제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

이미지의 속성 및 링크 정보를 이용한 이미지 검색 시스템

한기덕*, 정성원*, 윤근수**, 권혁철*
*부산대학교 정보컴퓨터공학부
**울산과학대학 컴퓨터정보학부
e-mail : templer@pusan.ac.kr, swjung@pusan.ac.kr,
ksyun@mail.ulsan-c.ac.kr, hckwon@pusan.ac.kr

Image Retrieval System Using Image Attributes and Links

Gi-deok Han*, Sung-won Jung*, Keun-soo Yun**, Hyuk-chul Kwon*
*Dept. of Computer Science and Engineering, Pusan National University
**Dept. of Computer Information, Ulsan College

요 약

컴퓨터와 네트워크의 처리속도 증가와, 인터넷의 발달로 인하여 이미지, 사운드, 동영상 등 각종 멀티미디어 정보가 인터넷상에 다수 등록되고 있으며, 이에 대한 검색 요구도 증가하고 있다. 그에 따라 다양한 멀티미디어 정보 검색을 위한 방법이 연구되고 있지만, 그에 대한 활용도는 미미하며, 데이터베이스에 등록된 단순 멀티미디어 정보 검색에 머물고 있는 실정이다. 이에 본 연구는 인터넷상의 멀티미디어 정보 중 이미지 정보를 능동적으로 수집, 정보를 추출하여 검색에 이용한다. 이를 위하여, 이미지에 대한 text정보와 이미지의 속성 및 Link 정보를 이용, 의미 있는 이미지와 의미 없는 이미지를 분류하여 검색의 효율을 높이고, 속성 및 Link 정보를 가중치로 사용함으로써 검색 시 이미지의 중요도를 평가할 수 있도록 한다.

1. 서론

현대의 사회는 빠른 속도로 정보 사회로 변화하고 있다. 인터넷의 빠른 보급으로 인해 국내의 웹 호스트 수가 2001년에 70만대에 달했고, 2002년 11월에는 정보통신부 국내 초고속 인터넷 가입자 수가 1,000만 명을 넘어섰다고 발표했다.[1] 또한, 디지털 산업의 성장에 발 맞춰 mp3 player, 디지털 카메라, 캠코더 등이 출현하였고, 사용자들은 기존의 text정보뿐만 아니라 멀티미디어 정보도 인터넷에 등록하기 쉽게 되었다.

이에 따라, 이미지, 동영상, 음성 등과 같은 멀티미디어 정보에 대한 검색 요구가 증가하였다[2]. 현재 일부 정보 검색 시스템에서는 이미지에 대한 검색을 서비스하고 있지만, 대부분의 경우에는 자동화된 정보 수집과 검색 방법을 사용하지 않고 인력에 의한 분류 작업을 통해서 서비스를 제공하고 있는 실정이고, 자동화된 정보 수집과 검색을 제공하는 시스템은 엄청난 이미지의 양에 대해서 효율적인 수집과 검색 성능을 나타내지 못하고 있다.

그래서 본 논문은 효율적인 멀티미디어 정보검색에 대한 연구 및 실험을 하였다. 구체적인 방법으로 멀티미디어 정보 중에서 의미가 있는 정보를 자동적으로 수집하고 이를 표현할 수 있는 핵심어(keyword)를 추출하는 정보 추출 에이전트(information extraction agent : Wrapper) 시스템[3]에 더욱 높은 성능을 위해서 이미지의 의미의 유무를 판단하는 시스템을 추가하여 그 성능을 테스트한다.

인터넷의 이미지에는 수많은 종류가 있으며, 우리는 크게 의미 있는 이미지와 의미 없는 이미지로 나눌 수 있다. 이 처리를 하기 위하여 이미지의 여러 가지 속성 중 의미의 유무를 판단할 수 있는 속성을 이용하여 주어진 이미지에 대해 의미가 있는지, 없는지를 결정한다.

이전의 시스템에서는 이미지의 파일 크기에 의해 필터링을 하였으나, 본 논문에서는 성능을 향상시키기 위해서 이미지의 의미 유무를 판단할 수 있는 속성으로 이미지의 넓이와 가로, 세로의 비, 사용된 색상 및 정보 추출 에이전트에서 얻을 수 있는 Link 정보를 이용하여, Hill Climbing 알고리즘[4]을 적용한 후 이미지의 의미의 유무를 판단한다. 이 결과는 검색기에서 가중치로써 재사용된다.

2. 관련 연구

Dunlop은 하이퍼텍스트(hypertext) 구조를 분석하여 멀티미디어 정보를 검색하는 모델을 제시하였다[5]. 즉, 링크 정보의 문자 정보를 텍스처 노드(textual node)로 그 외의 멀티미디어나 문서, 압축 파일 정보를 비텍스처 노드(non-textural node)로 정의한 후, 비텍스처 노드로 링크되는 텍스처 노드에 대한 전체 문서 집합에 대한 연결망(network)의 분석을 이용한 시스템을 제안하였다.

Dunlop은 비텍스트 노드의 주변 문맥을 caption, neighbouring caption, one step link text와 같이 3가지 집합으로 분리하여, 링크 구조와의 가중치를 계산하여 검색하는 방법을 이용하였다. 마지막으로 Dunlop은 이미지 정보에 대해서 적합하지 않은 정보를 기능성(functional) 이미지로 정의하였는데 이러한 이미지는 링크 연결망에서 높은 링크 수를 나타낸다는 특징을 이용하여 제거하는 방법을 사용하였다[6].

하지만, Dunlop system의 실험 결과를 보면 galley site와 같은 제한된 영역의 제한된 질의어에서만 우수한 성능을 보였다. 그리고 실제 자료가 자주 갱신되거나 사라지거나 생성되기도 하는 인터넷상에서 연결망을 구축하고 계산한다는 것은 그 한계가 있다.

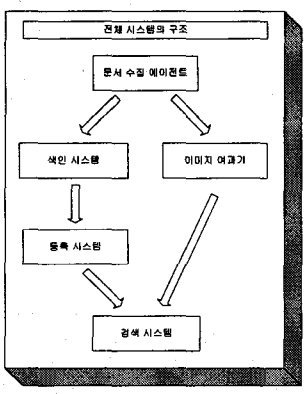
Google에서도 현재 이미지 검색 시스템을 구현하여 서비스를 하고 있다. Google이 사용하는 방식은 이미지 정보의 anchor text와 파일명, 그리고 이미지가 포함된 문서의 문자 정보에 대해서 자체적인 분석 알고리즘을 통해 이미지를 수집하고 검색하는 서비스를 제공하고 있다[7]. 하지만, 실제 Google 시스템을 사용해보면 불필요한 이미지가 결과로 제공되는 양이 너무 많다[8]. 특히 검색된 결과 이미지 중에는 크기가 아주 작은 이미지도 상당수가 존재한다. 이러한 결과는 사용자에게 많은 불편을 주기 때문에 불필요한 이미지를 제거하는 기능이 필요하다.

현재까지 국내에서는 네이버, 엠파스, 라이코스 코리아, 야후 코리아에서 이미지 검색을 제공하고 있다[9]. 하지만, 실제 검색을 해보면 자동적인 이미지 정보 수집에 의한 검색이 아닌 인력을 동원하여 분류 작업을 거친 시스템이 대부분이었다. 실제 야후 코리아 같은 경우는 이미지 전문 데이터베이스를 구축한 에이스파이더라는 업체의 시스템을 제공받고 있다[10].

이와 같은 문제점을 극복하기 위하여 선행 연구로써, 멀티미디어 정보의 추출 및 핵심어 결정 알고리즘을 적용한 정보 검색 에이전트를 개발했었다. 이 시스템에서 적용한 핵심어 결정 알고리즘은 멀티미디어의 파일명, Alt, Anchor, 디렉토리명, Title, 앞문장, 뒷문장에서 총 7개의 후보 핵심어를 추출한 후, 가중치에 의해 멀티미디어의 핵심어를 결정하는 것이었다. 선행 연구에 의해 개발된 이 시스템에 이미지 여과기 및 검색 기능을 향상시킨 시스템이 본 논문의 시스템이다.

3. 전체 시스템의 구조

전체 시스템은 문서 수집 에이전트 시스템, 이미지 여과기 시스템, 색인 시스템, 등록기 시스템, 검색 시스템으로 이루어져 있다. [그림 3-1]은 전체 시스템의 구조를 보여준다.



[그림 3-1] 정보 검색 에이전트 전체 구조도

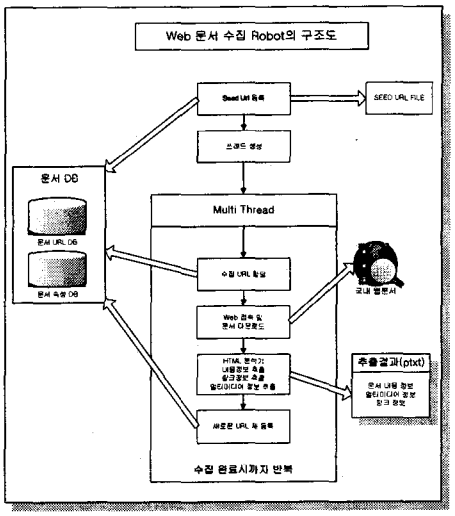
3.1 문서 수집 에이전트

문서 수집 에이전트는 크게 두 가지 기능을 수행한다.

첫째로 인터넷 문서를 수집하는 기능이다. Socket으로 해당 문서의 web server 접속한 후, HTTP 프로토콜에 따라서 문서를 요구하고 받아온다. 받아온 문서 URL을 데이터베이스로 관리하고 문서의 정보를 사전으로 구성하는 기능을 수행한다.

두 번째로 문서를 분석하는 기능이다. Html 파서(parser)를 이용해서 태그별로 분류하고 태그가 아닌 문서의 내용을 문서 정보검색을 위해서 저장한다. 그리고 문서의 링크 정보를 분석하여 멀티미디어 정보를 추출한다. 이러한 과정을 오프라인으로 수행할 수도 있지만, 일관성을 위하여 온라인에서 이러한 과정을 수행하는 것이 더 효율적이다[13].

문서 수집 에이전트에서 이미지의 Link정보를 추출하여 저장하게 되는데, 그 이유는 인터넷 상의 이미지에 걸린 Link는 이미지의 가치 여부를 판단하는 중요한 척도가 되기 때문이다. 예를 들면, 기능성 이미지의 경우 다른 이미지 또는 문서를 Link하는 이미지는 대부분이 기능성 이미지이며, 다른 이미지에 의해 Link가 걸리는 이미지는 다른 이미지에 비해 가치가 높다. 이런 이미지의 링크 정보 2가지를 문서 수집 에이전트의 문서 수집 과정에서 추출해 내어, 이미지의 가치를 평가하는 척도로써 사용한다. [그림 3-2]는 문서 수집 에이전트의 구조를 보여준다.

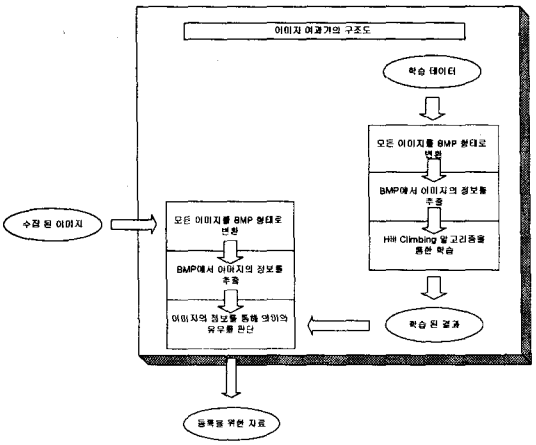


[그림 3-2] 문서 수집 에이전트의 구조

3.2 이미지 여과기

문서 수집 에이전트에서 수집한 이미지에는 기능성 이미지와 같은 불필요한 이미지가 많이 포함되어 있다. 따라서, 검색 효율 및 정확한 이미지의 제공을 위해서 불필요한 이미지를 제거해야 하는데, 이 처리를 하는 시스템이 이미지 여과기이다.

이미지 여과기에서 이미지의 가치를 평가하는 기준은 이미지의 넓이와 가로, 세로의 비, 사용된 색상 및 정보 추출 에이전트에서 추출한 Link를 사용한다. 이를 이용하여 Hill-Climbing 알고리즘을 적용하여, 이미지의 가치를 평가하고 가치가 낮은 이미지는 여과한다. [그림 3-3]은 이미지 여과기의 구조를 보여준다.



[그림 3-3] 이미지 여과기의 구조도

3.3 색인, 등록 시스템

색인 시스템은 정보검색 시스템의 저장 공간과 검색의 효율성을 고려하여 문서를 대표할 수 있는 색인어를 정확히 추출해야 한다. 등록 시스템은 문서에서 추출한 색인어를 검색에 용이한 형태로 사용하기 위해서 색인어 역파일, 포스팅 파일을 구성한다.

3.5 검색 시스템

일반 문서 검색에서는 문서 내의 색인어의 위치와 출현 빈도에 대해서 exact-matching 연산을 통해서 검색된 문서의 순위를 조정한다. 하지만, 멀티미디어 정보검색에서는 핵심어에 대한 색인만을 수행하기 때문에 출현 빈도가 사실상 유무로만 존재한다. 그러므로 다른 순위화 알고리즘을 사용해야 하는데, 현재 검색 시스템에서는 이미지의 의미의 유무를 판단하는데 사용한 Hill-Climbing 알고리즘의 적용 후의 결과로써 나온 수치를 이미지의 가중치로써 순위화를 구현하였다.

4. 이미지 여과기의 구현

인터넷 상의 이미지는 GIF, BMP, JPG 등 매우 다양한 포맷이 있다. 따라서 이런 모든 그림 파일로부터 파일의 정보를 얻기 위해서는 각각의 그림 파일 포맷에서 정보를 추출하는 시스템을 구현하거나 각각의 파일 포맷을 하나의 정해진 파일 포맷으로 바꾼 후, 바뀐 파일 포맷에서 그림 파일 정보를 추출하는 시스템을 구현하여야 한다.
본 논문에서는 2번째 시스템의 형태로 구현한다.

4.1 이미지 파일 포맷 변환기

이미지 파일 중 BMP 파일은 픽셀 정보에 대한 압축 처리를 하지 않은 형태로써, 이미지에 대한 정보를 추출하기에 용이하다. 그래서 인터넷 상의 다양한 이미지를 BMP 파일의 형태로 변환하여 이미지의 정보를 추출하기 위해 이미지 파일 포맷 변환기를 구현한다. 이 시스템은 BMP, JPG, GIF 이상 3가지를 처리하며, 각 이미지 포맷에서 다른 이미지 포맷으로 변환하는 기능이 있다.

4.2 BMP 파일 정보 추출기

BMP 파일은 픽셀에 대한 처리를 하지 않았기 때문에, 손쉽게 픽셀 정보를 얻을 수 있다. 본 논문에서 사용하는 이미지 관련 정보 중 3가지를 BMP 파일 정보 추출기에서 추출한다. 추출하는 정보는 이미지의 가치를 나타낼 수 있는 요소인, 그림의 면적(Width × Height), 사용한 색상, 가로 길이와 세로 길이의 비율(Width : Height)이다.

4.3 Hill Climbing 알고리즘의 적용

학습 대상 이미지에서 정보를 추출하여, 수작업으로 의미있는 그림과 의미 없는 그림으로 학습 데이터를 만든다. 분류한 정보와 추출한 정보를 이용하여 Hill Climbing 알고리즘을 적용하여, Local Maxima 값을 구한 후, 임의로 초기값을 변경하여, 몇 개의 Local Maxima을 구한다. 이 중에서 가장 높은 수치를 선택한다. Hill Climbing 알고리즘에 적용한 식은 다음과 같다.

적용한 식 :

$$S = \alpha C + \beta A + \gamma R + \delta L_a + \varepsilon L_b$$

S : 계산 결과값
C : 이미지에서 사용한 색상 수
A : 이미지의 면적
R : 이미지의 가로, 세로의 비
L_a : 이미지에 링크가 걸려서 다른 문서를 가리키고 있는지의 여부
L_b : 어떤 이미지가 자신에게 링크를 걸고 있는가의 여부
$\alpha, \beta, \gamma, \delta, \varepsilon$: 가중치

S의 계산 값이 Threshold보다 크면, 의미 있는 그림으로 간주하고, S의 계산 값이 Threshold보다 작다면, 의미 없는 그림으로 간주한다.

5. 실험 및 결과

본 논문에서는 Threshold를 0으로 설정하였고, 가중치 변화 값은 0.01로 설정하였다. 학습 및 테스트로 사용한 이미지는 인터넷상에서 문서 수집 에이전트에 의해 검색된 이미지를 사용하였고, 정보의 크기를 줄이기 위해 정규화 작업을 하였다. 이렇게 얻은 학습 데이터의 정보를 통해 Hill Climbing 알고리즘에 의한 학습을 시킨 후, 학습 결과의 가중치를 이용하여, 테스트 이미지의 의미의 유무를 판단하였다.
[표 5-1],[표 5-2]는 실험에 사용한 이미지 데이터이다.

	의미있는 그림	의미없는 그림	합계
사용한 Image 수	1012개	4998개	6000개

[표 5-1] 학습 데이터로 사용한 이미지

	의미있는 그림	의미없는 그림	합계
사용한 Image 수	8301개	67969개	76260개

[표 5-2] 테스트 데이터로 사용한 이미지

학습 데이터를 가지고 Hill Climbing 알고리즘을 적용하여 여러 개의 Local Maxima을 구한다.
[표 5-3]은 구한 Local Maxima의 예이다

	색상	면적	비	링크 A	링크 B	의미 없는 그림의 여과 성공 수	의미 있는 그림의 추출 성공 수
1	0.01	2.01	-2.07	-1.13	2.03	4992	1003
2	1.02	2.00	-2.03	-1.52	1.32	4044	671
3	1.53	1.10	-1.12	-1.91	1.10	4087	1011
4	**0.16**	**1.53**	**-1.54**	**-1.30**	**1.21**	**4997**	**1012**
5	0.34	0.51	-3.92	-2.41	0.52	4895	919
6	0.10	2.03	-2.03	-0.33	0.83	4982	1010
7	0.52	1.57	-0.92	-0.40	1.24	4730	758

[표 5-3] Local Maxima 값

위의 표에서 보는 바와 같이 4번째 실험에서 가장 높은 수치가 나타났다. 실험 결과 가장 성공 수가 높은 수치인 0.16 , 1.53, -1.54, -1.30, 1.21를 사용하여, 테스트 데이터의 이미지를 의미 있는 이미지와 의미 없는 이미지로 분류한다.

[표 5-4],[표 5-5]는 테스트 데이터를 처리한 결과이다.

	이미지의 수 (개)
의미 있는 이미지를 제대로 추출한 수	8149
의미 없는 이미지를 제대로 여과한 수	65168
의미 있는 이미지를 여과한 수(여과 실패)	161
의미 없는 이미지를 추출한 수(추출 실패)	2782
합계	76260

[표 5-4] 처리 결과

	성공률(%)
의미 있는 이미지의 추출 성공률	98.16
의미 없는 이미지의 여과 성공률	95.87

[표 5-5] 처리 성공률

이미지의 알고리즘 적용 결과(수치 값)는 검색 시, 각 이미지의 가중치로써 사용되면, 가중치가 높은 이미지의 경우, 웹상에서 보여줄 때, 상위 위치에서 보여주게 된다.

6. 결론 및 향후 과제

이 논문은 인터넷 이미지 정보를 자동적으로 추출하여, 사용자에게 정확한 정보를 제공하기 위한 시스템을 구현하기 위해 이미지의 속성 및 링크 정보를 이용, 이미지의 가중치를 부가하는 방법을 제안하였다. 이미지의 가중치는 이미지의 의미 유무에 의한 여과 및 이미지의 순위화에 사용되어, 검색 효율을 향상시켰다. 그러나 이미지의 의미 유무를 판단하는 기준은 본 논문에서 언급한 것 이외에도 있을 수 있으므로, 다른 기준도 찾아서 실험을 해 보아야 할 것이다. 또한, 이미지뿐만이 아니라, 음성, 영상 데이터 같은 멀티미디어에 대해서도 연구를 해야 할 것이다.

참고문헌

[1] 한국 인터넷 정보 센터, http://www.nic.or.kr/

[2] 신봉기, 김영환, "인터넷 정보검색 서비스 동향", 한국정보과학회 1998년 8월, Vol. 16, No. 8, pp. 16-20, 1226-2315.

[3] 박태원, "인터넷 멀티미디어 정보의 추출과 핵심어 결정 알고리즘", 부산대학교 전자계산학과 이학석사 학위논문, 2002년.

[4] Stuart Russell, Peter Norvig, "Artificial Intelligence A Modern Approach", Prentice Hall, 1995년

[5] M. D. Dunlop, "Multimedia Information Retrieval", Glasgow University Computing Science Research Report 1991/ R21, October 1991.

[6] V. Harmandas, M. Sanderson and M. D. Dunlop, "Image retrieval by hypertext links", Proceedings of SIGIR-97, 1997.

[7] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", 7th International World Wide Web Conference, 1998.

[8] Google 홈페이지, http://www.google.co.kr/

[9] 네이버 홈페이지, http://www.naver.co.kr/

[10] 야후 홈페이지, http://www.yahoo.co.kr/