형태소 분석과 Skin-Color 분포의 Human Detection 알고리즘을 이용한 유해사이트 자동 분류 시스템의 구현

이승만[O] 장영헌 임정환
고려대학교 컴퓨터과학기술 대학원
{scsmlee[O], andsimpson , aphros99}@korea.ac.kr

Implementation of a Harmful Website's Automatic Classification System based on Morphological Analysis and Skin-Color Distribution's Human Detection Algorithm

Seung-Man Lee[O] Young-Hun Jang Jung-Hwan Lim
Graduate School of Computer Science and Technology, Korea University

요 약

인터넷은 유익하고 건전한 정보의 유통이 대부분이지만 최근에는 익명성과 상업성으로 인해 유해 정보가 급속하게 늘어나고 있는 추세이다. 이러한 부정적인 영향으로부터 청소년들과 어린이들을 보호하기 위하여, 본 논문은 유해사이트 분류를 자동으로 할 수 있는 시스템을 제안한다. 기존의 유해사이트 구축은 검색 요원들이 유해사이트를 돌아다니며 일일이 데이터를 수집하여 분류하거나 유해사이트의 내용 중에 텍스트만을 추출하여 패턴 매칭 방법으로 분류하는 것이 대부분이었지만, 본 논문은 기존 방법의 문제점을 해결하기 위하여 형태소 분석을 이용한 사이트의 유해도 측정과 Skin-Color 분포의 분석 결과를 병합하여 95% 이상의 정확도(Precision) 성능을 보이며, 신뢰도가 높은 유해사이트 자동 분류 시스템을 구현할 수 있다는 것을 증명하였다.

1. 서 론

인터넷의 급속한 성장은 정보통신 환경에 큰 변화를 일으키고 있으며 인간의 삶의 방식과 가치관에 커다란 영향을 주어, 많은 사람들에게 새로운 기회와 도전의 장소를 제공하고 있다. 하지만 이러한 긍정적 측면의 이면으로는 인터넷의 익명성, 상업성으로 인해 유해 정보가 급속하게 증가되고 있으며, 이로 인하여 청소년이나 어린이들에게도 인터넷이 일반화되어 유해한 정보가 아무런 선별 없이 제공됨으로써 사회적으로 큰 문제를 야기시키고 있다[1].

청소년들이 이러한 유해 정보에 접하는 것을 막을 수 있는 방법으로는 개인 PC에 유해사이트의 IP와 URL을 차단해 주는 소프트웨어를 설치하여 사용하는 방법이 있고, 인터넷 서비스 제공자측에서 IP와 URL을 차단하는 방법이 있다. 이 두 가지 방법 모두 차단해야 할 IP와 URL의 구축을 필요로 한다. 기존의 유해사이트 구축은 검색 요원들이 유해사이트를 돌아다니며 일일이 분류를 하거나 웹 로봇 에이전트가 유해사이트의 내용 중 텍스트만을 추출해 패턴 매칭 (Pattern Matching) 방법으로 분류하는 것으로, 전자는 속도면에서 느린 대신 정확성은 높고, 후자는 속도면에서 빠르지만 정확성은 그다지 높지 않다. 그러므로 본 논문은 기존 방법의 문제점을 해결하기 위하여 형태소 분석을 이용한 유해사이트의 분류와 Skin-Color 분포의 Human Detection 알고리즘을 이용하여 신뢰도가 높은 유해사이트 자동 분류 시스템을 구현하고자 한다.

2. 기반 연구

2.1 형태소 분석을 이용한 문서 분류 알고리즘

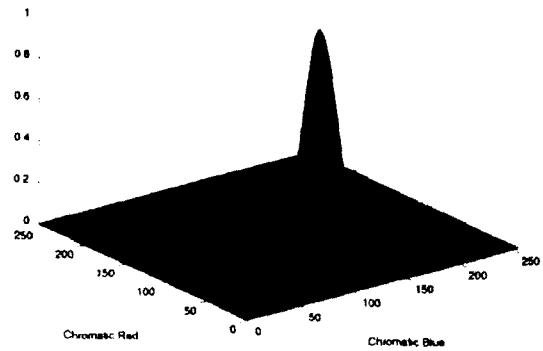
형태소 분석은 입력 문장에서 최소 의미 단위인 형태소를 추출하는 것으로, 전자 사전과 형태소 분석 문법을 이용하여 자연어 분석의 최소 단위를 제공하는 단계이다. 입력된 문장의 의미를 파악하기 위해서는 문장 내 사용된 어절의 품사 정보 및 의미 정보를 알아낸 후, 어절을 구성하는 형태소의 결합열을 찾아 각 형태소의 의미를 조합하여야 한다. 형태소 분석기가 어절로 구성되어 있는 문장을 형태소로 분리하여 다양한 형태소의 기본형 및 품사 정보를 제시하면, 태거(Tagger)는 확률 통계적 기법과 같은 다양한 정보를 활용하여 가장 가능성 높은 하나의 결과를 나타낸다[2].

형태소 분석의 결과가 제시되면 문서 분류를 실행하게 되는데, 문서 분류를 위한 기계학습 이론 중 대표적 알고리즘은 Naïve Bayesian, k-NN(k-Nearest Neighbor) 및 TFIDF 방법을 들 수 있다. Naïve Bayesian 방법은 이진속성벡터(Vector of Binary Attributes)로 문서의 통계적 확률을 이용하여 클래스를 결정하는 기법이고, k-NN 방법은 메모리 추론에 기반을 둔 학습 기법으로서 관련 문서들 간의 근접도를 이용하여 문서 분류가 이루어지는 기법이다. 마지막으로 TFIDF 방법은 문서에 출현하는 단어의 빈도와 특정 단어를 포함하는 문서의 역수를 특성으로 사용하여, 벡터 형태로 문서를 분류하는 알고리즘이다[3].

2.2 Skin-Color Detection 알고리즘

Skin-Color Detection 알고리즘은 인간만이 갖는 고유한 컬러 특성을 이용한 것으로, 대부분이 얼굴 인식 이나 얼굴 추적에서 사용되고 있으며, 이미 여러 분야에서 효과적인 방법임이 증명되었다. Skin-Color Detection 알고리즘에

대해서는 많은 연구가 진행되어 왔는데 RGB, HSV, YIQ, YcbCR, CIELuv 등이 사용되어 지고 있으며, 그 중 *Crowley*가 제안한 정규화 RGB 색상 공간에서 R(red)과 G(green)값을 이용하여 임계치를 기준으로 피부 영역을 추출하는 방법과 Saxe가 제안한 HSV 색상 공간에서 히스토그램의 Intersection을 반복적으로 사용하여 피부를 인식하는 방법이 많이 사용되어 진다[4].

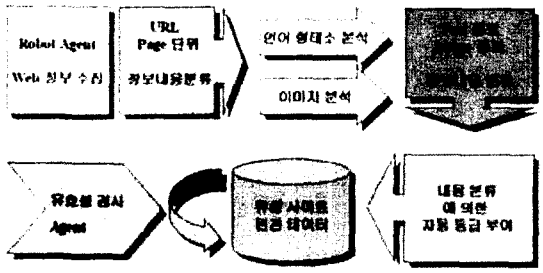


[그림 1] 살색 Gaussian 분포도

3. 제안 모델

3.1 형태소 분석을 이용한 TFIDF 유해 텍스트 등급 분류기

본 논문에서 제안한 분류기는 형태소 분석을 실시한 후, 미리 정의한 범주와 등급의 품사별 유해 키워드를 추출하여 TFIDF(Term Frequency Inverse Document Frequency) 알고리즘으로 해당 문서의 분류와 유해 정도를 자동으로 판단하는 것이다. 이로 인하여 수작업으로 문서를 분류하는데 소요되는 시간과 노력, 비용 등을 감소시킴으로써 효율적인 유해 등급 사이트 구축을 가능하게 한다.



[그림 2] 유해사이트 자동 분류 시스템 흐름도

TFIDF 텍스트 분류 알고리즘은 문서에 출현하는 TF(Term Frequency)와 DF(Document Frequency)의 특성을 이용하여 문서를 분류하는 방법으로, 다른 분류 알고리즘에 비해 계산량이 적고 성능이 우수하다는 장점이 있다. 단어의 관련도는 TF와 DF로 표현되는데, TF는 문서에서 특정 단어의 출현 빈도를 나타내며, 이는 특정 단어가 문서의 내용을 얼마나 잘 표현하고 있는지를 보여준다. DF는 특정 단어가 출현한 문서의 수를 나타내는 것으로, 특정 단어가 가지는 변별성 정도를 나타낸다. DF가 큰 단어는 관련 문서와 비관련 문서를 구분하는데 유용하지 못하지만 DF가 작은 단어는 구분의 유용함을 제공한다. TF와 DF를 이용하여 문서를 분류하는 기법에는 여러 가지가 있지만 주로 TFIDF 방식을 사용하는데, TF와 DF의 역수인 IDF(Inverse Document Frequency)를 곱하는 형태로 각 단어의 중요도와 유사도를 계산한다. 어떤 문서 $\mathbf{doc}_i$ 에서 단어 w_i 의 빈도수를 나타내는 TF, DF 및 IDF의 수식은 다음과 같이 표현된다[5].

$$tf(w_j, \mathrm{doc}_i) = \text{count of } w_i \text{ ococuring in document } \mathrm{doc}_i$$

$$idf(w_i) = \log\left(\frac{n}{df(w_i)}\right)$$

$$TFIDF = tf(w_i, doc_i) \cdot idf(w_i)$$

3.2 Skin-Color 분포를 이용한 유해 이미지 등급 분류기

제안된 알고리즘은 사람의 피부가 컬러 성분 중 Red 성분이 많은 비율을 차지하고 있는 것을 이용하는 것으로, 색상 공간 변환을 거친 이미지의 RGB 성분 비율 중, 각각의 비율이 특정 임계치 사이에 존재하면 살색으로 간주하고 그렇지 않으면 살색이 아닌 것으로 판별한다. 이는 아래의 수식과 조건을 만족하여야 하며, 간단하고 빠른 판별 속도를 제공하는 장점이 있다[6].

(R, G, B) is classified as skin if :
R > 95 and G > 40 and B > 20 and
Max{R, G, B} - Min{R, G, B} > 15 and
|R - G| > 15 and R > G and R > B

[그림 3] 유해 이미지의 Skin Color 분포 분석 결과

본 논문의 실험을 위해 일반 이미지와 유해 이미지를 수집하여 살색이 차지하는 비율을 조사해 본 결과, 유해 이미지는 일반 이미지와는 달리 전체 픽셀에서 살색이 차지하는 비율이 높은 결과를 보였고, 살색 비율이 전체에서 30~65% 정도의 분포를 보였을 경우에 대부분의 이미지가 유해 이미지로 판명되었다. 하지만 사람의 특정 부위를 접사하여 촬영한 경우이거나 살색 영역이 조명의 영향을 받은 경우에는 오분석율이 높은 단점을 보였다.

4. 실험 및 결과

4.1 실험 환경 및 입력 데이터

본 논문의 구현을 위해서 Dual CPU 1.2 GHz, Main Memory

1 Giga, Red Hat Linux release 9 (kernel 2.4.20-8) 운영체제에서, 형태소 분석기와 Skin Color 분포 추출기는 gcc 3.2.2-5를 이용하여 구현하였고 로봇 에이전트, 유효성 검사 에이전트 및 Wrapping 모듈은 Java (J2SE 1.4.2_03)로 구현하였으며, 하부 저장 구조로는 IBM DB2 Universal Database 8.1을 사용하였다. 웹 환경으로는 Apache 웹 서버 1.3.29, Servlet과 JSP 환경을 사용하기 위해 ApacheJServ 1.1.2 와 GnuJsp 1.0.1을 사용하여 구현하였다.

웹 로봇 에이전트의 출발 정보 URL은 Altavista (http://www.altavista.com) 검색 엔진에서 Sex 라는 키워드로 검색하여 나온 URL 리스트 1,000개를 사용하였고, 유해 키워드 전자 사전은 정보통신윤리위원회의 Safenet 등급 기준이 적용된 전자 사전을 이용하였다[7].

4.2 실험 평가 방법

구현 평가 방법으로는 얼마나 정확하게 분류되었는지 성능을 측정하기 위하여 형태소 분석 없이 패턴 매칭만 사용한 결과, 형태소 분석만 사용한 결과, Skin Color 분포만을 이용한 결과, 그리고 형태소 분석과 Skin Color 분포를 함께 사용한 결과를 아래와 같은 수식으로, 정확도 (Precision)와 재현율 (Recall) 그리고 F-measure 측정식을 이용하여 평가하였다[8].

$$\text{Precision} = \frac{\text{Categories assigned by the system and correct}}{\text{Total Categories assigned}}$$

$$\text{Recall} = \frac{\text{Categories assigned by the system and correct}}{\text{Total Categories correct}}$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

4.3 실험 결과

실험 결과를 보면 Skin Color 분포만을 이용하여 분석한 결과가 71%의 정확도로 가장 나쁜 성능을 보였고, 패턴 매칭 분석 84%, 형태소 분석을 이용한 분석이 91%를 보였으며 형태소 분석과 Skin Color 분석을 병합한 결과가 95%의 정확도를 보이며 가장 좋은 성능을 나타내었다. 전체적인 성능 비교를 위하여 F-measure 값을 측정하였으며, 이 또한 형태소 분석과 Skin Color 분석을 병합한 결과가 가장 우수함을 보였다.

[표 1] 성능 측정 결과표

	패턴 매칭 결과	형태소 분석 결과	Skin Color 분석결과	형태소+Skin Color 분석결과
RI	537	639	238	803
RL	456	587	171	770
NRD	41	15	185	5
Recall	0.520	0.670	0.195	0.878
Precision	0.849	0.918	0.718	0.958
F-measure	0.644	0.774	0.306	0.916

TI: 전체 data (1000 개의 사이트)

TL: 전체 data에서 실제로 유해하다고 판단된 data (876 개의 사이트)

RI: 전체 data에서 구현한 시스템이 유해하다고 분류한 data

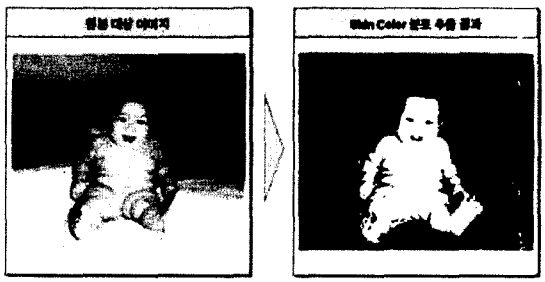
RL: 구현한 시스템이 유해하다고 분류한 data 중 실제로 유해한 data

NRD: 구현한 시스템이 유해하지 않다고 분류한 data 중 실제로 유해하지 않은 data

5. 결론 및 향후 연구

기존의 유해사이트 구축은 검색 요원들이 유해사이트를 돌아다니며 일일이 data를 수집 하여 분류하거나 유해사이트의 내용 중에 텍스트만을 추출하여 패턴 매칭 방법으로 분류하는 것이 대부분이었다. 더욱이 최근 유해사이트의 특징은 텍스트 정보는 줄이고 이미지 정보를 늘리는 추세이므로 기존의 방법으로는 효과적인 유해사이트 원천 데이터를 구축하기가 어렵다.

따라서 본 논문은 기존 방법의 문제점을 해결하기 위하여 형태소 분석을 이용한 사이트의 유해도 측정과 Skin-Color 분포의 Human Detection 알고리즘을 이용하여 95% 이상의 정확도 성능을 보이며, 신뢰도가 높은 유해사이트 구축 시스템을 구현할 수 있다는 것을 증명하였다. 여기서 중요한 점은 단순히 Skin Color 분포를 이용하여 분석한 결과는 좋은 성능을 보이지 않지만 형태소 분석 결과와 병합하여 이용하였을 경우에는 아주 좋은 성능을 보여준다는 것을 알 수 있다.



[그림 4] 일반 이미지의 Skin Color 분포 오분석 결과

그러나 본 논문에서 구현한 Skin Color 분포를 이용한 분석은 오직 살색 이미지가 원본 대상 이미지에서 차지하는 비율만을 통해 분석하는 알고리즘이기에 살색과 유사한 자연 이미지나 [그림 4]의 유해하지 않은 아기 사진등에서는 오분석을 일으킬 수 있는 많은 문제점을 갖고 있다. 그러므로 향후에는 인간의 특정 부위를 인식할 수 있는 개선된 Human Detection 알고리즘을 이용한 연구가 필요하리라 생각한다.

6. 참고 문헌

[1] Tim Jordan, "CyberPower - The Culture and Politics of Cyberspace and the Internet", Routledge, pp.60-68, 1999.

[2] Jeongwon Cha, Geunbae Lee, Jong Hyeok Lee, "Generalized Unknown Morpheme Guessing for Hybrid POS Tagging of Korean" Proc. SIXTH WORKSHOP ON VERY LARGE CORPORA in Coling-ACL98, 1998

[3] Steve Lawrence, C.Lee Giles, Kurt Bollacker, "Digital Libraries and Autonomous Citation Indexing", IEEE Computer Vloume32 Number6, IEEE, pp.67-71, 1999

[4] Michael J.Jones, James M.Rehg. "Statistical Color Models with Application to Skin Detection", Cambridge Research Laboratory, IEEE 1999

[5] Joachims T. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization" Proc.14th International Conference Machine Learning. 1997

[6] P.Peer, F.Solina, "An Automatic Human Face Detection Method", Proc.4th Computer Vision Winter Workshop, pp.122-130, Rastenfeld, Austria, 1999

[7] 정보통신부. "Research in Information Communication Ethics: Proliferation of Internet Content Rating System". 정보통신윤리위원회, pp.43-52, 2000

[8] Bekkerman R., El-Yaniv R., Tkshby N., Winter Y., "On Feature Distributional Clustering for Text Categorization", Proc.SIGIR 2001, SIGIR Conference, pp.146-153, 2001