# 지능적 웹 이미지 검색 엔진의 설계

박명선○, 이석호
mspark@db.snu.ac.kr, shlee@comp.snu.ac.kr
서울대학교 컴퓨터공학과

# Design of Intelligent Web Image Search Engine

Myungsun Park○, Sukho Lee
Deptartment of Computer Engineering., Seoul National University

요 약

기존의 웹 이미지 검색 엔진은 웹 이미지를 검색할 때 웹 이미지의 특징과, 웹 이미지를 포함한 HTML 문서의 텍스트를 이용한다. 그러나, 텍스트는 문맥에 따라 의미가 달라질 수 있으므로, 검색 대상을 미리 분류하면 검색 효율을 높일 수 있다. 본 논문은 웹 문서의 텍스트에서 이미지와 관련이 있는 이미지 설명 텍스트를 자동으로 추출하고, 검색 효율을 높이기 위하여 웹 이미지를 자동으로 분류하는 지능적 웹 이미지 검색 엔진을 제안한다. 지능적 웹 이미지 검색 엔진은 분류와 용어, 용어와 용어 사이의 연관도를 이용하여 분류의 정확도를 높인다.

## 1. 서론

월드 와이드 웹(World Wide Web)에 존재하는 문서의 수는 매년 폭발적으로 증가하고 있으며, 텍스트로 이루어진 웹 문서를 검색하기 위하여 다수의 검색 엔진이 사용되고 있다. 웹에는 텍스트 이외에 이미지, 비디오, 오디오 등의 다양한 형태의 정보가 존재하고 있는데 최근에 웹 상의 이미지를 효율적으로 검색하려는 연구가 시도되고 있다. 웹 이미지 검색에 관한 연구는 일반적으로 이미지의 내용에 기반을 둔 검색과 이미지 주위의 텍스트를 이용한 검색을 지원하고 있다. 웹 이미지 검색 엔진은 웹 이미지의 내용 기반 검색을 위하여 이미지 검색에 관한 기존 연구 내용을 응용하고 있으며, 텍스트를 이용한 이미지 검색을 위하여 웹 문서 내의 특정 텍스트를 활용하고 있다. 그러나, 텍스트는 문맥에 따라 의미가 달라지는 문제점이 있으므로, 검색 대상을 미리 분류해 놓으면 검색 효율을 높일 수 있다[9].

본 논문은 웹 문서의 텍스트에서 이미지 설명 텍스트를 자동으로 추출하고, 추출된 텍스트를 이용하여 이미지를 자동으로 분류하는 지능적 웹 이미지 검색 엔진을 제안한다. 지능적 웹 이미지 검색 엔진은 기존의 연구와 달리 이미지 분류와 키워드, 그리고 키워드와 키워드 사이의 연관도를 이용하여 분류의 정확도를 높인다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구로 기존의 웹 이미지 검색 엔진에 대해 설명하고, 3장에서 본 논문이 제안하는 지능적 웹 이미지 검색 엔진의 설계 내용을 기술한다. 4장에서 이미지 설명 텍스트를 자동으로 추출하는 방법을 설명하며, 5장에서 결론을 맺는다.

## 2. 관련 연구

웹 상의 이미지를 검색하는 시스템은 몇 년 전부터 등장하기 시작하였는데 최근에는 상용 검색 엔진도 부수적인 서비스로 웹 이미지 검색을 지원하고 있다.

Yahoo의 Image Surfer(Interpix Software)[1]와 WebSeek[2]는 키워드 기반의 분류 트리(Category tree)를 생성하고 이를 이용하여 웹 이미지를 분류하고 검색한다. 이 시스템은 이미지와 관련된 텍스트를 이용하여 반자동으로 이미지를 분류한다. 이 시스템을 이용하면 특정 분류 안에 들어있는 이미지를 대상으로 색상 히스토그램 기반 검색을 할 수 있다. Lycos 미디어 검색 툴과 WebSeer[3]는 이미지 URL과, 이미지가 포함된 웹 문서에서 키워드를 자동으로 추출한다. WebSeer 시스템은 웹 이미지 내용을 분석하여 얻어진 이미지 헤더, 화일 종류, 크기, 날짜 등과 이미지 주위의 텍스트 정보를 이용하여 웹 이미지를 검색한다. 또한 이 시스템은 사람의 얼굴과 수평선 등의 객체를 이미지에서 자동으로 인식하고, 색상과 질감 등의 정보를 사용하여 하나의 이미지를 여러 조각으로 분할한다.

ImageRover[4]는 이미지 수집, 해석을 담당하는 이미지 수집 서브시스템과, 질의 서버와 사용자 인터페이스로 이루어진 이미지 질의 서브시스템으로 구성되어 있다. PicToSeek[5]는 이미지 카탈로그를 이용한 시각적 브라우징과 이미지 예제, 이미지 특징 질의를 지원하는 웹 이미지 검색 시스템이다. WISE[7]는 이미지 설명 텍스트와 이미지 색상 히스토그램을 이용하여 웹 이미지 검색을 지원하는 시스템이다. 이미지 설명 텍스트는 이미지 표제, 이미지 주위의 텍스트, 링크 텍스트, alt 텍스트 등을 추출하여 구성하며, 사용자가 and 또는 or로 연결한 키워드를 이용하여 검색한다.

웹 문서의 텍스트와 하이퍼텍스트 구조에 기반을 둔 웹 기반 이미지 검색 모델을 제안하고 있는 논문으로는 [6]이 있다.

## 3. 지능적 웹 이미지 검색 엔진의 설계

지능적 웹 이미지 검색 엔진은 크게 웹 문서 탐색 에이전트와 웹 문서 분석기, 텍스트 처리기, 분류 지식 생성기, 이미지 분류기, 이미지 처리기로 구성된다. 구조는 그림 1과 같다.

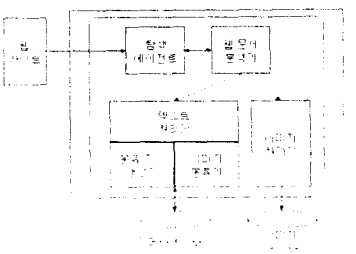

그림 1 지능적 웹 이미지 검색 엔진의 구조

* 본 연구는 한국과학재단 특성기초연구(98-0102-06-01-3)의 지원을 받았음

### 3.1 탐색 에이전트

탐색 에이전트는 웹 사이트를 방문하여 이미지와 텍스트를 수집한다(그림 2). 탐색 에이전트의 작동 순서는 아래와 같다.

- Seed URL(초기값)부터 웹 사이트를 탐색한다.
- URL을 방문하고 웹 문서를 다운로드한다.
- 다운로드한 웹 문서는 웹 문서 분석기에 보낸다.
- 웹 문서 분석기에서 돌아온 URL 집합 중 이미 방문하였거나 중복된 것을 제거하고 URL 저장소에 저장한다.
- URL 저장소에서 URL을 하나 얻는다.
- URL을 가진 웹 문서를 방문한다.

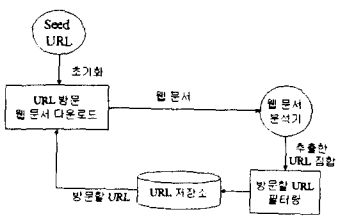


그림 2 탐색 에이전트

### 3.2 웹 문서 분석기

웹 문서 분석기는 웹 문서를 분석하여 이미지 설명 텍스트를 추출하고, 설명 텍스트와 이미지를 각각 텍스트 처리기와 이미지 처리기에 넘겨준다(그림 3).

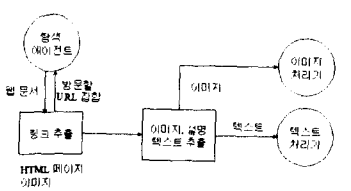


그림 3 웹 문서 분석기

웹 문서 분석기의 작동 순서는 아래와 같다.

다운로드한 웹 문서가 HTML 페이지이면 먼저 링크를 추출한다. 추출한 링크는 웹 문서 탐색 에이전트에 보낸다.

- 링크가 제거된 텍스트에서 이미지 설명 텍스트를 추출하여 텍스트 처리기에 보낸다.
- 웹 문서가 이미지이면 이미지 처리기에 보낸다.
- 웹 문서의 종류는 HTTP 헤더를 조사하면 알 수 있다. content type이 text/html이면 HTML 문서이고, image/jpeg 또는 image/gif이면 이미지이다.

### 3.3 텍스트 처리기

텍스트 처리기는 추출한 이미지 설명 텍스트를 용어로 변환한다[8](그림 4).

- 형태소 분석 : 텍스트를 word, token 단위로 분해한다.
- 불용어 필터링 : 인덱싱에 불필요한 the, of, and, to 등의 단어를 삭제한다.
- stemming : 단어의 형변화를 하나의 단어로 인덱싱하여 인덱스의 크기를 줄인다. 예를 들어 playing, played, players 등을 하나로 인덱싱한다.
- 시소러스 사용 : 용어 사이의 관계를 분석하여 종합적이고 정확한 단어를 인덱스에 유지한다. 동의어는 하나의 단어로 변환하여 단어의 처리비용을 줄인다. 이 과정에서 생성한 용어는 분류 지식 생성기에 보낸다.

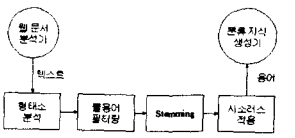


그림 4 텍스트 처리기

### 3.4 이미지 처리기

탐색 에이전트가 다운로드한 이미지는 먼저 색상 히스토그램을 계산하고, 이미지 인덱싱을 위하여 계산된 특징을 저차원으로 변환한다. 이미지 URL, 특징, 저차원 변환된 특징을 데이타베이스에 저장한다(그림 5).

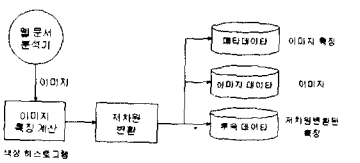


그림 5 이미지 처리기

### 3.5 분류 지식 생성기

분류 지식 생성기는 텍스트 처리기에서 보낸 용어에서 "대표 용어"와 "확장 용어"를 계산한다(그림 6).

대표 용어는 특정 분류에 속한 웹 문서들 중 특정 빈도 이상으로 나타나는 용어를 말한다. 용어의 분류에 대한 지지도가 특정 값보다 클 때, 대표 용어가 된다[9].

특정 분류에 나타나는 대표 용어의 개수는 제한되므로, 더 많은 용어가 분류와 연관이 있도록 하려면, 용어 사이의 연관 관계를 이용한다[9,10]. 즉, A라는 용어와 연관된 용어 B가 특정 분류에 대한 지지도가 크다면, A도 역시 특정 분류와 연관도가 높다고 할 수 있기 때문에 분류를 할 수 있게 된다. 이런 경우 A를 확장 용어라고 부른다.

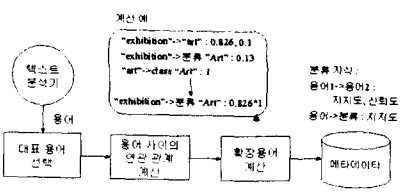


그림 6 분류 지식 생성기

예를 들어,

"exhibition" -> "art" : 신뢰도 0.826, 지지도 0.1
"exhibition" -> 분류 "Art" : 지지도 0.13
"art"-> 분류 "Art" : 지지도 1

이라면, "exhibition"은 대표 용어가 될 수 없다. 그런데, "art"-> 분류 "Art"의 지지도가 1이라는 성질을 이용하면 "exhibition" -> 분류 "Art"의 지지도를 0.826 × 1로 계산할 수 있고 이 값이 주어진 임계값보다 크다면 용어 "exhibition"은

확장 용어가 되어 이 용어를 포함한 이미지 설명 텍스트는 "Art"로 분류할 수 있다.

분류 지식 생성기에 사용되는 데이타는 다음과 같다.
- 용어 사이의 연관도 : 용어1, 용어2, 신뢰도, 지지도
- 용어, 분류 사이의 연관도 : 용어, 분류, 지지도

이미 분류된 사이트의 웹 이미지를 모두 탐색하고 나면 용어와 분류 사이의 연관도를 모두 구할 수 있고, 분류 계층에 있는 모든 분류에 대해 구한 용어와 분류 사이의 연관도가 분류 지식이 된다. 본 시스템은 Yahoo의 분류 계층을 기준으로 웹 이미지를 분류한다.

### 3.6 이미지 분류기

이미지 분류기는 분류 지식 베이스를 구축한 다음 분류되지 않은 이미지에 대해 이미지 분류를 수행한다(그림 7).

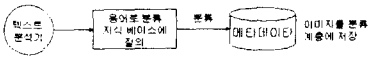


그림 7 이미지 분류기

분류되지 않은 웹 이미지의 설명 텍스트에서 추출한 용어를 가지고 분류 지식 베이스에 질의하여 가장 높은 연관도를 갖는 클래스로 이미지를 분류한다. 그림 8은 상품 카탈로그 이미지에 대한 이미지 분류 계층의 예이다.


그림 8 이미지 분류 계층의 예

## 4. 이미지 설명 텍스트 추출 기법

이미지 설명 텍스트는 이미지와 연관성이 높은 텍스트를 웹 페이지에서 선정한다[7,6]. 이러한 조건을 만족하는 텍스트는 다음과 같은 유형에서 찾을 수 있다. 그림 9는 웹 페이지에 나타나는 이미지 설명 텍스트 유형의 예이다.

- 이미지 아래에 덧붙은 캡션
- 웹 페이지 안의 텍스트 중에서 '그림 1은 ... 이다'라고 서술한 부분
- 이미지가 존재하는 디렉토리 이름, 이미지 화일 이름
- 이미지 바로 옆에 나타나는 텍스트
- 하나의 웹 페이지에 여러 이미지가 들어 있는 프레임과 그 이미지들의 링크를 가지고 있는 프레임으로 구성되어 있는 경우
- 일반적으로 이미지 링크에 덧붙인 텍스트
- (img src="...")에 나타나는 alt="..." 텍스트

이러한 유형을 갖는 텍스트를 이미지 설명 텍스트로 선정하고, 이미지 분류를 위해 텍스트를 처리할 때 유형에 따른 가중치를 용어의 발생 빈도로 곱하여 유형에 따른 설명 텍스트의 연관성을 반영한다.

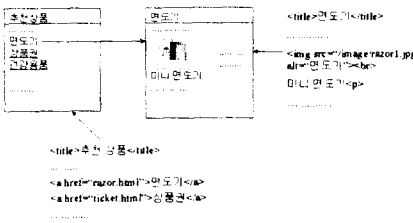


그림 9 이미지 설명 텍스트 유형의 예

## 5. 결론

웹 이미지 검색에 대한 기존 연구는 웹 이미지의 특징과 함께, 웹 이미지가 포함된 HTML 문서에서 특정 위치의 텍스트를 검색에 이용하고 있다. 그러나, 텍스트는 문맥에 따라 의미가 달라지므로 텍스트를 이용하여 웹 이미지를 분류하고 질의할 때 대상 이미지의 분류를 지정하면 검색 효율을 높일 수 있다.

본 논문은 이미지 설명 텍스트를 이용하여 웹 이미지를 계층으로 분류하고 질의할 때 분류를 명시함으로써 검색 효율을 높이는 지능적 웹 이미지 검색 엔진에 대해 설명하였다. 지능적 웹 이미지 검색 엔진은 데이타마이닝의 연관 규칙 탐사 기법을 이용하여 용어 사이의 연관 관계를 계산하고 이를 이용해 분류 효율을 높인다. 본 검색 엔진은 C와 Perl을 이용하여 부분 구현되어 있다.

## 참고 문헌

[1] Yahoo's Image Surfer, http://ipix.yahoo.com.
[2] J. R. Smith, S. -F. Chang, "An Image and Video Search Engine for the World-Wide Web", Symposium on Electronic Imaging: Science and Technology - Storage & Retrieval for Image and Video Databases V, San Jose, CA, February 1997.
[3] Charles Frankel, Michael J. Swain, Vassilis Athisos, "WebSeer: An Image Search Engine for the World Wide Web", TR 96-14, U. Chicago, 1996.
[4] Stan Sclaroff, Leonid Taycher, Marco La Cascia, "ImageRover: A Content-Based Image Browser for the World Wide Web", Proc. IEEE Workshop on Content-based Access of Image and Video Libraries, 1997.
[5] Theo Gevers, "PicToSeek: A Content-Based Image Search System for the World Wide Web", Proc. Visual '97 1997.
[6] V. Harmandas, M. Sanderson, M. D. Dunlop, "Image retrieval by hypertext links", ACM SIGIR '97, 1997.
[7] 박명선, 송병호, 이석호, "WISE : WWW 이미지 검색 엔진", 정보과학회 논문지, 제4권, 제3호, pp.305-313, 1998.
[8] Frakes, W. B., Baeza-Yates, R., "Information Retrieval-Data Structures & Algorithms", Prentice Hall, 1992.
[9] Shian-Hua Lin, Chi-Sheng Shih, Meng Chang Chen, Jan-Ming Ho, "Extracting Classification Knowledge of Internet Documents with Mining Therm Associations: A Semantic Approach", ACM SIGIR '98, 1998.
[10] Rakesh Agrawal, Ramakrishnan Srikant, "Fast Algorithm for Mining Association Rules", Proceedings of the 20th VLDB Conference, 1994.