웹 로봇 에이전트의 하이퍼링크 분석기법을 이용한 음란메일 차단 시스템의 구현

이승만○ 정희석 한상 송우석 이도한 홍지영 반의환 양준영
정보통신윤리위원회 기술개발팀
{smlee○, hsjung, hs, wssong, dhlee, jyhong, baneh, jyyang}@kiscom.or.kr

Implementation of Anti-Porn Spam System based on Hyperlink Analysis Technique's of the Web Robot Agent

Seung-Man Lee○ Hui-Sok Jung Sang Han Woo-Seok Song
Do-Han Lee Ji-Young Hong Eui-Hwan Ban Joon-Young Yang
Technology Development Team, Korea Internet Safety Commision

요 약

이메일은 누구나 쉽게 정보를 교환할 수 있는 편리함 때문에 인터넷에서 가장 중요한 수단으로 사용되고 있다. 그러나 순수한 의사소통의 수단이 아닌 스팸메일의 범람은 성인뿐만 아니라, 어린이·청소년에게도 무차별적으로 전송됨으로써 심각한 부작용을 낳고 있다. 본 논문은 점차 지능화 되는 신 유형의 음란 스팸메일로부터 청소년을 보호하기 위하여 새로운 방법의 음란메일 차단시스템을 제안하고자 한다.

기존의 스팸메일 차단시스템은 사용자가 직접 음란한 메일이라고 판단되는 메일에 대해 일일이 키워드를 설정하거나, 메일 내용 중에 텍스트만을 추출하여 패턴 매칭방법으로 분류하는 것이 대부분이었지만, 본 논문은 기존 방법의 문제점을 해결하기 위하여 이미지 내 Skin-Color분포의 Human Detection 알고리즘과 웹 로봇 에이전트의 하이퍼링크 분석기법을 사용하였다. 성능 측정결과, 형태소 분석과 Human Detection 알고리즘을 병합하여 적용한 경우 성능 측정에서 90% 정도의 F-measure를 보였지만, 추가적으로 웹 로봇 에이전트의 하이퍼링크 분석기법을 병합하여 적용한 경우 97% 이상의 F-measure를 보이며, 신뢰성이 높은 음란스팸메일 차단 시스템을 구현할 수 있다는 것을 증명하였다.

1. 서 론

인터넷의 급속한 성장은 정보통신 환경에 큰 변화를 일으켜 왔으며, 인간의 삶의 방식과 가치관에 커다란 영향을 주어 많은 사람들에게 새로운 기회와 도전의 장을 제공하고 있다. 그 중에서도 이메일은 전 세계 누구나 쉽게 정보를 교환할 수 있는 편리성과 간편함 때문에 그동안 인터넷에서 가장 중요한 수단으로 사용되어 왔다.

그러나 이러한 저비용·고효율의 이메일을 이용한 마케팅이 발달하게 되면서 원하지 않는 스팸메일이 범람하게 되고, 그에 따른 부작용은 심각한 수준에 이르게 되었다. 지난 2003년 네티즌 1인당 하루 평균 50통이나 발송되던 스팸메일은 정부의 강력한 규제와 단속으로 그 수의 증가가 둔화되고는 있으나, 광고의 내용이나 전송기법은 오히려 더욱 악성적으로 지능화되어 실제로 수신자가 체감하는 피해의 정도와 경제적 손실은 이보다 더 심각해지고 있다.

특히 이 중, 성인, 어린이, 청소년 등에게 무차별적으로 전송되고 있는 음란스팸메일은 정서적으로 발달과정에 있는 어린이, 청소년의 건전한 도덕관념을 왜곡시킴은 물론, 모방범죄도 유발시키는 등 사회적으로 큰 문제를 야기시키고 있다.

이에 따라, 본 논문은 점차 지능화 되고 있는 신 유형의 음란스팸메일에 적극적으로 대처 할 수 있도록 새로운 방법의 음란메일 차단시스템을 제안하고자 한다. 이 시스템은 음란 단어, 음란이미지 및 음란사이트 주소(URL)를 인식할 수 있는 기능이 있으며, 특히, 웹 로봇 에이전트를 이용하여 음란메일에 링크된 페이지의 음란성까지 검사한다는 점에서 기존의 차단시스템들과 차별화된다 하겠다.

본 논문의 2장에서는 Web-Robot과 형태소분석 및 Skin-Color Detection 알고리즘에 관련된 기존 연구에 대해 기술하고 3장에서는 음란메일 차단의 정확도를 높일 수 있는 알고리즘과 시스템에 대해 제안하였다. 4장에서는 실험 환경과 결과에 대해 서술하고 5장에서는 결론 및 향후 연구과제를 제시하였다.

2. 기반 연구

2.1 Web-Robot

웹 로봇은 웹에서 사용자의 질의를 받아 시스템적으로 문서를 탐색하고 그것들의 관련성을 평가해 사용자에게 리스트를 넘기는 역할을 하는 것이다. 웹은 directed graph와 유사한 구조를 가지고 있기 때문에 graph-traversal 알고리즘을 사용하여 정보탐색을 한다.

웹의 traversal방법은 3가지로 분류된다. 첫 번째로 최초 탐색에 Seed URL을 주고 문서를 추출한 다음, 그 이후에 재귀적으로 탐색을 하는 방법이다. 재귀적인 탐색시에 breadth-first 또는 depth-first방법 중 하나를 사용한다. 다음으로 특정 URL의 집합을 통해 이를 탐색하는 기법이다. 이는 웹사이트의 인기도와 검색도가 높은 URL이 좀 더 많은 URL들을 가리키고 있을 것이라는 가정하에 정보를 탐색하는 기법이다. 마지막으로 URL의 구분에서 국가 또는 할당된 지역에 기초하여 웹 공간을 분배하고, 그 공간에 대해서 로봇을 분배하는 방법이 있다. 마지막 방법이 앞의 두 가지 방법보다는 좀 더

널리 이용되고 있다. [1]

문서탐색에 있어서 현존하는 Robot Agent들이 주로 사용하는 방법은 색인을 사용하는 것이다. IR(Information Retrieval) 문맥에서 인덱싱은 색인이나 content descriptor를 할당하여 문서의 representation을 개발하는 과정이다. 색인에는 objective 색인과 non-objective 색인 두 가지 형식이 있다. 전자는 의미론과 관련이 없는 저자명, 문서URL, 출판일 등이 해당되고, 후자는 문서내의 정보를 반영하는 content terms를 말한다. 이러한 색인을 추출하기 위한 자동인덱싱 방법이 IR시스템에 있어서 성능을 크게 좌우하는데, 이는 Single-Term기법과 Multi-Term기법으로 나뉘고, 각 기법에 대해 통계학적, 개연론적, 언어학적인 접근방법들이 존재한다. 위의 두 인덱싱 스키마 기법 중에서 Multi-Term기법이 문맥 상에서 2가지 이상의 의미를 가지는 단어의 중의성 때문에 좀 더 이상적인 기법이라 할 수 있다.

문서탐색의 모델은 문서와 질의를 위한 representation, 사용자의 질의에 대한 문서의 적절성에 평가를 위한 매칭전략과, 랭킹질의 결과에 대한 방법 그리고 user-relevance 피드백을 위한 메커니즘에 의해 결정된다. 이러한 모델은 집합론적(Set Theoretic)모델, 대수학적(Algebraic)모델, 개연론적(Probabilistic)모델과 하이브리드(Hybrid)모델로 분류된다. [2]

2.2 형태소 분석을 이용한 문서 분류 알고리즘

형태소 분석은 입력된 문자열을 분석하여 형태소로 분류하는 작업을 말한다. 이러한 형태소를 자동으로 분석하는 기법에는 사전기반형과 휴리스틱기반형으로 나뉘어진다. 전자는 사전에 저장된 단어들에 대한 정확한 결과를 보증하기 위해 어간사전을 사용하고, 후자는 이전에 보지 못한 단어들에 대한 결과를 유추하기 위하여 휴리스틱 규칙을 사용한다. 그러나 언어는 새로운 단어가 지속적으로 나타나기 때문에 모든 사전은 완전하다고 말할 수 없다.

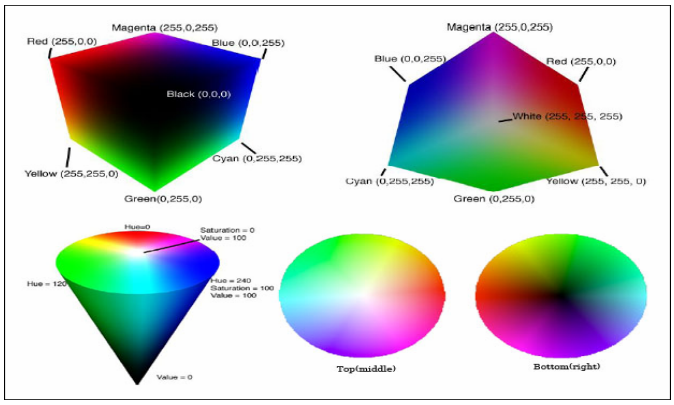
형태론적인 관점에서 언어는 굴절어와 교착어 등으로 구분할 수 있는데, 어떠한 분석기법을 사용하더라도 각 언어마다 문법 체계가 다르기 때문에 타 언어에서 사용되는 모델을 통한 형태소 분석에는 많은 문제점이 발생하게 된다. 그렇기 때문에 각 언어마다 분석기법이 다양하게 존재한다. 이러한 분석 모델로서 two-level 모델(Koskenniemi, 2002), Gelbukh (2003), Maximun Entropy(Ratnaparkhi, 1996), Pos-tagging (Church, 1988)모델 등이 사용되고 있다. [3]

위의 형태소 분석 모델을 이용하여 형태소가 분리되면 이를 바탕으로 문서를 분류하는데, 대표적인 방법으로는 naïve Bayesian기법과, k-NN(k-Nearst Neighbor)학습기법 및 TFDIF(Term Frequency Inverse Document Frequency)분류 기법 등을 들 수 있다. naïve Bayesian기법은 베이스 정리 (Bayes theorem)를 바탕으로 한 확률모델을 이용하고, k-NN학습기법의 경우 각 문서들간의 유클리드 거리를 계산하여 분류대상 문서와의 근접도를 바탕으로 문서를 분류를 하는 기법이다. 마지막으로 TFIDF 기법은 문서 내 특정단어의 출현 빈도수(TF)와 특정 단어가 검색된 N개의 문서의 수에 대한 역수(IDF)를 바탕으로 문서를 분류하는 알고리즘이다. [4]

2.3 Skin-Color Detection 알고리즘

Skin-Color Detection 알고리즘은 이미지로부터 사람의 피부색 픽셀을 찾는 알고리즘으로서, 이는 얼굴탐색 및 웹 이미지 컨텐츠의 검색 및 필터링을 하는 다양한 어플리케이션들에서 중요한 역할을 하고 있다. 현재 Skin-Color Detection에 대한 연구가 많이 진행되고 있는데, 이를 위해 다양한 통계학적인 색상모델이 사용되고 있다. [5]

이러한 색상모델로서 RGB, Normalized RGB, HSV, TSL, YCrCb, CIELuv등이 사용되고 있는데, 그 중에서, RGB 모델의 가장 큰 장점은 그것의 단순성과 처리속도이다. 위의 색상모델을 통한 연구모델로서, Non-parametric 피부분류 모델로는 Bayes classifier, Self Oraganizing Map (SOM) 등이 있고, Parametric 피부분류 모델로는 Single Gaussian, Gaussian Mixture Model, Elliptic Boundary Model등이 있으며, Dynamic Skin Distribution모델로는 Online Expectation Maximization 과 Dynamic Histogram, Gaussain Distribution Adaptation등이 있다. [6]



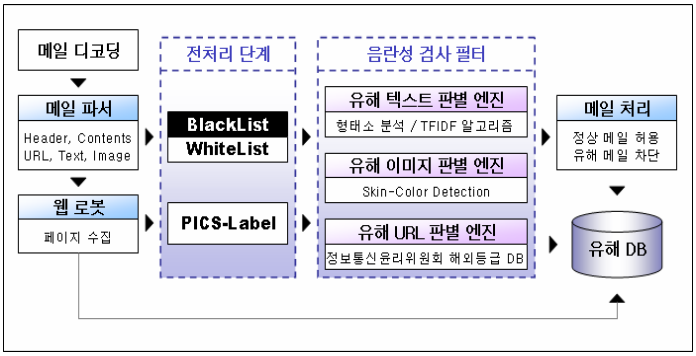
[그림 1] RGB Cube(上) · HSV Cone(下)

3. 제안모델

3.1 시스템 구조

본 논문에서 제안한 시스템은 메일 제목, 내용뿐만 아니라 메일에 링크된 페이지들의 음란성까지 검사하여 해당 메일의 유해성을 판별한다. 이를 위하여 음란 텍스트 및 이미지 판별 엔진과 웹 로봇 에이전트의 구현을 필요로 하였다.

판별과정에 있어서 우선적으로 메일을 파싱하고, 파싱한 헤더 정보를 정보통신부 스팸메일 방지 가이드라인과 블랙 리스트, 화이트 리스트를 이용하여 검사한 후 음란성 검사 필터를 이용하여 각 메일에 대해서 필터링을 한다. 화이트 리스트에 포함 되지 않았을 경우 웹 로봇 에이전트를 이용하여 메일에 링크된 페이지를 수집하고 수집된 페이지를 인터넷 내용선별 기술 표준을 이용하여 전처리 한 후 메일의 음란성 검사와 마찬가지로 필터를 이용하여 최종적으로 메일의 음란성에 대해 판별을 한다.



[그림 2] 시스템 구조

음란성 검사 필터는 입력되는 데이터에 대하여 정보통신윤리위원회의 해외등급DB를 이용하여 링크된 URL을 검사하고 형태소 분석을 이용한 텍스트 검사 그리고 Skin-Color 분포를 이용한 이미지 검사를 통해 최종적인 판단에 이르게 된다.

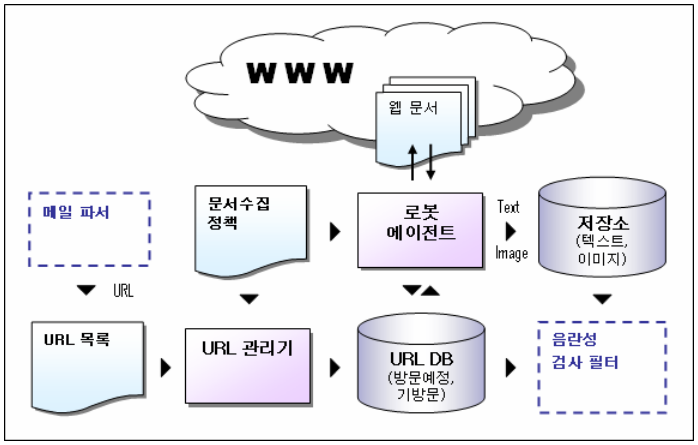
3.2 웹 로봇 에이전트를 이용한 웹 페이지 수집

웹 로봇 에이전트를 통한 하이퍼링크의 탐색이 음란메일 차단 시스템에 있어서 중요한 점은 유해·음란사이트의 경우 하이퍼링크를 거의 포함하고 있으며, 링크된 사이트 또한 유해 사이트일 확률이 매우 높다는 것이다.

본 논문에서 제안한 웹 로봇은 메일파서를 통해 추출된 URL을 시작점으로 하여, HTTP 프로토콜을 이용해 웹 공간에 분산되어 있는 웹 페이지에 대한 정보를 읽고 분석한다. 이때 에이전트가 제공받는 문서수집정책에 의해 필요한 정보만 수집하게 되며 기방문 DB내에 존재하는 페이지에 대해서는 데이터 수집을 하지 않는다. 새로운 페이지에 대해서는 핵심 컨텐츠를 가져 오고, 불필요한 광고나, 메뉴 등에 대한 내용은 배제하여 추출된 데이터의 질을 높인다.

재귀적인 URL 수집시에 링크된 웹 페이지가 쇼핑몰이나, 게시판으로 판별될 경우, 차후에 사용될 수 있는 링크는 없는 것으로 간주하여 URL에 대한 수집을 중단하게 된다. 이는 의미 없는 수집활동에 대한 이동을 최소화 하는 것이다.

이러한 웹페이지 수집을 통해 얻어진 텍스트 및 이미지 데이터는 텍스트·이미지 저장소에 저장하게 되는데, 이는 음란성 검사 필터의 소스로 사용되어진다.



[그림 3] 로봇 에이전트 구조도

3.3 전처리 과정

정보통신부에서 제공하는 스팸메일 방지 가이드라인에 따르면 광고성 전자우편 전송자는 광고성 전자우편 전송시 정보통신망법시행규칙 개정(안)에 따라 '(광고)' 또는 '(성인광고)'를 표기하도록 하였다.

정보통신부 가이드라인을 준수한 메일을 일차적으로 분류하고 그와 함께 블랙·화이트 리스트를 통해서 우선적인 필터링을 하게 된다.

또한 웹로봇 에이전트를 통해 수집된 웹페이지의 메타태그 내에서 W3C가 제정한 기술표준인 PICS(인터넷 내용 선별 기술표준: Platform for Internet Content Selection) 및 정보통신윤리위원회의 SafeNet등급과 국제적인 RSACi 및 ICRA 등급을 가지고 있는 HTML의 메타 태그(정보등급 부분)를 인식하여 분류한다.

```
<META http-equiv="PICS-label" content= '(PICS-1.1
"http://www.safenet.ne.kr/rating.html" l gen true[false] for
"정보제공자 자율등급표시 URL 명" r(n 1 s 1 v 2 l 3 i 0 h 1))'>
```

이러한 전처리 과정을 거침으로 해서 음란성 검사를 최소화 하여 시스템의 부하를 줄인다.

3.4 음란성 검사

정보통신윤리위원회에서는 국내법의 적용을 받지 않는 해외 음란·폭력정보에 대해 인터넷내용등급서비스(Safenet)의 등급 기준에 따라 해당 사이트의 등급을 설정하여 DB로 구축하고 있다. 이 해외등급DB를 이용하여 음란성 검사 필터가 메일 또는 수집된 웹 페이지에 연결된 URL의 음란성 여부를 판별한다.

텍스트의 경우 음란성 판별을 하기 위해 형태소 분석을 실시한 후, 미리 정의한 범주와 등급에 해당하는 품사별 음란 키워드를 추출하여 각 키워드에 대해 TFIDF 알고리즘을 사용한다.

다른 알고리즘에 비해 계산량이 적고 성능이 우수한 TFIDF 텍스트 분류 알고리즘은 문서에 출현하는 TF와 DF의 특성을 이용하여 문서를 분류하는 방법으로 TF와 DF의 역수인 IDF(Inverse Document Frequency)를 곱하는 형태로 각 단어의 중요도와 유사도를 계산한다. [7]

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

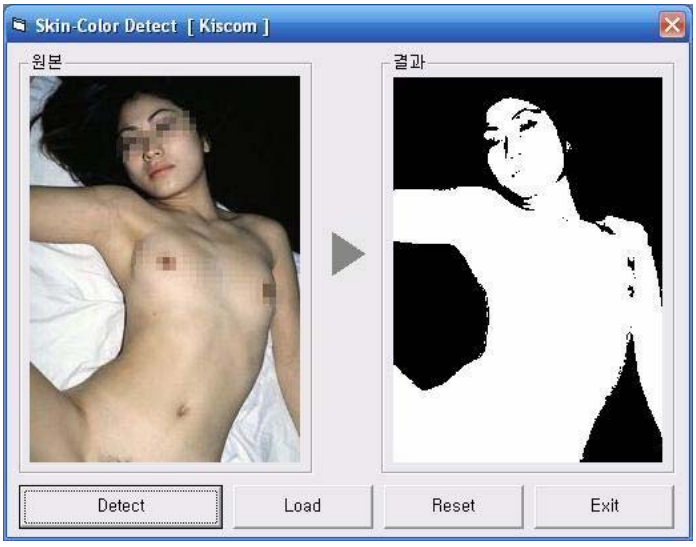
$$idf(w_i) = \log\left(\frac{n}{df(w_i)}\right)$$

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

이미지 판별은 사람의 피부색을 검출하여 그 분포에 따라 해당 이미지의 음란성을 판별한다. 피부색 판별에 있어서 RGB 모델은 각각의 성분이 빛의 세기를 포함하고 있으므로 빛의 변화에 의해 컬러의 검출을 방해한다. 그러므로 본 논문에서는 빛의 영향으로부터 RGB 모델에 비해 자유로운 HSV모델을 이용하여 피부색의 임계값을 설정하였다.

이는 아래의 수식과 조건을 만족하여야 하며 간단하고 빠른 판별 속도를 제공하는 장점이 있다.[8]

$$S > 10;\quad V > 40;\ S < -H - 0.1V + 110$$
$$H < -0.4V + 75$$
$$\text{if } H > 0\quad S < 0.08(100-V)H + 0.5V$$
$$\text{else } S < 0.5H + 35$$



[그림 4] 피부색 검출 결과

4. 실험 및 결과

4.1 실험 환경 및 입력 데이터

본 논문의 구현을 위해서 CPU 3.0 GHz, Main Memory 1 Gigabyte, Windows XP Professional 운영 체제에서, 음란성 검사 필터와 웹 로봇 에이전트는 Microsoft Visual C++ 6.0을 이용하여 구현하였고 사용자 인터페이스와 메일 수집 유닛, 정보 관리 유닛은 Microsoft Visual Basic 6.0으로 구현하였으며, 하부 저장 구조로는 My SQL 5.0.19를 사용하였다.

검사를 위한 음란 스팸메일은 2006년 6월 ~ 12월까지 수집한 음란메일 800건과 비음란메일 200건을 사용하였다. 음란 키워드 사전은 정보통신윤리위원회의 Safenet등급기준이 적용된 전자사전을 이용하였으며 음란 URL 목록은 정보통신윤리위원회 해외등급 DB를 이용하였다.

4.2 실험 평가 방법

실험 평가 방법으로는 얼마나 정확하게 분류되었는지 성능을 측정하기 위하여 메일의 음란성만 검사한 결과, 메일에 링크된 페이지들(2 Depth)의 음란성만을 검사한 결과, 메일의 음란성과 함께 링크된 페이지들(2 Depth)의 음란성까지 검사한 결과를, 아래와 같은 수식으로 정확도(Precision)와 재현율(Recall) 그리고 F-measure 측정식으로 평가하였다. [9]

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

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

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

4.3 실험 결과

실험 결과를 보면 메일 내 텍스트와 이미지 만을 검사한 분석결과에서는 성능평가 기준인 F-measure가 각각 84.0%와 51.1%로서 메일판별의 정확도에 있어서 다소 떨어지는 것을 확인할 수 있다. 웹 로봇 에이전트의 수집페이지 분석결과에서 링크된 페이지들(2 Depth)의 음란성만 검사한 결과가 96.1%의 정확도와 79.1%의 재현율을 보였고, 메일의 음란성만을 검사한 결과가 96.1%의 정확도와 85.6%의 재현율을 보였으며 메일의 음란성과 함께 링크된 페이지들(2 Depth)의 음란성까지 검사한 결과가 96.2%의 정확도와 97.6%의 재현율을 보이며 가장 좋은 성능을 나타내었다.

[표 1] 성능 측정 결과표

	Text 분석결과	Image 분석결과	Mail 분석결과 (A)	웹 로봇 수집 페이지 분석결과 (B)	통합 시스템 분석결과 (A + B)
RI	612	289	713	659	812
RL	593	278	685	633	781
NRD	181	189	172	174	169
Recall	0.741	0.348	0.856	0.791	0.976
Precision	0.969	0.962	0.961	0.961	0.962
F-measure	0.840	0.511	0.905	0.868	0.969

TI :전체 data (1000개 메일)
TL :전체 data에서 실제로 유해하다고 판단된 data (800개 메일)
RI :전체 data에서 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 스팸메일 차단시스템은 사용자가 직접 음란한 메일이라고 판단되는 메일에 대해 일일이 키워드를 설정하거나, 메일 내용 중에 텍스트만을 추출하여 패턴 매칭방법으로 분류하는 것이 대부분이었다. 더욱이 최근의 스팸메일의 특징은 텍스트 정보를 줄이고, 이미지와 링크를 제공하는 추세이므로 기존의 방법으로는 효과적인 스팸메일 차단 시스템을 구축하기가 어렵다.

따라서 본 논문은 기존 방법의 문제점을 해결하기 위하여 이미지 내 Skin-Color분포를 이용한 Human Detection 알고리즘(A)과 웹 로봇 에이전트(B)를 이용하였다. 형태소 분석과 전자(A)를 병합하여 적용한 경우 성능 측정에서 90% 정도의 F-measure를 보였지만 추가적으로 후자(B)와의 병합을 적용한 경우 성능 측정에서 97% 이상의 F-measure를 보이며, 신뢰성이 높은 스팸메일 차단 시스템을 구현할 수 있다는 것을 증명하였다. 여기서 중요한 점은 형태소 분석 및 Skin-Color 분포만을 이용하여 분석한 결과는 만족할 만한 성능을 보이진 않지만 웹 로봇 에이전트를 통한 분석과 병합하여 적용하였을 경우에는 아주 좋은 성능을 보여 준다는 것을 알 수 있다.

그러나 본 논문에서 구현한 웹 로봇 에이전트를 이용한 분석은 웹 로봇의 탐색시간 때문에 판별 시간이 늘어나는 문제점을 갖고 있다. 그러므로 향후에는 웹 로봇 성능 개선에 대한 연구가 필요하리라 생각된다.

6. 참고 문헌

[1] Rajashekar TB , “Web Search Engines”, Resonance: Journal of Science Education Volume 3 Number 11, pp. 40-53, Nov. 1998

[2] Gudivada, V.N.; Raghavan, V.V.; Grosky, W.I.; Kasanagottu, R. “Information retrieval on the World Wide Web” , IEEE JNL,Volume 1, Issue 5,Page(s):58 – 68, Sept.-Oct. 1997

[3] Alexandar Gelbukh, Grigori Sidorv, “Morpho logical Analysis of Inflective Languages through Gene ration”, J. Procesamiento de Lenguaje Natural, No 29, September 2002

[4] In Cheol Kim, Soo Sun Cho, “A Learning Agent for Automatic Bookmark Classification”, KIPS, VOL. 8-B NO.05 pp.0455~0462, Oct 2001

[5] H. Zheng, M. Daoudi and B. Jedynak, “Blocking Adult Images Based on Statistical Skin Detection”, Electronic Letters on Computer Vision and Image Analysis, Volume 4, Number 2, pages 1-14, 2004.

[6] Vezhnevets V., Sazonov V., Andreeva A., "A Surv ey on Pixel-Based Skin Color Detection Techniques". Proc. Graphicon-2003, pp. 85-92, Moscow, Russia, September 2003.

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

[8] Garcia, C., Tziritas, G. , “Face detection using quantized skin color regions merging and wavelet packet analysis” Multimedia, IEEE Transactions on, Volume 1, Issue 3, Page(s):264-277, Sept 1999

[9] 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