Journal of Korea Multimedia Society Vol. 8, No. 3, March 2005(pp. 355-361)

다중스레드를 이용한 분산 환경에서의 이미지 검색 에이전트

차상환[†], 김순철[††], 황병곤[†††]

요 약

본 논문에서는 분산 환경에서 이미지 정보를 수집하고 검색하기 위해 다중스레드를 사용한 이미지 검색 에이전트를 구현하였다. 본 논문에서 제안한 이미지 검색 에이전트는 정보의 효과적인 검색을 위해 다중스레드를 사용함으로써 처리기의 이용률을 높일 수 있었고 웹상의 정보를 전달받는데 소요되는 대기시간과 처리시간을 줄일 수 있었다. 이미지 검색을 위한 에이전트는 플랫폼에 독립적인 자바 언어를 사용하여 분산 환경에 적합하도록 하였고, 검색한 이미지 저장을 위해 JDBC를 사용하여 데이터베이스에 연결하였다. 또한 이미지 자체는 분산된 에이전트의 데이터베이스에 저장하고 이미지의 인덱스만 인덱스 서버에 저장함으로써 검색 시간을 줄일 수 있었다.

Multi-Thread Based Image Retrieval Agent in Distributed Environment

Sang-Hwan Cha[†], Soon-Cheol Kim[††], Byung-Kon Hwang[†††]

ABSTRACT

This paper proposed a system collecting image information by agents in multi-threaded environment and then retrieving them with content based image retrieval. This system uses multi threads to retrieve web information effectively, then improves efficiency of CPU cycles to reduce latency time, which is the time requesting queries, executing communication processing that the retrieval agents perform and filtering the retrieval results. Also, the agents for image retrieval use Java language, which is platform independent, to be suitable for distributed environment. Using JDBC to save the retrieved images, the agents are connected to database. The images themselves are stored in distributed agents' databases, and only the image indexes are stored in an index server so that the efficiency of storage and retrieval time can be improved.

Key words: Multi-Thread(다중 스레드), Agent(에이전트), JDBC(자바데이터베이스접속)

1. 서 론

2004년 1월 기준으로 전 세계 호스트의 수는 약 2억 3천여개이다[1]. 이는 조사를 시작한 1993년 130만개 정도와 비교하여 무려 177배나 증가하였다. 이러한 호스트의 규모, 즉 인터넷망이 증가함에 따라 사용자들은 텍스트, 이미지 그리고 비디오 등과 같은 원하는 정보를 인터넷 상의 호스트에서 정확하고 신속하게 검색하기는 매우 어려운 일이다. 그러므로 웹상에 존재하는 방대한 정보를 좀 더 효과적으로 저장

※ 교신저자(Corresponding Author) : 황병곤, 주소 : 경북 경산시 진량읍 내리리 15번지(712-714), 전화 : 053)850-6580, FAX : 053)850-6589, E-mail : bkhwang@taegu.ac.kr
접수일 : 2004년 7월 10일, 완료일 : 2005년 1월 20일
[†] 준회원, 대구대학교 일반대학원 컴퓨터정보공학과
(E-mail : cha1977@webmail.daegu.ac.kr)
[††] 정회원, 대구대학교 컴퓨터IT공학부 조교수
(E-mail : kimsc@daegu.ac.kr)
[†††] 종신회원, 대구대학교 컴퓨터 정보공학부 정교수
※ 본 논문은 2004년도 대구대학교 교내 학술연구비 지원에 의해 연구되었음.

하고 검색할 수 있는 방법에 관한 연구가 필요하게 되었다.[2,3].

초기의 검색 방법은 웹상에 존재하는 정보를 검색하기 위한 작업을 서버에서 수행하였다. 이러한 환경에서는 계속적으로 증가되는 정보를 저장하기 위한 데이터베이스 용량의 증가 및 서버의 부하가 늘어나게 된다는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위한 방법으로 검색 작업을 여러 대의 컴퓨터에 의한 분산처리를 이용한 에이전트 시스템이 연구되었다[4,5]. 이러한 시스템은 웹상의 정보를 검색하기 위해서 메타 검색엔진을 사용하는데 클라이언트에 검색 모듈들을 분산시킴으로써 네트워크에 트래픽이 집중되는 문제를 해결하고 검색 서버의 과도한 부하를 개선하였다[4]. 그러나 메타 검색엔진은 검색작업에서 질의문 생성이나 결과 취합 등 몇 단계의 검색작업이 추가되고 그에 따라 검색 대기시간도 증가하게 된다. 더불어 사용자가 질의를 하는 순간부터 질의문을 생성하고 각 검색엔진으로부터 결과를 취합하기 때문에 이미지를 탐지하고 분석, 자름, 그레이스케일로의 변환 등과 같은 이미지 프로세싱 작업이 필요한 내용기반 이미지 검색시스템에서는 메타 검색엔진을 이용하는 것은 부적절하다. 이미지 프로세싱 작업은 많은 처리 능력을 요구하기 때문에 사용자가 질의를 한 순간부터 검색작업을 시작하면 텍스트를 처리하는 시간보다 검색 시간이 월등히 증가하게 된다. 따라서 내용기반 이미지 검색시스템에서는 사용자가 질의를 하기 전에 이미지 검색 에이전트에 의해 수집된 이미지 인덱스를 가지고 검색에 이용되어야 한다. 따라서 내용기반 이미지 검색 시스템에서는 자신의 데이터베이스를 갖고 있지 않은 메타 검색엔진 보다는 자신의 데이터베이스를 갖고 있는 일반적인 검색엔진을 이용하는 것이 더 타당하다[6-8]. 그러나 이러한 일반적인 검색엔진을 사용하더라도 CPU 사이클의 효율성이 대기시간에 많은 영향을 준다[2,9].

이러한 문제점을 해결하기 위해 본 논문은 다중스레드를 사용하여 CPU 사이클의 효율성을 높여 대기시간을 줄였고, 결과적으로 이미지 프로세싱 처리시간을 줄여 전체 검색시간을 줄였다.

논문의 구성은 2장에서 웹 검색 에이전트에 관해 설명하고, 3장에서는 제안한 분산 환경에서의 이미지 검색에 관해 설명한다. 4장에서는 CPU 사이클이 검색에 미치는 영향과 검색 에이전트에 의해 수집된 이미지를 내용기반 이미지 검색 시스템으로 테스트한 결과를 설명한다. 그리고 마지막 5장에서는 결론을 맺는다.

2. 웹 검색 에이전트

웹 검색 에이전트는 웹의 하이퍼텍스트 구조를 돌아다니며 자동적으로 문서의 정보를 수집해 검색엔진이 검색할 수 있도록 만들어주는 프로그램이다[10].

2.1 웹 검색 에이전트 동작 알고리즘

웹 검색 에이전트 동작 알고리즘은 그림 1과 같고 다음과 작은 동작 순서를 가진다.

① 최초 지정된 웹 사이트를 탐색한다.
② URL의 호스트 이름에 따라 robots.txt에 접근한다.
③ robots.txt 파일의 내용을 받아온다.
④ 받아온 robots.txt 파일을 분석한다.
⑤ 만일, 검색 에이전트의 접근을 배제하는 사이트가 아니면 해당 URL에 다시 접근한다.
⑥ 접근한 URL에서 웹문서를 받아온다.
⑦ 수집된 문서를 분석하여 URL을 추출하여 URL 데이터베이스에 저장한다.
⑧ 수집한 문서에서 정보를 추출하고 인덱스 데이터베이스에 저장한다.
⑨ 연결된 문서에서 사용자가 지정한 카운트를 벗어나지 않는 범위에서 ② 혹은 ③부터 반복하여 탐색을 한다.

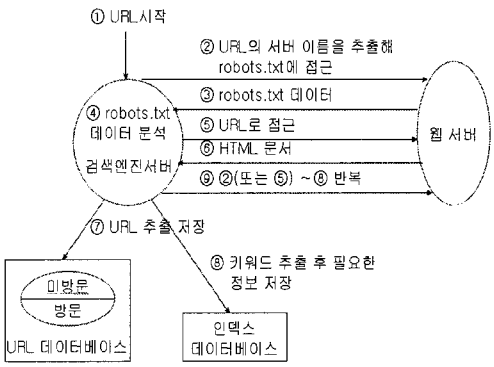


그림 1. 웹 검색 에이전트의 동작 순서

2.2 웹 검색 에이전트의 데이터베이스 접속

웹 검색 에이전트가 JDBC를 이용하여 데이터베이스에 접속하기 위해서는 우선 드라이버의 유형과 데이터베이스의 종류에 맞는 드라이버를 메모리에 로드해야 한다.

이후, 그림 2와 같이 MySQL에 연결하기 위해 Class.forName() 메소드를 이용하여 메모리에 로드했고 JDBC 인터페이스의 DriverManager라는 클래스의 getConnection 메소드를 이용하여 JDBC 드라이버를 자바 응용 프로그램과 데이터베이스에 연결한다. 연결을 위해서는 데이터베이스의 위치를 기술하기 위한 URL, 데이터베이스의 사용자 이름, 암호 등을 필요로 한다.

```
package imagesagent;
import java.sql.*;
public class DBConnection{
  private Connection conn = null;
  private Statement stmt = null;
  private static DBConnection instance = null;
  private DBConnection(){  // 외부에서 생성
    try{                   // 할 수 없게 private
  // 드라이버 로드 및 DB 접속
    Class.forName("org.gjt.mm.mysql.Driver");
    conn = DriverManager.getConnection
              ("jdbc:mysql://URL/", "UserName",
               "PassWord");
        stmt = conn.createStatement();
    }catch(ClassNotFoundException e){
      System.out.println("JDBC 드라이버를 찾을 수
없습니다");
    }catch(SQLException e){
      System.out.println("DataBase에 연결할 수 없습
니다");
    }
  }
```

그림 2. JDBC 데이터베이스 연결

3. 분산 환경에서의 이미지 검색

검색엔진은 일반적으로 검색 시간의 70% 가까이를 웹 서버에서 클라이언트로 자료를 받는데 소비한다[2]. 따라서 이미지 검색 에이전트가 이미지들을 전달받을 때, 요청하는 수에 비례하여 기하급수적으로 이미지 수집 시간이 증가하게 된다. 이미지 수집을 위해 웹 페이지를 요청하고 응답받는 시간까지 CPU는 정지(idle)상태가 되고 이로 인해 CPU 사이클은 낭비된다.

다중스레드 프로세스의 중요한 목적은 이러한 CPU의 정지상태를 없애거나 최소화하기 위해 병렬수행(parallel processing)을 하여 처리속도를 증가시키는데 있다[2,6]. 즉, 하나의 스레드가 웹페이지를 전달받기 위해 기다리는 동안 다른 스레드는 전달받은 페이지를 처리한다. 더욱이 이미지를 탐지하고 분석, 자름, 그레이스케일로의 변환 등과 같은 이미지 프로세스 작업은 많은 프로세싱 능력을 요구하게 되어 전체 응답시간에 부담을 준다. 따라서 빠른 응답시간을 얻기 위해서 웹 서버로 동작하고 있는 서로 다른 컴퓨터를 이용하여 작업 처리를 하여 계산 능력을 분산 시켰다.

3.1 이미지 검색 구조

그림 3은 분산 환경에서의 다중스레드를 이용한 이미지 검색 구조로 다음과 같이 동작한다.

① URL 서버가 URL 목록을 작성하기 위해 초기에 방문할 URL을 관리자가 URL 서버에게 제공한다.

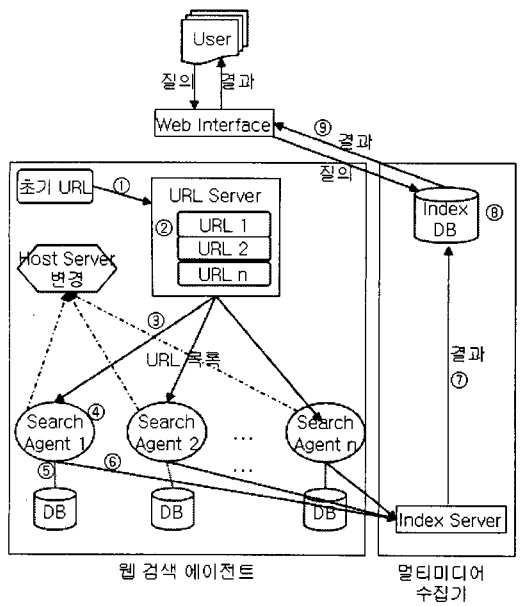


그림 3. 이미지 수집 구조

② URL 서버는 중복되지 않은 URL를 받아와 URL 목록을 만들어 스택에 저장한다.
③ URL 서버는 URL 서버에 저장된 목록을 n개의 검색 에이전트에 각각 분배한다.
④ 검색 에이전트들은 다중스레드로 구성되어 있고, 분배 받은 목록들을 검색하고 해당 웹 페이지를 접근해 분석하여 이미지를 전달받는다.
⑤ 전달받은 이미지를 자신의 데이터베이스에 저장한다. 이때, 이미지 저장 시간도 데이터베이스에 저장한다.
⑥ 검색 에이전트는 이미지의 이름과 경로, 수집한 시간, 에이전트 이름, URL, 이미지 수집시간 등 이미지의 인덱스를 인덱스 서버로 보낸다.
⑦ 인덱스 서버는 이미지의 인덱스를 데이터베이스에 저장한다.
⑧ 인덱스 데이터베이스는 사용자의 질의와는 상관없이 반복해서 에이전트로부터 이미지 인덱스를 업데이트 하고 사용자 질의에 대한 응답을 제공한다.
⑨ 서버를 통해 사용자가 질의를 보내면 웹 인터페이스는 인덱스 데이터베이스에서 가장 유사도가 큰 순서로 이미지들을 제공한다.

3.2 웹 인터페이스

웹 인터페이스는 사용자가 편리한 방법으로 질의를 할 수 있도록 GUI (Graphical User Interface) 를 제공한다. 사용자가 질의하는 방법은 스케치에 의한 질의와 예제이미지에 의한 질의방법을 제공한다. 그리고 그 결과를 유사도가 높은 이미지순으로 사용자에게 전달한다.

3.3 URL 서버

URL 서버의 상세도는 그림 4와 같다. 초기 접속할 URL을 제공받아 URL 목록을 모은다. 이후 에이전트의 요청에 의해 URL 목록을 전달하는 역할을 한다.

이미지 검색 에이전트에게 URL 목록을 전달할 때는 다중스레드를 이용해 병렬처리하여 통신이 지연되는 것을 방지한다. 또한 다수의 에이전트들이 하나의 URL 서버의 URL 목록을 요청하여 URL을 받아올 때, 전송도중 데이터의 무결성이 깨져버릴 수 있다. 특히, 에이전트의 수가 늘어나면 날수록 무결성

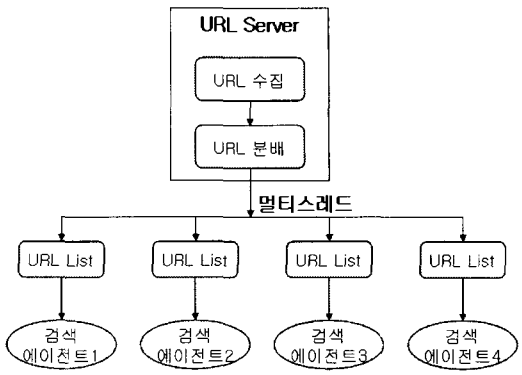


그림 4. URL 서버 모듈

이 깨져버릴 확률은 높아진다. 이를 방지하기 위해 자바 키워드인 synchronized 블록을 사용해서 동기화하여 데이터의 무결성을 지켰다.

3.4 이미지 검색 에이전트

이미지 검색 에이전트는 자바 어플리케이션으로 구현했으며, 그림 5에서와 같이 URL 목록으로부터 받은 웹 페이지를 전달받아 분석하는 작업을 한다.

이러한 작업에는 많은 CPU 사이클을 요구하기 때문에 시스템 자원을 효율적으로 사용해야 한다. 따라서 병행 처리하는 다중스레드 시스템을 사용했다. 다중스레드는 요청과 응답에 독립적으로 통신하는 환경으로 CPU, 메모리 등의 시스템 자원을 능률적으로 사용하게 한다. 이렇게 다중스레드를 이용함으로써 CPU 사이클의 정지 상태를 줄여 전체적인 이미지 검색 시간을 단축할 수 있다.

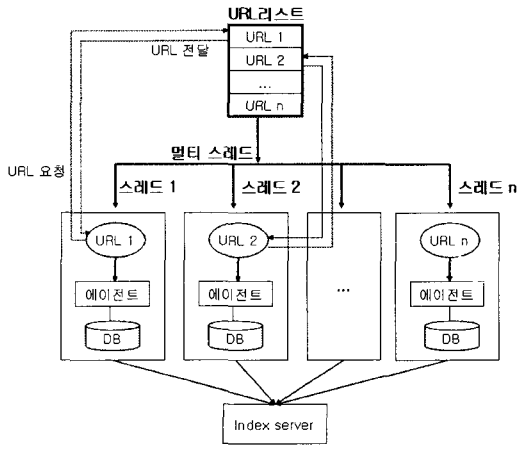


그림 5. 이미지 검색 에이전트 구조

또한, 각각의 에이전트들은 이미지를 자신의 데이터베이스에 저장하고 이미지 인덱스를 유지한다. 이러한 구조를 구성한 이유는 이미지들이 네트워크상의 많은 서버로 나누어 저장됨으로 이미지 저장 공간의 효율성도 높일 수 있기 때문이다.

3.5 인덱스 서버

인덱스 서버는 각기 다른 검색 에이전트로부터 받은 이미지 인덱스들을 저장하고 인덱스 데이터베이스를 업데이트한다.

표 1은 인덱스 서버가 이미지 검색 에이전트로부터 전달받은 이미지 인덱스의 예를 나타내었다.

표 1. 이미지 인덱스 내용

이미지의 이름	무궁화.jpg
에이전트 이름	Thread-1
이미지 경로	http://wstatic.naver.com/w3/무궁화.jpg
URL	http://watatic.naver.com
이미지 수집시간	Fri Feb 27 22:11:20 KST 2004

3.6 인덱스 데이터베이스

인덱스 데이터베이스는 에이전트의 데이터베이스에 저장되어 있는 이미지 원본과 관련된 이미지의 인덱스를 저장하고 있다. 그림 6은 인덱스 데이터베이스의 구조로 이미지 프로세싱 처리를 완료한 상태를 나타내었다.

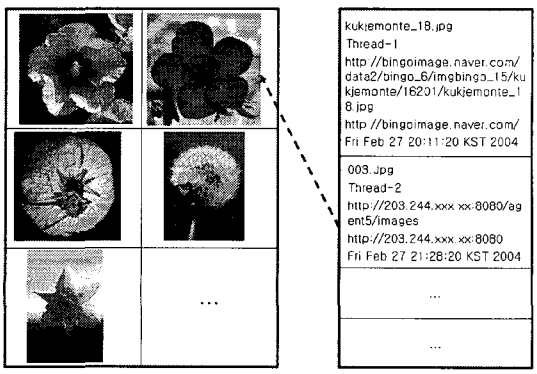


그림 6. 인덱스 데이터베이스 구조

4. 실험결과

본 논문에서 제안된 시스템의 구현 환경은 Windows 2000 Server를 운영체제로 하는 P4 1.6MHz 상의 컴퓨터 5대로 실험하였다. 1대는 초기 URL을 'http://www.naver.com'로 제공한 인덱스 서버이고 나머지 4대는 이미지 검색 에이전트들이다. 또한, 분산 환경에서의 이미지 검색 에이전트의 성능평가를 위해 에이전트가 웹 페이지 수집시 발생하는 부가적인 웹 서버와 네트워크의 부하는 무시였다.

내용기반 이미지 검색시스템에 적용하기 위한 테스트 이미지는 다양한 색과 영역의 부분 또는 전체를 차지하여 영역별 변화가 많은 꽃 이미지를 사용하였다.

4.1. 다중스레드에 의한 URL 저장 시간

하나의 이미지 검색 에이전트에서 단일스레드와 다중스레드의 URL 저장 시간을 비교하였다. 이는 CPU의 정지상태(idle)를 최소화한 상태에서 URL을 저장하는 시간을 알 수 있다. 그림 7에서와 같이 단일스레드를 이용했을 때 2,000개의 URL을 수집하는데 약 32,000초가 걸렸으며, 2개의 스레드를 이용했을 때에는 약 14,700초가 소요되었다. 그리고 4개의 스레드를 이용했을 때에는 약 3,500초가 소요되었다. 이는, 단일스레드를 이용했을 때 보다 2개와 4개의 스레드를 이용했을 때의 성능이 각각 210%와 910%의 성능 향상이 있었다.

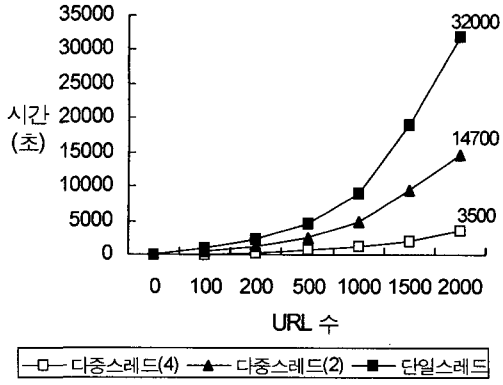


그림 7. 다중스레드에 의한 URL 저장 시간

4.2 분산 환경에서의 인덱스 저장 시간

그림 8은 검색 에이전트를 2대와 4대의 컴퓨터로

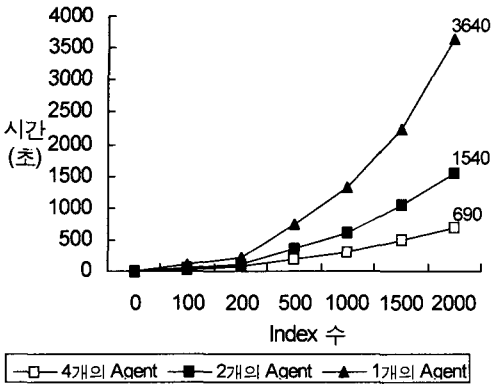


그림 8. 분산 환경에서의 인덱스 저장 시간

분산처리를 했을 때 인덱스 데이터베이스에 저장되는 인덱스 저장 시간을 나타내었다. 인덱스 저장 시간은 인덱스 서버로부터 전송받은 이미지 수집시간을 이용하였고, 분산된 에이전트에서 인덱스 서버로 이미지 수집 시간을 전달하는데 소요되는 시간은 경미하여 무시하였다.

1대의 이미지 검색 에이전트를 가동했을 때, 2,000개의 인덱스를 저장하는 시간은 약 3,640초이고, 2대의 이미지 검색에이전트를 가동했을 때와 4대의 이미지 검색 에이전트를 가동했을 때의 시간은 각각 약 1,540초 와 약 690초로 2.2배와 5.2배 이상의 인덱스 저장 시간의 차이를 나타내었다.

4.3. 내용기반 이미지 검색 결과

그림 9는 이미지 검색 에이전트에서 수집한 이미지를 내용기반 이미지 검색 시스템을 이용하여 테스트한 결과이다.

검색된 그림 중 가장 유사도가 높은 것은 오른쪽에 큰 그림을 한개 보여주고 아래쪽 화면에는 다음 유사도가 높은 것부터 낮은 것 순으로 4개씩 작은 그림으로 나타내었다. 아울러 화면 밑에 작은 그림을 클릭하면 오른쪽 위부분에 있는 큰 화면으로 그림이 나타나도록 했다.

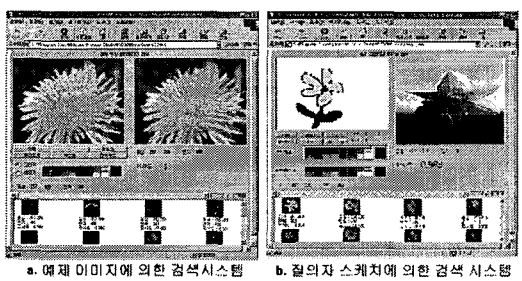


그림 9. 내용기반 이미지 검색 결과

5. 결 론

본 논문은 이미지 검색 에이전트를 플렛폼에 독립적인 자바 언어로 구현하여 분산 환경에 적합하게 하였다. 또한, 웹상의 이미지를 다중스레드를 이용한 분산 환경을 이용해 정보를 요청하고 응답받는 시간 동안 CPU 사이클의 정지 상태를 줄임으로써, 2,000개의 인덱스를 저장하는 시간이 1개의 에이전트의 경우 약 3,640초이고 4개의 에이전트를 이용해 분산처리 했을 경우는 약 690초로 분산처리 했을 경우 인덱스 저장 시간을 단축하였다. 이는 다중스레드를 이용한 분산 환경에서 이미지 수집을 520% 이상 효율적으로 한다는 것을 보여준다. 각각의 에이전트에 수집한 이미지를 JDBC를 이용하여 데이터베이스에 연결하여 저장하게 함으로써 이미지 저장 공간의 효율성을 더했다. 그리고 질의 예제에 의한 이미지 검색 및 사용자가 질의할 내용을 직접 스케치하여 이미지를 검색하는 내용기반 이미지 검색 시스템에 이용하였다.

본 논문은 다중스레드를 이용한 분산 환경에서의 이미지 검색이 목적이므로 내용기반 이미지 검색 시스템의 성능향상을 위한 인덱스 설정문제와 이미지 특징 연구를 수행할 계획이다.

참 고 문 헌

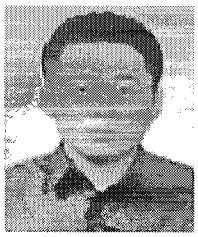
[1] http://www.isc.org.

[2] Y. Alp Aslandogan and Clement T. Yu, "Multiple Evidence Combination in Image Retrieval: Diogenes Searches for People on Web", In Proceedings of ACMSIGIR 2000, Athens, Greece, 2000.

[3] Henry A. Rowley, Shumeet Baluja, and Takek Kanade, "Neural Network Based Face Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998.

[4] Erik Selberg and Oren Etzioni, "Multi-Service

Search and Comparison Using the MetaCrawler", The Fourth International World Wide Web Conference December 11-14, Boston, Massachusetts USA., 1995.
[5] Zhihong Lu, "Scalable Distributed Architecture for Information Retrieval", University of Massachusetts, Amherst, 1999.
[6] 박명선, 이석호, "지능형 웹 영상 검색 엔진의 설계", 한국정보과학회 가을 학술 발표논문집, Vol.26, No.2, 1999.
[7] 박명선, "WISE : WWW 이미지 검색 엔진의 설계 및 구현", 서울대학교 석사논문, 1997.
[8] Remco C. Veltkamp and Mirela Tanase, "Content-Based Image Retrieval Systems: A survey", Utrecht University, 2002.
[9] Bernard K. Gunther, "Multithreading with Distributed Functional Units", IEEE, 46, 399-411, 1997.
[10] Brandon Cahoon and Kathyrn S. McKinley, "Performance evaluation of a Distributed Architecture for Information Retrieval, University of Massachusetts, Amherst, 1996.



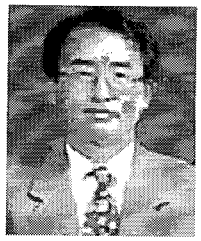
차 상 환

2002년 대구대학교 식품영양학과 (공학사)
2004년 대구대학교 컴퓨터정보공학과(공학석사)

관심분야 : 분산 멀티미디어 시스템, 영상정보검색

김 순 철

1990년 서울대학교 컴퓨터공학과 (공학사)
1992년 서울대학교 컴퓨터공학과 (공학석사)
1998년 서울대학교 컴퓨터공학과 (공학박사)
1998년 서울대학교 컴퓨터신기술 공동연구소 특별연구원
1999년~현재 대구대학교 컴퓨터·IT공학부 조교수
관심분야 : 운영체제, 분산시스템, 멀티미디어시스템



황 병 곤

1974년 경북대학교 전자공학과 (공학사)
1980년 경북대학교 전자공학과 (공학석사)
1990년 경북대학교 전자공학과 (공학박사)
1982년~현재 대구대학교 컴퓨터·IT공학부 교수
관심분야 : 멀티미디어 정보검색, 컴퓨터 그래픽스, 인터넷 응용