기계학습 기반의 웹 이미지 분류

조 수 선[†]·이 동 우[††]·한 동 원[†††]·황 치 정[††††]

요 약

HTML 페이지로 대표되는 웹 문서에서 이미지는 매우 큰 비중을 차지하고 있지만 이에 대한 분석 및 이해에 관한 연구는 활발하게 진행되지 못하고 있다. 여러 가지 웹 이미지들은 중요한 정보를 전달하기도 하지만 그렇지 않은 것들도 있다. 본 논문에서는 현재 서비스중인 인터넷 사이트의 웹 이미지들을 수집하여 기계학습(machine learning)에 기반한 분류(classification)를 통해 제거 가능한 이미지와 제거 불가능한 이미지의 두가지 클래스로 분석해 본다. 이를 위해 16개의 독특하고 풍부한 웹 이미지 특징들을 발굴하고 베이지안 기법과 결정 트리 기법을 사용하여 실험하였다. 그 결과 각각의 기법에서 87.09%, 82.72%의 F-measure 값을 얻었으며 특히, 특징 그룹의 비교 실험을 통해 본 연구에서 추가한 특징들이 매우 유용한 것임을 입증하였다.

A Machine Learning Approach to Web Image Classification

Soosun Cho[†]·Dongwoo Lee[††]·Dongwon Han[†††]·Chi-Jung Hwang[††††]

ABSTRACT

Although image occupies a large part of importance on the Web documemts, there have not been many researches for analyzing and understanding it. Many Web images are used for carrying important information but others are not used for it. In this paper we classify the Web images from presently served Web sites to erasable or non-erasable classes, based on machine learning methods. For this research, we have detected 16 special and rich features for Web images and experimented by using the Baysian and decision tree methods. As the results, F-measures of 87.09%, 82.72% were achived for each method and particularly, from the experiments to compare the effects of feature groups, it has proved that the added features on this study are very useful for Web image classification.

키워드 : 이미지 분류(Image Classification), 기계학습(Machine Learning), 웹 이미지 특징(Features of Web Images), 베이즈 분류기(Bayes Classifier), 결정 트리(Decision Tree)

1. 서 론

오늘날 인터넷은 시공간의 제약을 극복하여 언제 어디서나 어떤 매체를 통해서도 자유로이 접근할 수 있도록 발전하고 있다. 따라서 인터넷 컨텐트의 대표 역할을 하는 웹 문서는 여러가지 포맷으로 다양한 단말을 통해 나타날 수 있어야 하고 이때 웹 문서에서 큰 비중을 차지하는 이미지들이 선택적으로 표현될 수 있어야 한다. 따라서 웹 문서에 포함되어 있는 다수의 이미지들이 제거될 수 없는 고유의 정보를 가지는지 그렇지 않은지 자동 분류할 수 있다면 웹 이미지의 선택적 표현을 위한 중요한 역할을 할 수 있을 것이다.

웹 이미지 분석에 대한 연구는 여러 가지 목적으로 진행되어 왔다. 웹 이미지의 트랜스코딩을 위해 JPEG, GIF 등 이미지 데이터의 자체적인 특성(characteristics)을 분석하여 통계치를 내는 연구 방향이 있었고[7, 8] 검색 엔진에서 텍스트 외에 이미지를 이용하기 위해 사진, 그래픽 등으로 웹 이미지를 분류하는 연구도 있었다[5]. 이와 같은 연구에서는 웹 이미지의 분류를 위한 각종 휴리스틱을 개발하고 임의의 기준값을 사용하여 웹 이미지를 분류하는 방법을 사용함으로써 적용되는 이미지의 종류에 의존적이며 보편적인 방법을 제공하지 못하는 한계가 있다. 실제로 연구[5]도 사람의 얼굴 이미지에만 적용한 것이다.

본 연구에서는 기계학습에 기반한 분류법을 사용하여 웹 이미지를 분류함으로써 웹 이미지 분석을 위한 기존의 연구들에 비해 보다 보편적이며 효과적인 방안을 제시한다. 기계학습 기법을 이용하여 웹 문서의 텍스트나 테이블 등을 분석하기 위한 연구[2, 11]는 활발하게 발표되고 있지만 웹 이미지를 분석하는 연구는 양적으로나 질적으로 매우 빈약하다. 웹 문서상에서 중요한 비중을 차지하는 이미지에 대한 기계학습 기반의 분류는 웹 마이닝, 웹 컨텐트 압축, 협대역 단말(narrow-bandwidth devices)로의 컨텐트 전송 등에 응용되어 시공간을 극복한 인터넷 컨텐트의 전파에 기여할 수 있다. 본 연구에서는 이를 위해 여러가지 웹 이미지

† 정 회 원 : 한국전자통신연구원 정보가전연구부 선임연구원
†† 정 회 원 : 한국전자통신연구원 정보가전연구부 연구원
††† 정 회 원 : 한국전자통신연구원 책임연구원 휴대클라이언트 연구팀장
†††† 종신회원 : 충남대학교 컴퓨터과학과 교수
논문접수 : 2002년 8월 12일, 심사완료 : 2002년 11월 11일

특징(feature)들을 추출하여 적용하였으며 실제로 서비스 중인 인터넷 쇼핑 웹 사이트들에 적용 실험하여 의미있는 결과를 얻을 수 있었다.

본 논문은 이후 다음과 같이 전개된다. 서론에 이어서 제 2장에서는 웹 이미지 분류에 대한 관련 연구들을 알아보고 제 3장에서는 본 연구에서 채택한 두 가지 분류법인 베이지안 기법과 결정 트리 기법에 대하여 간단히 소개한다. 제 4장에서는 적용한 웹 이미지 특징들을 설명하고 제 5장에서는 실험 및 결과를 소개하며 제 6장에서 결론을 맺는다.

2. 관련 연구

본 연구는 휴리스틱 방법을 사용한 기존의 웹 이미지 분류에 대한 한계를 극복하기 위해 주로 웹 문서의 텍스트나 테이블의 분류에 사용되어 온 다양한 기계학습 기법을 웹 이미지 분류에 이용하였다. 이를 위해 다양하고 풍부한 웹 이미지 특징들을 발굴하여 적용하였으며 어떤 종류의 특징들이 웹 이미지 분류에 더 효과적인지 알아보고자 하였다. 또한 대표적인 두 가지 기계학습 기법인 베이지안 기법과 결정트리 기법을 사용하여 그 결과를 서로 비교하였다.

많지는 않지만 웹 이미지 분류에 기계학습을 적용한 기존의 몇몇 연구가 있다. 대표적인 것으로 연구[9]는 웹 문서상의 이미지의 목적을 찾기 위한 것이다. 이 연구에서는 웹 이미지를 컨텐트 이미지와 비컨텐트 이미지로 분류하기 위해 '파일 포맷', '컬러 수', '그래이 레벨', '명도', '채도', '가로 길이', '세로 길이' 등 이미지 고유의 특징들과 함께 HTML에 포함되어 있는 라벨 정보 등을 특징으로 이용하였으며 결정 트리 기법을 사용하여 84%의 정확도로 웹 이미지들을 분류하였다. 이 연구는 결정 트리 기법 한가지만을 사용하여 전체 특징 그룹을 한꺼번에 적용하였으므로 여러 가지 분류 기법을 비교 실험하거나 특징들의 종류별로 유효성을 검증하지는 못하였다. 본 논문에서는 연구[9]에서 사용된 웹 이미지 특징들 중 대표적인 것들을 포함하면서도 보다 풍부하고 의미있는 16개의 웹 이미지 특징 집합을 도출하였으며 이 특징들을 종류별로 묶어서 적용함으로써 어떤 종류의 특징들이 웹 이미지 분류에 더 많은 기여를 하는지 실험하였다. 그 결과 기존 연구에서 사용되지 않았고 본 연구에서 새로이 도출된 많은 특징들이 분류의 정확성을 나타내는 척도인 F값의 향상에 상당히 기여하는 것을 알 수 있었다. 자세한 것은 '제 5장 실험 및 결과'에서 다룬다.

또 다른 기계학습 기반의 웹 이미지 분류로서 영상분류기 개발에 관한 연구[12]가 있다. 이 연구에서는 유해 영상물 분류를 위해 베이지안 및 K-nearest-neighbor 기법을 사용하여 두 가지 기법을 비교 실험하였으며 특히 2단계 투표 방식을 채택해 정확도를 높인 것으로 나타나있다. 하지만 대상으로하는 JPEG 이미지의 특징에 대한 설명이 전혀 없고 2단계 투표 알고리즘 및 그 구현에 대해서도 기술된 바가 없다. 실험 결과는 80%에 가까운 정확도를 보이고 있지만 수행한 실험 및 이미지 특징에 대한 구체적인 설명이 없는 점이 한계이다.

한편 웹 이미지가 아닌 웹 문서의 테이블을 분류한 연구들은 최근 활발하게 발표되고 있다[2, 11]. Yalin Wang[11]은 웹 페이지 내에서 테이블이 원래 고유한 테이블의 목적으로 사용된 경우와 그렇지 않은 경우로 분류하였는데 이를 위해 결정 트리 기법과 SVM(Support Vector Machines)을 사용하였다. 본 연구는 웹 문서의 효과적인 분석을 위해서는 테이블 못지 않게 웹 이미지가 주요한 역할을 할 것임에 착안하여 웹 이미지의 분류를 위한 풍부한 특징들을 발굴하여 종류별로 적용 실험하였으며 현재 서비스되고 있는 여러 인터넷 쇼핑 사이트의 웹 문서를 수집하여 이에 포함된 다양한 웹 이미지들을 분류하였다.

3. 분류 기법

일반적으로 기계학습에 기반한 클래스 분류는 훈련(training)과 분류(classification), 두 단계로 구성된다. 분류 가능한 클래스가 미리 알려져 있고 이 알려진 클래스로의 분류가 수행되면 이를 'supervised calssification'이라 한다. 본 연구에서는 '제거 가능', '제거 불가능'의 두 가지 클래스를 미리 정하고 있으므로 이에 해당한다. 문서 분류나 정보 검색 분야에서는 오랫동안 다양한 분류 기법들이 사용되어 왔다. 본 연구의 웹 이미지 분류에서는 통계적 가정 및 확률 이론에 기반을 둔 나이브 베이즈 분류기(naive Bayes classifier)와 이질적인 특징들을 가진 데이터의 분류에 적합한 결정 트리(decision tree)를 이용한 분류 기법을 채택한다. 이 기법들은 문서 정보 등의 'supervised calssification'을 위해 가장 널리 채택되어 왔고 그 효과가 오래전부터 검증되어 왔다[10].

3.1 나이브 베이즈 분류기

베이지안 학습 기법은 가정에 대한 명확한 확률을 계산하고 이를 분류에 이용한다. 분류하고자 하는 데이터의 각 케이스들은 여러 가지 속성들(attributes)을 값으로 가지고 있는데 이 속성값들을 조건으로 하는 클래스의 확률값으로부터 각 케이스가 속하는 클래스를 결정하는 것이다. 베이즈 분류기의 기초가 되는 베이즈 정리(Baye's Theorem)로 이를 설명하면 다음과 같다. 분류 클래스를 C라고 하고 각 속성들을 A_i라고 할 때, 클래스 C는 일반적인 값으로 c_j를 갖고 속성값의 집합은 $e_k = \{A_1 = a_{1k}, \cdots, A_m = a_{mk}\}$로 표시할 수 있다. 이때 각 속성값이 주어진 조건하에서 클래스 값의 사후 확률(posterior probability)은 다음과 같은 계산식으로 구할 수 있고 베이즈 분류기는 이 사후 확률 값을 최대로 만들어주는 클래스 값 c_j를 채택하여 해당 케이스를 그 클래스에 분류한다.

$$p(c_j \mid e_k) = \frac{\prod_{i=1\cdots m} p(a_{ik} \mid c_j)\, p(c_j)}{\sum_{h=1\cdots c} \prod_{i=1\cdots m} p(a_{ik} \mid c_h)\, p(c_h)} \qquad (1)$$

위 식 (1)의 오른쪽 항에 나타나는 $p(c_j)$는 사전 확률(prior probability)이라 하는데 속성값에 상관없는 클래스값의

확률을 나타낸다. 일반적으로 베이즈 분류기에서는 균등확률(uniform probability)을 사용한다.

베이즈 분류기에서는 두가지 기본적인 가정을 한다. 첫 번째는 클래스들이 상호 배타적이고 포괄적(mutually exclusive and exhaustive)이어야 하는 것이고 두 번째는 클래스가 주어진 조건에서 속성들이 서로 독립이어야 하는 것이다. 본 연구의 웹 이미지 데이터가 최종적으로 속하게 되는 두가지 클래스는 위의 첫 번째 가정을 만족한다. 두 번째 가정인 데이터 속성 즉, 웹 이미지 특징들의 조건부 독립성을 만족시키기 위해서는 특징 추출에 몇가지 제약이 필요하다. 자세한 것은 '제4장 웹 이미지 특징 추출'에서 설명한다.

3.2 결정 트리

결정 트리는 각각의 데이터 케이스를 트리의 루트로부터 클래스를 나타내는 리프 노드에 이르도록 정렬하는 방법이다. 결정 트리의 각 노드는 특징의 테스트를 의미하고 브랜치는 그 특징의 가능한 값에 해당한다. 하나의 케이스는 루트 노드로부터 시작하여 특징의 테스트를 거치면서 그 값에 대응하는 브랜치를 통해 하위 노드로 이동하며 최종적으로 리프 노드에서 클래스의 결정이 이루어진다.

결정 트리에서 각 인스턴스는 고정된 특징 집합과 그들의 값으로 구성된다. 따라서 분류를 위한 가장 간단한 경우는 특징 값들이 적은 수의 서로 겹치지 않는 이산 값을 가질 때이다. 그러나 많은 경우 연속 값을 가지는 특징들이 존재하는데 이를 결정 트리에 이용하기 위해서 문턱값(threshold) 기반의 불(boolean) 특징을 생성한다. 연속 값을 가지는 특징으로부터 자동 생성된 불 특징은 이산 값을 가지는 다른 특징들과 경쟁하면서 결정 트리를 키워나간다[10]. 본 연구의 웹 이미지의 특징은 연속값을 가지는 것이 많이 포함되어 있지만 이와 같은 방법으로 결정 트리에 적용될 수 있다. 또한 웹 이미지의 특징들은 대부분 균질(homogeneous)의 것이 아니므로 결정 트리에 적합하다.

4. 웹 이미지 특징 추출

웹 이미지 분류에 관한 기존의 연구에서 많은 이미지 특징들이 분석되고 이용되었다. 이미지 트랜스코딩을 위한 분류에서는 주로 이미지 파일 헤더로부터 '이미지 포맷', '파일 크기', '컬러 수', '그레이 레벨' 등과 같은 특징들을 이용하였다[3, 7]. 기계학습을 이용하여 웹 이미지 분류를 수행한 연구[9]에서는 이미지 자체의 특징과 함께 HTML에 포함된 해당 이미지의 정보로서 '웹 페이지에서의 이미지 배치 순서', '이미지 라벨' 등이 이용되기도 하였다.

본 연구에서는 기존의 연구에서 채택된 이미지 파일 헤더로부터의 특징들중 분류하고자하는 클래스에 영향을 줄 것으로 예상되는 것은 포함시키되 그렇지 않은 것은 과감히 제외하였다. 대신 기존 연구에서는 채택된 바 없는 HTML 및 랜더링 화면에 나타나는 의미있는 특징들을 포함시켰다. 특징의 개수가 늘면서 발생할 수 있는 'Curse of Dimensionality'를 피하기 위해 가능한 모든 특징들을 포함시키는 것보다는 클래스 결정에 영향력이 클 것으로 추정되는 특징들만을 선택하였다. <표 1>은 추출된 웹 이미지 특징들에 대한 간략한 정보와 함께 이들의 종류를 구분하여 나타낸 것이다.

〈표 1〉 웹 이미지 특징 표

특징 ID	설 명	값 범위	특징의 출처
1	이미지 파일 크기	0.0 - 최대 크기	이미지 파일에서 획득 (I)
2	이미지 포맷	{gif, jpg}	
3	높 이	0.0 - 최대 높이	
4	넓 이	0.0 - 최대 넓이	
5	넓이/높이 비율	0.0 - 최대 비율	
6	HTML에서의 순서	{1, 2, …, 최대 개수}	HTML에서 획득 (H)
7	HTML에서의 상대 순서	0.0 - 1.0	
8	테이블 깊이	{0, 1, …, 최대 깊이}	
9	링크 여부	{t, f}	
10	반복 사용 여부	{t, f}	
11	광고 라벨 사용 여부	{t, f}	
12	아이콘 여부	{t, f}	랜더링된 화면에서 획득 (R)
13	웹 페이지상의 위치	{top, main, bottom, left, center, right, …}	
14	링크된 관련 텍스트 유무	{t, f}	
15	같은 행에 3개 이상 유사 이미지 배치 여부	{t, f}	
16	같은 열에 3개 이상 유사 이미지 배치 여부	{t, f}	

본 연구에서 도출된 16개의 특징들은 3가지 종류로 구분된다. 1번부터 5번까지의 특징들은 이미지 파일 헤더로부터 얻을 수 있는 이미지 고유의 정보들이다. 이를 특징 그룹 I로 표시한다. 6번부터 11번까지의 특징들은 HTML 파일 내에서 이미지가 사용된 방법으로부터 얻어지는 특징들이다. 이들은 HTML 문서 자체를 파싱하여 획득할 수 있다. 특징 그룹 H로 표시한다. 마지막으로 12번부터 16번까지의 특징들은 HTML 페이지가 화면상에 랜더링된 후 나타나는 특징들이며 특징 그룹 R로 표시한다. 이들 특징들 중 관련 연구[9]에서도 사용된것은 1, 2, 3, 4, 7, 11번 특징들이다. 1, 2, 3, 4번 특징들이 특징 그룹 I에 속하므로 특징 그룹 I는 대부분 기존 연구에서도 사용한 특징들이라고 할 수 있고 나머지 두 가지 그룹은 대부분 기존 연구에서 사용된 적이 없는 특징들로 구성되었음을 알 수 있다. 따라서 I 그룹 뿐만 아니라 H 또는 R 그룹에 속하는 특징들이 '제거 가능' 및 '제거 불가능' 클래스로의 분류에 얼마나 영향을 주는지 알아봄으로써 도출된 특징들의 유효성을 평가할 수 있다. 이와 같이 웹 이미지 분류를 위한 독특하고 풍부한 특징들을 발굴하고, 이를 종류별로 그루핑하여 적용하는 분류 실험을 하며, 그 결과를 통하여 도출된 특징들의 유효성을 검증하는 것이 본 연구의 주 목적이다. 특징 추출을 위한 이와 같은 접근 방식은 기존의 연구와 차별되는 것으로서 보다 높은 분류 성공률을 위한 향후 연구의 기초가 된다.

본 연구에서 도출된 16개의 특징들 중 <표 1>의 설명만

으로는 이해가 어려운 몇가지 특징들을 자세히 설명하면 다음과 같다.

- 8번 특징 '테이블 깊이'는 해당 웹 이미지가 HTML에서 배치될 때 속하는 테이블의 깊이를 뜻한다. 이것은 대부분의 웹 이미지가 레이아웃 배치를 위해 테이블의 한 셀에 포함된다는 것에 착안하여 테이블 깊이의 정도가 클래스 분류에 영향을 주는지 확인하고자 하였다. 대부분의 웹 이미지는 테이블의 한 셀에 포함되어 있으므로 실제로 수집된 데이터에서도 '테이블 깊이'가 0인 경우는 거의 찾아볼 수 없었다.
- 10번 특징 '반복사용 여부'는 동일 이미지가 페이지내에서 여러 번 사용된 것인지 아닌지를 나타낸다. 장식의 목적으로 사용되는 아이콘 등은 반복 사용되는 경우가 많고 '제거 가능' 클래스로 분류될 가능성이 높다.
- 11번 특징 '광고 라벨 사용 여부'는 이미지 이름에 광고나 배너를 의미하는 텍스트(예를 들면, ad, ban 등)가 포함되어 있는지 아닌지를 나타낸다. 광고의 종류에 따라서 '제거 가능' 클래스에 속할 수도 있다.
- 13번 특징 '웹 페이지상의 위치'는 랜더링된 웹 화면을 가로, 세로로 각각 3레벨, 모두 9개 영역으로 나누어 볼 때 어느 영역에 속하는지를 나타낸다. 이때 가로, 세로를 같은 길이로 3등분 하는 것이 아니고 화면의 레이아웃 배치 상에서 뚜렷한 구분 영역을 찾아낸다. 예를 들면 아래 그림에서 웹 화면은 세로 방향으로 top-main-bottom의 세 영역으로 나누어지고 main영역이 다시 left-center-right의 3영역으로 나누어진다. 따라서 아래 그림에서 각 이미지들이 속할 수 있는 영역은 모두 5개(top, left-main, center-main, right-main, bottom)가 된다. 나누어지는 영역의 개수는 웹 페이지마다 다르며 최대 9가 된다.

(그림 1) 웹 이미지 영역 구분

- 14번 특징 '링크된 관련 텍스트 유무'는 9번 특징 '링크 여부'가 이미지 자체의 링크를 뜻하는 데 반해, 웹 이미지가 화면에 배치되었을 때 같은 내용을 나타내는 링크된 텍스트가 인접하고 있는지를 나타낸다. 아래 그림은 링크된 관련 텍스트가 해당 이미지에 인접해 있는 경우를 보여준다. 이와 같은 경우 해당 이미지는 '제거 가능' 클래스에 속할 수 있다.

(그림 2) 웹 이미지와 링크된 관련 텍스트 예

- 15번 특징 '같은 행에 3개 이상 유사 이미지 배치 여부'는 위 (그림 1)의 center-main 영역에 속하는 상품 이미지들이 이에 해당한다. 6개가 1행씩 총 12개의 이미지가 2행에 배치되어 있고 그것이 다시 2번 반복되어 있다. 16번 특징 '같은 열에 3개 이상 유사 이미지 배치 여부'는 (그림 1)에서 left-main 영역의 이미지들에 해당된다.

이상과 같은 웹 이미지 특징들은 베이즈 분류기에서는 분포에 따른 확률값으로 클래스 결정을 위한 기준치를 생성할 것이고 결정 트리에서는 그 영향력의 크기에 따라 서로 경쟁하면서 클래스 결정에 이를 것이다. 단, 베이즈 분류기에서는 특징들이 주어진 클래스 조건하에서 서로 독립이어야 하는 가정이 있으므로 이 때 사용되는 특징들은 전체 16개 특징들로부터 조정이 필요하다. 문제가 될만한 것은 3, 4, 5번 특징과 6, 7번 특징이다. 3번 특징은 이미지의 높이를, 4번 특징은 이미지의 넓이를 나타내고 5번 특징은 이들의 비율이므로 3, 4, 5번 특징이 서로 독립이라고 할 수가 없다. 마찬가지로 6번 특징은 HTML에서 나타나는 순서를 의미하고 7번 특징은 이들의 상대적인 순서를 나타내므로 서로 독립성이 없다. 이들간의 독립성을 확실히 보장하기 위하여 본 연구에서 사용하는 베이즈 분류기에서는 5번과 6번 특징을 제외한 총 14개의 특징만을 사용한다.

5. 실험 및 결과

5.1 데이터 수집 및 실험 방법

웹 이미지의 분류를 위하여 현재 실제로 서비스 중인 웹 사이트로부터 HTML과 이에 포함된 이미지들을 수집하였다. '야후 쇼핑', '엠파스 쇼핑', 'HP 쇼핑' 등 인터넷 쇼핑 사이트를 주 대상으로 하였으며 총 30개의 웹 페이지로부터 2445개의 이미지 데이터를 수집하였다. 웹 이미지와 HTML 분석을 위해서는 웹 브라우저 플러그인 프로그램인 'HTML Analyser[1]'를 이용하였고 베이즈 분류기와 결정 트리 프로그램으로는 The Open University의 'RoC[4]' 및 RuleQuest의 'See5[6]'를 각각 이용하였다.

보다 공정한 실험을 위하여 전체 2445개의 이미지 집합을 3등분하여 815개의 부분 집합으로 만든 후, 분류기의 훈련(training) 단계에서는 2개의 이미지 부분 집합을 사용하고 분류(classification) 단계에서는 나머지 하나의 부분 집합을 사용하였다. 부분 집합이 3가지이므로 훈련과 분류 단

계에서 사용되는 부분 집합의 조합을 바꾸어가며 모두 3번의 실험을 통해 각각의 성능 척도값을 얻은 후 그 평균 값을 계산하여 최종 결과로 사용하였다.

5.2 분류 기법 비교 실험

분류기의 성능 척도는 다음과 같이 계산된다. 실험의 결과로 배치된 클래스는 원래의 사실과 비교되어 <표 2>와 같은 비교값이 도출된다. 비교표에서 행은 사실 클래스를 나타내고 열은 실험의 결과 배치된 클래스를 나타낸다.

〈표 2〉 사실 클래스와 배치된 클래스의 조합

사실 클래스	배치된 클래스	
	제거 가능 이미지(yes)	제거 불가능 이미지(no)
제거 가능 이미지(yes)	N_{yy}	N_{yn}
제거 불가능 이미지(no)	N_{ny}	N_{nn}

비교표에 나타난 값들을 이용하여 3가지 성능 척도, Recall Rate(R), Precision Rate(P), 그리고 F-measure(F)가 각각 다음과 같이 계산된다.

$$R = \frac{N_{yy}}{N_{yy} + N_{yn}} \quad P = \frac{N_{yy}}{N_{yy} + N_{ny}} \quad F = \frac{R + P}{2}$$

실험은 베이즈 분류기 및 결정 트리를 이용하여 각각 14개, 16개의 특징을 적용하였으며 아래 <표 3>과 같은 결과를 얻었다. <표 3>의 R, P, F 값들은 훈련 및 분류에 사용되는 이미지 데이터를 앞 절에서 설명한 3가지 조합으로 적용하여 얻은 후 평균한 값들이다. 베이즈 분류기가 결정 트리를 이용하였을 때보다 더욱 높은 성공률을 보여준다.

〈표 3〉 웹 이미지 분류 실험 결과

	베이즈 분류기 (14개 특징 적용)	결정 트리 (16개 특징 적용)
R(%)	78.37	75.01
P(%)	95.81	90.42
F(%)	87.09	82.72

5.3 특징 그룹 비교 실험

이어지는 실험은 더욱 효과적인 것으로 판명된 베이즈 분류기를 사용하여 웹 이미지 특징들을 종류에 따라 그루핑하여 적용해 본다. 특징들의 종류는 제 4장의 <표 1>에 나타나 있으며 '이미지 파일로부터 획득된 것(I)', 'HTML로부터 획득된 것(H)', '랜더링된 화면으로부터 획득된 것(R)'의 3가지이다. 본 실험에서는 각각의 특징 그룹 및 이의 조합으로 웹 이미지 분류에 적용해 보고 각각의 성능 척도를 서로 비교해 본다. 특징 그룹의 조합은 I, H, R 단독 그룹 3가지, IH, IR, HR 두 그룹 조합 3가지 그리고 전체 IHR 3그룹 조합까지 포함하여 모두 7가지이다. 이 때 첫 번째 실험에서와 같이 3가지의 이미지 부분 집합의 조합으로 훈련 및 분류에 반복 적용하여 평균을 내는 방법은 쓰지 않는다. 대신 모든 특징을 사용한 첫 번째 실험에서 가장 높은 성능을 나타낸 부분 집합의 조합을 사용한다. 이 실험에서는 특징의 종류에 따른 분류기의 성능을 알아보고자 하는 것이므로 특징 그룹간의 상대적인 비교에 더 의미가 있기 때문이다.

결과는 <표 4>에 나타난 바와 같다. F값을 비교해 보았을 때 I, H, R 각각의 단독 그룹에서는 I 그룹이 82.79%로 가장 높은 결과치를 나타내었다. 이것은 특징 그룹을 그 출처에 따른 종류로 볼때 이미지 파일로부터 획득된 정보들이 클래스 분류에 가장 큰 영향력을 미친다는 의미이다. H 및 R 그룹은 두 그룹으로 묶여서 적용되었을 때 더 큰 효과를 가져온다. IH, IR, HR의 두 그룹 조합에서는 각각 77.22%, 93.99%, 93.44%의 결과치를 나타내었다. H 및 R 그룹이 단독으로는 I 그룹 만큼 큰 효과를 가지지 못하지만 조합된 그룹 IR에서나 HR에서는 모두 93%가 넘는 좋은 성능을 보여 주었다. 따라서 본 연구에서 비교하고자 하였던 특징의 종류별 효과를 입증할 수 있게 되었다. 기존 연구에서 사용하지 않았던 독특하고 풍부한 특징들을 HTML 파일 및 랜더링 화면으로부터 발굴해 내어 웹 이미지 분류에 추가로 적용함으로써 이미지 파일에서 획득한 특징들만 사용한 경우보다 훨씬 높은 성능을 이끌어 낼 수 있었다. 마지막으로 IHR 조합으로 전체 특징을 모두 적용했을 때 91.61%의 F값을 얻었으며 이는 최고 값인 IR 조합의 93.99%보다는 약간 낮은 결과치이다.

〈표 4〉 특징 그룹별 분류 실험 결과

	I	H	R	IH	IR	HR	IHR
R(%)	69.57	67.39	62.96	68.84	88.81	90.32	85.04
P(%)	96.00	83.78	85.00	85.59	99.17	96.55	98.18
F(%)	82.79	75.59	73.98	77.22	93.99	93.44	91.61

6. 결론 및 향후 연구

웹 이미지가 전체 웹 문서에서 매우 중요한 역할을 하고 있고 그 사용 빈도는 날로 증가하고 있음에도 불구하고 웹 문서에 포함된 이미지의 분석에 대한 연구는 활발하게 이루어지지 못하고 있다. 웹 이미지 분석은 그 목적에 따라 이미지의 특성별로 통계를 내거나 각종 휴리스틱을 이용하여 분류하는 방식으로 연구되어 왔다. 또 기계학습을 이용한 웹 이미지의 분석 또는 분류에 대한 연구도 몇몇 찾아볼 수는 있으나 분류에 이용하는 이미지 특징들이 빈약하며 실험 방법이 구체적이지 않아 여러 가지 한계를 가지는 것이 사실이다.

본 연구에서는 이와 같은 기존 연구의 제한성을 극복하기 위해 기계학습 기반의 웹 이미지 분류를 위한 보다 구체적이고 다양한 실험을 계획하고 실행하였다. 웹 이미지 분류를 위해 대표적인 두 가지 기계학습 기법인 베이지안 기법과 결정 트리 기법을 사용하여 비교 실험하였으며 독특하고 풍부한 웹 이미지 특징들을 다양한 출처로부터 발굴하

여 적용 실험하였다. 실험 결과 기존의 기계학습 기반의 웹 이미지 분류에서는 사용된 적이 없는 특징들이 분류기의 성능 향상에 높은 효과를 내는 것으로 드러났다. 이는 본 연구에서 도출한 웹 이미지 특징들이 기존 연구의 특징들과 함께 사용될 때 웹 이미지의 분류에 더욱 큰 영향을 미칠 수 있으며 따라서 본 연구에서 매우 유의미한 특징들을 발굴한 것임이 입증된 것이다. 한편 본 연구에서 사용한 베이지안 기법과 결정 트리 기법에서의 평균 F값은 각각 87.09% 및 82.72%로서 참조한 이미지 분류 연구들[9, 12]과 비교할 때 전자는 높은 결과치를, 후자는 약간 낮은 결과치를 보여줌으로써 전반적으로 큰 차이가 없음을 알 수 있다. 이것은 전체 특징 그룹을 모두 적용했을 때이므로 IR이나 HR 조합 그룹을 적용했을 때는 더 높은 결과치를 얻을 수 있을 것으로 기대된다.

이상과 같은 다양한 실험 및 그 결과에도 불구하고 본 연구는 적용 웹 사이트와 데이터 크기면에서 제한적인 것이 사실이다. 이어지는 연구에서는 더욱 유용한 웹 이미지 특징들을 도출하고, 대상 웹 사이트 및 문서를 더욱 확대하여 웹 이미지 분석에서 일반화 될 수 있는 분류 방법으로 발전시켜야 할 것이다. 또한 자동 분류 시스템 개발에 대한 연구도 병행되어야 할 것이다.

참 고 문 헌

[1] ADEW, "HTML Analyser," http://www.htmlanalyser.com/.

[2] G. Penn, J. Hu, H. Luo and R. McDonald, "Flexible web document analysis for delivery to narrow-bandwidth devices," In Proc. 6th International Conference on Document Analysis and Recognition, Seattle, WA, USA, pp.1074-1078, September, 2001.

[3] J. R. Smith, R. Mohan and C-S Li, "Content-Based Transcoding of Images In The Internet," In Proc. IEEE Inter. Conf. Image Processing, October, 1998.

[4] Knowledge Media Institute and The Open University, "RoC : The Robust Bayesian Classifier," http://kmi.open.ac.uk/projects/bkd/.

[5] M. J. Swain, C. Frankel and V. Athitsos, "WebSeer : An Image Search Engine for the World Wide Web," In Proc. IEEE Computer Vision and Pattern Recognition Conference, June, 1997.

[6] Rulequest Research, "Data Mining Tools See5 and C5.0," http://www.rulequest.com/see5-info.html.

[7] S. Chandra, A. Gehani, C. S. Ellis and A. Vahdat, "Transcoding Characteristics of Web Images," In Proc. Multimedia Computing and Networking, San Jose, CA, Vol.4312, pp. 135-149, January, 2001.

[8] S. Chandra and C. S. Ellis, "JPEG Compression Metric as a Quality Aware Image Transcoding," In Proc. USENIX 2nd Symposium on Internet Technologies and Systems, Boulder, CO, pp.81-92, October, 1999.

[9] S. Paek, "Detecting image purpose in World-Wide Web documents," In Proc. IS&T/SPIE Symposium on Electronic Imaging : Science and Technology Document Recognition, San Jose, CA, USA, January, 1998.

[10] T. M. Mitchell, 'Machine Learning', McGraw-Hill, 1997.

[11] Y. Wang and J. Hu, "A Machine Learning Based Approach for Table Detection on The Web," In Proc. The 11th International World Wide Web Conference, Honolulu, Hawaii, USA, pp.242-250, May, 2002.

[12] 김명관, "2단계 분류기법을 이용한 영상분류기 개발", 한국컴퓨터산업교육학회논문집, Vol.3., No.5, pp.605-610, 2002.

조 수 선

e-mail : scho@etri.re.kr

1987년 서울대학교 계산통계학과 졸업(학사)

1989년 서울대학교 대학원 계산통계학과 졸업(석사)

1989년~1994년 (주)웅진미디어 CBE개발부 연구원

1994년~현재 한국전자통신연구원 정보가전연구부 선임연구원

관심분야 : 모바일 웹지원 기술, 웹 컨텐트 분석 및 이해

이 동 우

e-mail : hermes@etri.re.kr

1995년 경북대학교 전자공학과 졸업(학사)

1997년 경북대학교 대학원 전자공학과 졸업(석사)

1997년~2001년 (주)현대전자 전장사업본부 대리

2001년~현재 한국전자통신연구원 정보가전연구부 연구원

관심분야 : 지능정보 단말, 멀티모달 브라우저

한 동 원

e-mail : dwhan@etri.re.kr

1982년 숭실대학교 전자공학과 졸업(학사)

1992년 한남대학교 대학원 전자공학과 졸업(석사)

1995년 충남대학교 컴퓨터과학과 박사과정 수료

1982년~현재 한국전자통신연구원 책임연구원 휴대클라이언트 연구팀장

관심분야 : 멀티미디어 휴대정보단말, 웨어러블 컴퓨터, 편재형 컴퓨팅 분야

황 치 정

e-mail : cjhwang@ipl.cnu.ac.kr

1975년 서강대학교 수학과 졸업(학사)

1979년 서강대학교 대학원 수학과 졸업(석사)

1985년 Univ. of Connecticut 전산학과 졸업(석사)

1987년 Univ. of Connecticut 전산학과 졸업(박사)

1987년~1988년 Univ. of Connecticut 객원교수

1988년~현재 충남대학교 컴퓨터과학과 교수

관심분야 : 이미지 프로세싱, 패턴 인식 등