 및 위치정보, 데이터베이스 내에 존재하는 각각의 단어들에 대한 부가정보, 데이터베이스 내에 각각의 블록들이 사용중인지 아닌지에 관한 정보, 파일의 삭제와 재저장(update)을 위하여 이용되는 정보, 하나의 파일에 여러 개의 섹션정보 등이 존재하며, 각 섹션에 따라 검색



(그림 10) 자동색인의 흐름도



(그림 11) 검색엔진의 수행과정



(그림 12) 환경 설정

이 가능하다. 또한 가상영역과 실제영역에 대한 정보를 저장하고 있으며, 특정한 영역에 대한 정보의 검색이 가능하며, 가상영역이 존재하여 여러 개의 영역을 동시에 검색할 수도 있다.

6. 검색엔진의 수행과정

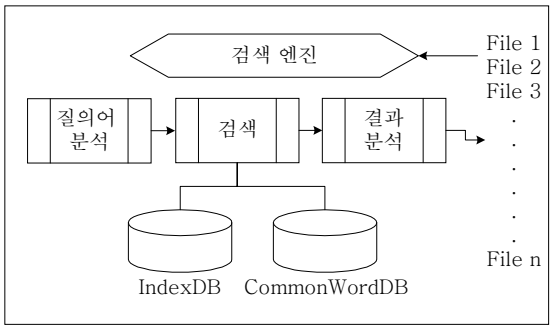
검색시스템의 검색엔진은 단어의 개수나 사용자 인터페이스, 질문의 복잡도, 그리고 질의어 내의 각 주제어의 분산 정도에 따라 검색속도나 정확도에 약간의 차이가 있을 수 있으나, 비교적 빠른 검색을 위하여 설계가 되어야 하며, 수행과정은 (그림 11)과 같다.

가. 질의어 입력

질의어는 주제어 또는 검색키의 내용에 따라 AND, OR 등의 불리언 연산을 이용하여 입력하도록 하며, 예를 들면 “이동통신 AND 미국” 또는 “이동통신 OR 미국” 등의 형태이다.

나. 환경 설정

환경 설정은 (그림 12)와 같이 검색엔진을 작동하는 데 필요한 다음과 같은 참조파일들을 설정한다. 영역은 자유자제로 설정할 수 있으며, 여러개의 영역을 만들 수 있다. 또한 영역마다 독립적인 내용을 가지고 있으며, 다른 영역과 연관을 갖게 하려면 영역간의 오퍼레이션이 필요하다. DocumentPATH는 찾고자 하는 주제어에 관한 인덱스 파일이 들어가 있는 디렉토리를 지정하고, CommonWordPATH는 검색엔진이 참조하는 불용어 사전이 들어있는 디렉토리를 지정한다. 또한 IndexPATH는 검색엔진이 참조하는 여러가지 정보들, 즉 파일 인덱스 번호, 주



(그림 13) 검색엔진의 처리과정

제어 인덱스 번호 등이 들어 있다.

다. 변환과정 I

질의의 전 처리과정의 하나로서 질의어가 들어오면 내부적으로 쉽게 처리가 가능하도록 Postfix 형태의 질의어로, 즉 "이동통신 AND 미국"은 "이동통신 미국 AND" 의 형태로 변환한다.

라. 변환과정 II

변환과정 I에서 처리한 결과를 가지고 검색엔진에 맞게 다시 변환한다. 여러가지 옵션들을 다시 설정하고, 오퍼레이터와 오퍼랜드를 질문에 알맞는 순서대로 나열을 한다. Option 1에서는 정확도를 나타내는 옵션 등을 줄 수 있고, Option 2에서는 AND, OR 등이 올 수 있다.

예) [Option1] [Option2] Operands

마. 검색

검색엔진은 (그림 13)과 같이 직접 도큐먼트 파일을 이용하지 않고, 그것의 정보를 갖고있는 인덱스 파일을 이용한다.

1) 질의어 분석

질의어 분석(query analysis)은 변환과정 II를 거친 질의어를 가지고 해석을 한다. 확장자(wildcard)가 있는지 그리고 오퍼레이터(AND, OR)는 무엇인지를 판단한다. 확장자가 있으면 확장을 하고 브래킷([,])이 있으면 이것에 알맞는 확장을 한다.

2) 검색

검색은 실제로 이용자의 검색 요구를 처리하는 부분으로, 먼저 주제어의 루트작업(root를 찾는 작업, 영어인 경우 어근 추출, 한글인 경우 명사 사전 이용)을 한 다음, 이 주제어에 해당하는 인덱싱 번호를 찾은 다음, 이것으로부터 여러가지 정보를 얻는다. 이러한 정보들을 이용하여 결과를 찾아낸다.

3) 결과분석

결과분석(result analysis)에서 나오는 결과는 중간 결과로, 이 중간 결과를 가지고 한 번의 과정을 거쳐야 최종 결과가 나오게 된다.

바. 검색결과 제공

검색결과 제공부분에서는 검색에서 얻은 결과들을 가지고 마지막 마무리 작업을 한다. 중간 결과의 불필요한 부분을 제거하고, 질의어와 매칭되는 수가 많은 순서대로 출력을 한다.

V. 결 론

웹 기반의 한글 정보검색시스템을 구현하고 정보를 제공하기 위해서는 지금까지 살펴보았듯이 인터넷 통신환경을 접속하기 위한 네트워크 환경 및 웹 서버 구성, 그리고 대용량의 정보를 색인하고 검색할 수 있는 한글 검색엔진 등의 개발이 이루어져야 한다. 이에 본 고에서는 한글 검색엔진을 자체 개발하고자 할 경우에 있어서 갖추어야 할 기능 및 방법들에 대해서 서술하였다.

정보검색시스템은 제공되는 정보의 내용이 기존의 PC 통신환경 기반의 텍스트 정보에서 TCP/IP 인터넷 통신환경 기반의 텍스트, 이미지, 오디오, 비디오 등의 멀티미디어 정보로 변화되면서 이용자 인터페이스 및 검색시스템의 유형 등에도 많은 변화를 가져왔다. 따라서 이를 지원할 수 있는 하이퍼텍스트 및 GUI(Graphic User Interface) 기반의 웹 브라

우저들과 이와 상호 인터페이스할 수 있는 웹 서버들이 개발되었다.

또한 검색시스템의 유형을 보면, 현재도 서비스되고 있는 텍스트 기반의 DIALOG, BRS 등 대표적인 검색시스템들은 주제어 중심의 불리언 검색이 주를 이루고 있으나 인터넷이 확산되면서 주제어 중심의 불리언 검색 외에 하이퍼텍스트 기반의 주제어 카탈로그 검색, 각기 다른 사이트의 검색엔진들로부터 질의한 결과를 통합하여 제공하는 지능형 검색시스템 등이 등장하게 되었다.

그러나 현재 인터넷 상에서 운용되고 있는 국외의 야후, 알타비스타 등과 심마니, 미스 다찾니 등 국내의 검색시스템들은 로봇 에이전트를 이용하여 인터넷 사이트에 등록되어 있는 다양한 분야의 홈페이지 정보들을 수집/분석하여 관련 사이트를 연결해 주는 방식의 메타 검색엔진들로서 불필요한 정보들까지 제공함에 따라 이용자들이 필요로하는 정보를 찾기에는 너무 많은 노력과 시간을 소모하게 되는 문제점을 안고 있다.

따라서 본 고에서는 웹 기반의 한글 정보검색시스템을 구현하는 데 있어서 핵심 부분이 되는 한글 검색엔진이 갖추어야 할 기능 및 구현 방법, 특히 명사, 조사, 불용어 등 각종 한글 사전 등을 이용하여 한글의 특성에 맞는 형태소 분석을 이용하여 검색의 정확성 및 재현율을 향상시키는 방법과 부가적인 기능으로 맞춤정보서비스, 주제별 카탈로그 서비스 등에 대해서 기술하였다.

참 고 문 헌

[1] D.S. Haverkamp, S. Gauch, "Intelligent Information Agents: Review and Challenges for Distributed Information Sources," *J. of the American Society for Information Science*, Vol. 49, No. 4, 1998, pp. 304 – 311.

[2] 전인걸, 박영우, 이은석, "개인 적응형 에이전트를 이용한 정보검색과 필터링," 한국정보과학회 '97 학술발표논문집 (B) 24호, 1997, pp. 555 – 558.

[3] 김병학, 이광형, 조충호, "정보검색을 위한 인텔리전트 웹 에이전트," 정보과학회지 제14호, 1997, pp. 12 – 21.

[4] 이현아, 홍남희, 이종혁, 이근배, "한국어 형태소 구조 규칙에 기반한 색인시스템의 구현," 한국정보과학회 학술발표논문집, 1995, pp. 933 – 936.

[5] Charles T. Meadow, *Text Information Retrieval Systems*, Academic Press, Inc., 1992, pp. 129-156.

[6] A.S. Pollitt, *Information Storage and Retrieval Systems*, Ellis Horwood Limited, 1989, pp. 59-106.

[7] J.H. Ashford, P. Willett, *Text Retrieval and Document Databases*, Chartwell-Bratt, Bromley, 1988.