지역정보망을 위한 실시간 검색 엔진의 설계 및 구현

정 용훈, 김 일, 김 성후, 박 규석
경남대학교 컴퓨터공학과

Design and Implementation of Realtime Search Engine for Local Information Network

Yong-Hun Jung, Il Kim, Sung-Hoo Kim, Kyoo-Seok Park
Dept. of Computer Science, Kyungnam University

대부분의 정보는 시간 제약을 가지므로 사용자는 필요 정보를 일정 시간 내에 얻어야만 한다. 시간적 제약을 가진 정보를 적절하게 서비스하려면 분산된 정보를 실시간에 검색하고 서비스해 주는 실시간 검색시스템이 필요할 뿐 만 아니라, 필요에 따라 실시간으로 검색시스템의 정보를 갱신해 주어야 한다.

본 논문에서는 인터넷을 이용한 지역 특성에 맞는 인터넷 정보 시스템 모델을 제안하고, 지역정보망을 위한 실시간 제어 검색엔진을 설계 및 구현하였다. 따라서 분산되어 있는 정보를 쉽게 모아서 갱신해 줄 수 있는 지능적인 로봇의 구현과 주기적으로 네트워크의 부하를 모니터링하고 모니터링한 데이터를 기반으로 로봇의 동작 정책을 수립하여 전체 시스템의 성능을 향상시킬 수 있는 실시간 로봇 제어 정책이 가능하다.

1. 서론

정보화 사회가 진전되면서 정보화가 타지역의 문화와 관습들에 대한 접촉 가능성을 높여주고 자기 지역에 부족하거나 또는 개발 가능한 부분에 대한 인식을 확산시킴으로써 지역을 활성화시킬 수 있다는 것은 누구도 의심치 않고 있다.

국가적 차원에서 정보화추진 10대 과제에 지역정보화가 포함되어 있고, 특히 지방자치단체가 지역정보화에 각별한 관심을 갖게 된 현 시점에서, 지역의 특성에 맞는 정보망을 구축하고 이를 효율적으로 관리해 나가야 함은 절실하다[1].

본 논문에서는 인터넷을 이용하여 지역 특성에 맞는 인터넷 정보 시스템 모델을 제안하고 지역정보망을 위한 실시간 제어 검색엔진을 설계 및 구현하였다. 본 시스템은 네트워크 부하에 대한 모니터링으로 네트워크의 부하를 고려하여 웹 서버의 정보를 가져올 수 있어 네트워크 부하를 조정할 수 있으며, 로봇의 동작 시간 조정으로 네트워크의 과부하를 줄일 수 있다. 본 논문의 2장에서는 관련 연구에 관하여 설명하고, 3장에서는 제안한 시스템의 설계 과정을 기술한다. 4장에서는 구현내용을 기술하고, 5장에서 결론을 내린다.

2. ASP(Active Server Pages)

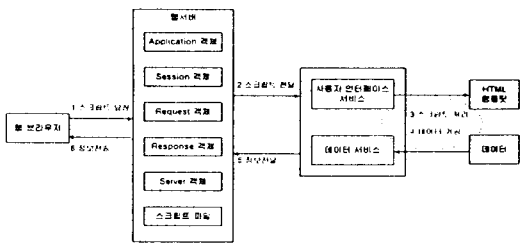
1996년 12월 발표된 ASP는 IIS(Internet Information Server) 버전 3.0에서 사용할 수 있도록 한 것으로 서버측 프로그래밍 기술로 가장 큰 장점은 처리 시간을 줄일 수 있다는 것이며, 고도로 동적이

며 상호 대화적인 애플리케이션을 작성할 수 있다. ASP 페이지는 스크립트와 HTML 코드로 결합되어 구성되며, 이들 스크립트와 HTML 코드는 ASP가 지원하는 내장 객체에 대한 호출을 포함할 수 있다.

웹 브라우저가 ASP 페이지를 호출하면 웹 서버는 웹 브라우저로부터의 요청을 ASP 엔진으로 넘긴다. 요청을 넘겨받은 ASP 엔진은 스크립트를 처리하고 결과를 HTML 스트림으로 삽입한 다음에 요청을 보낸 웹 브라우저로 그 결과를 반환한다. 이 모든 과정은 서버에서 이루어지고 모든 결과 출력은 HTML 형태가 되므로 일반적인 브라우저의 사용에도 문제가 되지 않는다[2].

2.1 ASP의 동작원리

그림 1에서처럼 클라이언트 브라우저에서 요구한 데이터는 서버측의 ASP와 애플리케이션 컴포넌트를 통해 가공/처리되며, 처리된 결과는 HTML 형식으로 바뀌어 클라이언트인 브라우저로 반환된다.

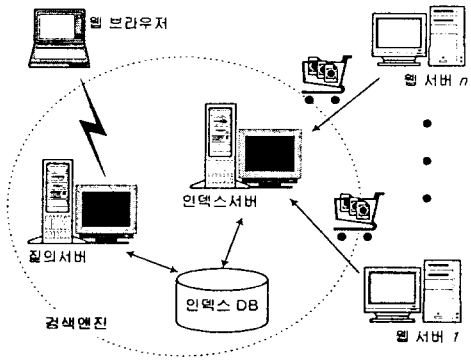


<그림 1> ASP의 동작원리

그리고 액티브 서버 프레임워크, 애플리케이션 컴포넌트, ASP는 각각 다른 일을 수행하는데, 액티브 서버 프레임워크는 ASP의 근간을 이루는 것으로서 5가지 기본 객체에 대한 동작을 정의한다. 그리고 애플리케이션 컴포넌트는 ASP에서 지원하는 여러 가지 컴포넌트들로 데이터베이스에 직접 접근하거나 질의 실행 등을 도와주는 객체이다. 마지막으로 ASP는 문서속에 포함되어 있는 ASP 코드들을 분석, 이를 적절한 서비스로 넘겨주는 역할을 한다.

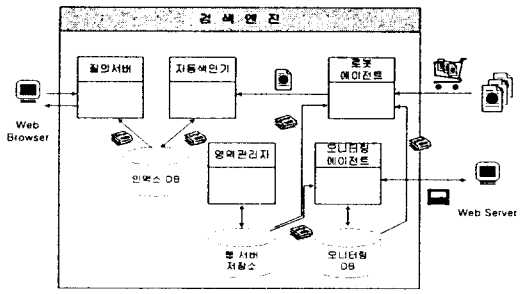
3. 지역정보화를 위한 검색 시스템의 구조

웹을 이용한 정보검색시스템은 정보를 검색할 수 있는 검색엔진과 정보제공자인 다수의 웹서버, 그리고 정보를 요구하게 되는 클라이언트인 웹 브라우저로 구성되며 그 구성도는 그림 2와 같다.



<그림 2> 정보검색시스템의 구조

검색엔진은 크게 사용자의 질의를 처리하여 결과를 돌려주는 질의서버(Query Server)와 특정 영역을 관리하는 영역관리자(Region Manager), 영역 내의 HTML 문서를 수집하는 로봇 에이전트(Robot Agent), 그리고 수집한 문서를 인덱싱하는 자동색인기(Automatic Index Builder), 등록된 웹서버간의 네트워크 부하를 모니터링하는 모니터링 에이전트(Monitoring Agent)로 구성되며 그림 3과 같이 구성된다.

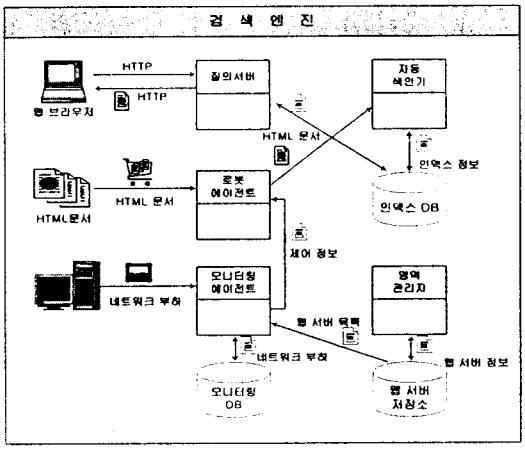


<그림 3> 지역 정보를 위한 검색 시스템의 구성도

3.1 서브시스템별 요구 데이터의 흐름

그림 4는 검색 엔진의 각 서브 시스템별 요구 데이터의 흐름을 나타낸 것으로 질의서버는 클라이언트인 사용자로부터 질의어를 입력받아 질의어에 대한 결과를 사용자에게 되돌려준다. 자동색인기는 로봇에이전트가 수집한 HTML 문서에 대한 인덱스 정보를 자동으로 추출하여 관리하고 로봇 에이전트는 모니터링 에이전트가 제공하는 로봇 제어정보에

의해 각 웹서버별 HTML 문서를 수집하여 자동색인기에 넘겨준다. 영역관리자는 정보를 제공하게될 각 웹서버를 등록·관리하며, 모니터링 에이전트는 웹서버에 대한 네트워크 부하 모니터링을 실시하여 로봇 제어 정보를 로봇 에이전트에게 넘겨준다.



<그림 4> 각 서버 시스템별 요구 데이터의 흐름

3.2 영역관리자

영역관리자는 검색 엔진에서 서비스할 웹서버를 검색엔진에 등록하거나 삭제하는 기능을 수행하며 웹서버의 상태정보를 가진다. 웹서버의 등록 및 삭제는 수동으로 관리자가 작업하는 경우와 IP 어드레스 범위만 주어지면 자동으로 웹서버를 찾아서 영역을 관리하는 경우로 나뉜다.

관리할 영역이 주어지면 영역 내에 있는 웹서버를 자동으로 찾아 등록하며, 등록된 웹서버는 영역관리자가 수작업에 의해 웹서버에 대한 정보를 수정하고 웹서버에 대한 등록 및 삭제에 대한 관리 권한을 가진다.

3.3 로봇 에이전트

웹 로봇을 동작시키기 위한 알고리즘은 다음과 같다.

Backward link를 기반으로 한 알고리즘

```
enqueue(Going_Queue, starting_url)
while (not empty(Going_Queue))
   url = dequeue(Going_Queue)
   page = going_page(url)
   enqueue(Gone_Queue, (url, page))
   url_list = extract_urls(page)
   for i = 1 to url_list
      enqueue(links, (url, u))
      if [u not in Going_Queue] and
        [(u, -) not in gone_pages] Then
         enqueue(Going_Queue, u)
   Next i
Wend
```

Function Description

enqueue(queue, element) : 큐의 끝에 하나의 원소를 추가한다.

dequeue(queue) : 큐의 첫 원소를 제거하고 그 값을 돌려준다.

extract_urls(page) : 문서내의 링크정보를 추출한다.

3.4 자동색인기

검색시스템에서 색인이란 자료로부터 색인어를 추출해 내는 작업과 그 색인어에 대한 정보의 위치를 지시함으로써 효율적인 정보검색의 기반을 제공하는 색인(또는 색인파일) 작성의 작업으로 이루어진다.

자동색인기는 수집기에서 가져온 HTML 문서를 HTML 태그를 제거하고 남는 문서의 내용을 가지고 타이틀 및 색인어를 추출하여 색인데이터베이스를 만들고, 가져온 HTML 문서는 문서 등록 데이터베이스에 등록한다.

3.5 모니터링 에이전트

모니터링 에이전트는 접속할 웹서버의 네트워크 부하상태를 고려하여 검색 엔진의 수집기가 웹서버에 접속하여 URL 정보를 가져올 시간을 결정하는데 사용된다.

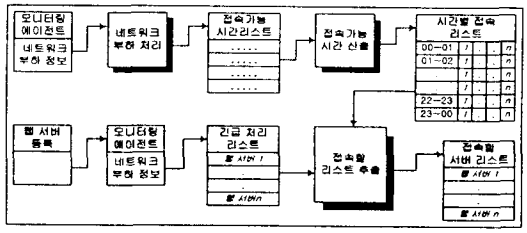
네트워크의 부하상태를 고려하여 웹서버에 접속하는 이유는 네트워크의 부하 상태를 고려하여 네트워크의 부하가 적은 시간대에 웹서버에 접속하여 웹서버의 정보를 가져오면 네트워크에 가중되는 부하를 줄일 수 있고 웹서버의 부하가 적은 시간대에 접속하면 더욱 향상된 성능을 가져올 수 있다.

또한 각 웹서버에 대한 네트워크의 모니터링된 결과는 데이터베이스에 주기적으로 기록되고 기록된 데이터는 네트워크의 부하 상태를 한 눈에 알 수 있도록 보여주는데 이용된다.

네트워크의 부하상태는 대상 서버와의 통신속도를 측정하여 ms 단위로 기록되고 서버의 부하 상태는

서버로부터 시스템 자원의 사용 상태를 전송 받아 기록한다.

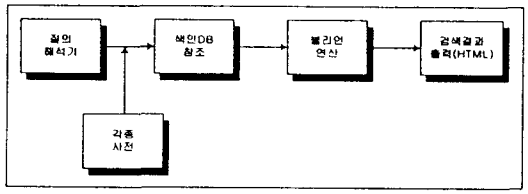
그림 5는 모니터링 에이전트에 의한 네트워크 모니터링 데이터를 기반으로 접속할 웹서버의 스케줄을 산출하기 위한 제어 흐름도이다.



<그림 5> 모니터링 제어 흐름도

3.6 질의서버

질의서버는 사용자가 입력한 질의어를 분석하여 각각의 질의어에 대해 색인 데이터베이스를 참조하여 적합한 문서를 검색하고, 검색 결과에 대해 검색 연산자(AND, OR,...)등을 적용하고 가중치를 적용하여 사용자가 원하는 정보를 제공해 준다.



<그림 6> 질의서버의 검색 과정

질의서버의 검색 과정에서, 사용자가 입력한 질의어는 질의어 해석기를 통해 검색 문법에 맞는지 분석한 다음 그림 6과와 같이 색인 데이터베이스를 참조하여 주어진 질의어에 적합한 문서를 검색한다.

검색 결과는 질의어에 사용된 검색연산자에 의해 불리언 연산을 수행한 후, 최종 검색결과를 HTML 형태로 가공하여 사용자에게 출력한다.

4. 구현 및 평가

본 논문에서 제안하고 구현한 시스템은 크게 사용자의 질의를 받아 검색 결과를 HTML문서로 사용자에게 보내주는 질의서버, 지역 내에 있는 웹서버를 관리하는 영역 관리자, 수집한 문서에 대한 인덱싱을 하는 자동색인기, 웹서버에 대한 네트워크 부하를 모니터링하는 모니터링 에이전트, 웹서버로부터 웹문서를 수집하는 로봇 에이전트로 구성된다.

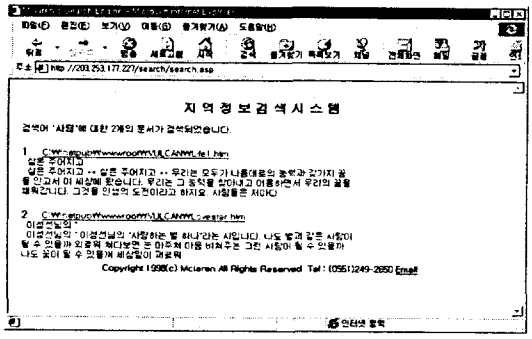
제안한 시스템의 설치 및 운용방법은 다음과 같다.

지역정보망에 웹서버를 등록하려면 검색 엔진의 영역 관리자에서 웹서버를 등록한다. 등록을 마치면 웹서버에 대한 네트워크 부하상태를 모니터링하여 최적의 상태를 산출하여 접속할 시간을 설정한다. 설정된 시간에 웹서버에 접속하여 네트워크의 부하상태를 검사하고 접속 가능 상태로 판단되면, 로봇 에이전트는 해당 웹서버의 URL 정보를 수집하여 새로운 URL 정보의 추출 및 HTML 문서에 대한 인덱싱을 실시한다.

등록된 모든 웹서버에 대한 인덱싱이 완료되면 질의서버로부터 해당 지역에 대한 정보서비스가 시작된다.

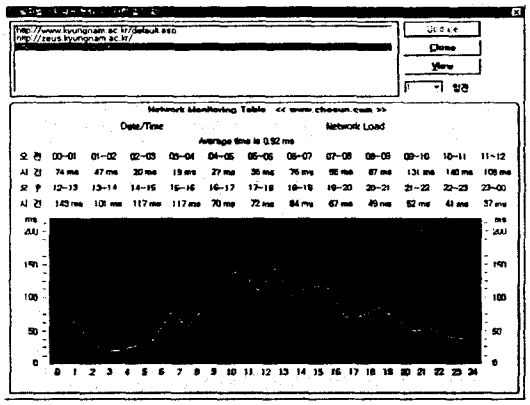
그림 7은 검색어 "사람"에 대한 처리 결과를 보여주는 화면으로, 사용자가 질의어를 입력하면 질의서버는 질의어 해석기를 통해 질의어가 검색 문법에 맞는지 분석한 다음 색인 데이터베이스를 검색한다.

검색 결과는 질의어에 사용된 검색연산자에 의해 불리언 연산을 수행한 후 최종 검색결과는 HTML 형태로 변환하여 사용자에게 출력된다.



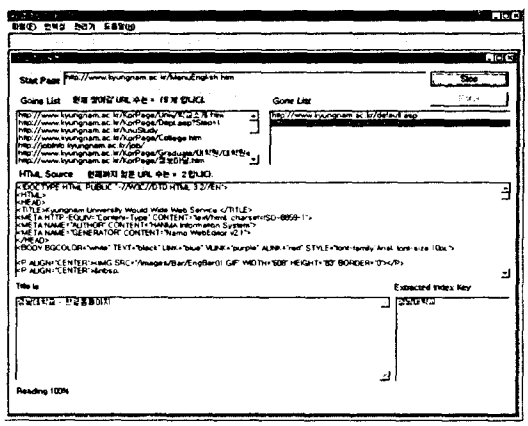
<그림 7> 검색어 '사람'에 대한 검색 결과 화면

모니터링 에이전트는 웹서버의 부하상태와 네트워크의 부하상태를 모니터링하고 모니터링한 결과를 데이터베이스에 주기적으로 기록하며 모니터링 상황을 관리자가 확인할 수 있도록 그림 8과 같이 그래프로 나타내주어 내부적으로 처리되는 모니터링을 관리자가 비주얼하게 확인 가능하도록 구현하였다.



<그림 8> 네트워크 부하모니터링 화면

그림 9는 자동색인기를 실행한 후의 결과화면이다.



<그림 9> 자동색인기 실행화면

5. 결론

지방자치제가 시행되고 있는 현 시점에서 지역 특성에 맞는 인터넷 정보 서비스망을 구축하기 위한 시스템을 제안하고 이 시스템에 필요한 실시간 제어 검색엔진 및 분산된 정보를 모아서 갱신해 주는 지능적인 로봇, 네트워크 부하 및 웹서버의 부하를 모니터링하여 전체 시스템의 성능을 향상시킬 수 있는 모델을 설계 및 구현하였다.

본 논문에서 제안한 시스템은 분산된 다수의 웹서버와 웹서버의 부하, 네트워크의 부하를 고려하지 않고 로봇을 동작시키고, 2~3주 간격으로 정보를 갱신하는 기존의 검색엔진과는 다르게 지역의 특성화를 위하여 3~4일 정도의 짧은 기간에 모든 정보를 최신의 정보로 갱신하며, 주기적으로 네트워크의 부하를 모니터링하고 모니터링한 데이터를 기반으로 로봇의 동작 정책을 수립하여 전체 시스템의 성능 향상 및 실시간 제어를 가능하게 함으로서 검색엔진의 부하를 줄이고 지역 정보에 대한 실시간 갱신의 효과를 가져올 수 있다.

참고문헌

[1] 정 국환, 이 석재, 류 승호, 김 형민, "지역정보화사업 평가와 추진 방안", 한국전산원, 1997

[2] 사이버게이트8, "한글 윈도우 NT에서의 ASP 활용", 사이버출판사, 1998. 9

[3] 여 찬기, 송 관호, 유 용석, "한국인터넷디렉토리 시스템 최종개발완료보고 서", 한국전산원, 1997

[4] 권 혁진, 김 영민, 김 형근, 이 상엽, 정 일형, 조 강래, 신 봉기, 송 주원, 장 희순, "국내 웹 정보검색 기술의 동향", 정보과학회지, pp.16-22, 1997.10, 제 15권, 제 10호,

[5] 고 형대, 유 재수, 김 병기, "효율적인 정보 검색 시스템 구축을 위한 새로운 프로세스 구조, 한국정보처리학회 논문지, pp.76-86, 1997.1, 제 4권 제 1호.

[6] 오 수철, 정 상화, 류 광렬, "분산 메모리 다중 프로세서 시스템에서의 병렬정보검색", 한국정보과학회논문지, pp.1078-1089, 1997.11, 제 24권 제 11호.

[7] 박 영몽, 김 민구, 이 정태, "지식 기반의 정보검색 시스템", 한국정보 과학회논문지, pp.2090-2098, 1994.11, 제 21권 제 11호.

[8] 충남대학교 컴퓨터공학과 데이터베이스 연구실, "Robot agents and Search Engine", http://sharon.comeng.chungnam.ac.kr/~bluefrog/ws2_html.html

[9] 한 선영, 이 강준, 한 기준, "World-Wide Web을 위한 한글 병렬 서치 엔진", 한국정보과학회 학술발표논문집, April 20 1996. pp.139-142.