특집

유해정보방지기술

김영수, 남택용, 정종수 (한국전자통신연구원 정보보호연구단 네트워크보안그룹)

I. 서 론

21세기 정보화 혁명을 주도하고 있는 인터넷은 사람들로 하여금 시간과 공간을 뛰어넘어 언제, 어디서든 손쉽게 유용한 정보를 획득할 수 있게끔 하였다. 하지만 인터넷은 유용한 정보와 손쉬운 활용이라는 순기능과 더불어 사회적으로 통제를 필요로 하는 유해한 정보 역시 인터넷을 이용하는 이용자들에게 무차별적으로 제공함으로써 역기능을 발생시키고 있다. 사회적인 보호를 받아야 하는 청소년을 비롯한 판단력과 절제력이 부족한 인터넷 이용자들이 인터넷의 유해 정보를 아무런 제재 없이 접근할 수 있게 되어서 개인뿐 아니라 사회적인 문제가 되고 있다. 한편, 인터넷은 전 세계적으로 연결된 개방망이므로 유해정보 제공자를 각국의 법적, 제도적 장치를 이용하여 규제하는데 한계가 있다. 그리고, 유해 사이트, 유해성 스팸 메일, P2P 등 다양한 경로를 통해 유해 정보를 접할 수 있기 때문에, 어떤 시스템에 특화된 유해정보 분류기술을 개발하는 것은 바람직하지 않다. 따라서, 유해정보의 내용 자체에 기반하여 유해 여부를 자동으로 분류할 수 있는 유해정보 분류를 위한 핵심 기술의 연구 및 개발의 중요성이 점차 부각되고 있다. 이에 본고에서는 유해정보의 형태 중 텍스트와 이미지에 대하여 기존에 연구되었거나 현재 연구되고 있는 유해정보 분류기술에 대하여 고찰하도록 한다.

II. 유해 텍스트 분류 기술

텍스트 분류(text categorization)는 범주 정형화, 자질 추출 그리고 텍스트 분류 세 단계로 구성된다. 범주 정형화는 분류 대상이 되는 범주와 각 범주를 구분하는 기준을 정하는 단계이다. 자질(feature) 추출 단계는 문서에 나타나는 내용어(content word) 중 문

〈그림 1〉 텍스트 분류 기술맵

서 분류에 유용하게 사용될 만한 내용어를 선택하는 단계로서 형태소분석 기반 자질 추출 방법과 구문분석 기반 자질 추출 방법이 있다. 그리고, 텍스트 분류는 추출된 자질을 통해 텍스트를 분류하는 단계로 기계 학습 분야에서 사용되는 알고리즘들이 사용되는데, 크게 규칙 기반 모델(Rule based Model)과 학습 기반 모델(Inductive Learning Model), 그리고 유사도 기반 모델(Similarity based Model)로 분류할 수 있다.

1. 범주별 문서 수집

분류 대상이 되는 범주를 미리 정하고, 각 범주를 구분할 수 있는 기준을 정한다. 국내의 한 웹 검색 엔진의 디렉토리 서비스 메뉴를 예로 들면, [표 1]과 같은 12개의 범주와 이를 세분화한 하위 범주들로 구성할 수 있다.

〈표 1〉 웹 검색 엔진의 디렉토리 서비스 범주

가정, 여성	건강, 의학	게임
교육, 참고 자료	레크리에이션	북한
비즈니스, 경제	쇼핑	스포츠
엔터테인먼트	컴퓨터, 인터넷	학문, 과학

상위 12개의 범주를 시작으로 이를 세분화하여 불필요하거나 중복되는 범주를 임의로 제거하고, 최종적으로 60여개의 문서 범주로 분류할 수 있다. 상위 범주로부터 세부 범주들로 세분화한 예는 다음과 같다.

예) 스포츠 : 축구, 야구
건강, 의학 : 의학, 전통 의학, 증상, 질병
비즈니스, 경제 : 무역, 부동산, 투자, 금융, 재테크

범주가 결정이 되면, 각 범주에 해당하는 문서를 수집해야 한다. 수집된 문서들은 학습 모델을 생성하기 위한 학습 샘플로 활용되며, 각 범주의 특징을 대표하는 문서들을 정확하고 풍부하게 수집하는 것이 문서 분류의 성능에 큰 영향을 미친다.

2. 자질 추출

문서에 나타나는 내용어(content word) 중 문서 분류에 유용하게 사용될 만한 내용어를 선택하는 단계이다. 학습 문서에 나타나는 내용어의 수는 수만에서 수십만에 이르기 때문에, 모든 내용어가 자질(feature)로 선택될 경우 학습 및 분류 시간이 매우 오래 걸리게 되며 성능도 보장할 수 없다. 성능 저하 없이 자질의 수를 줄이기 위하여, 학습 문서에 나타나는 내용어의 정보량을 계산하고 정보량이 큰 내용어만을 자질로 선택하려는 연구가 활발히 진행 중에 있다. 자질 추출 방법은 형태소 분석 기반 자질 추출과 구문 분석 기반 자질 추출의 두 가지로 이루어진다.

1) 형태소 분석 기반 자질 추출

형태소 분석에 의한 자질 추출은, 형태소 분석에 의해 자질어 후보를 추출한 후에 통계적 기법에 의해 자질어를 선별하고, 용어 통제표를 이용하여 대표어를 추출함으로써 적합한 자질어 리스트를 선정하게 된다. 자질어 리스트에 있는 자질어들이 문서에서 차지하는 중요도를 판단하여 가중치를 부여함으로써, 가중치가 큰 자질어가 분류에 더 많은 영향을 줄 수 있도록 한다. 일반적으로 사용하는 자질어 선별 기법은 다음과 같다.[1]

① TF(Term Frequency) 이용: 전체 텍스트 집합 중 특정 용어가 출현한 빈도를 나타내며 가장 단순한 방법이라 할 수 있다.

② DF(Document Frequency) 이용: 전체 텍스트 집합 중 특정 용어가 출현한 문서의 수를 의미하며, 일정 빈도(임계치) 이하의 텍스트에서 출현하는 용어들은 텍스트에서의 중요도가 낮다고 판단하고 이를 제거한다. DF를 이용할 때의 기본 가정은 DF가 아주 작은 용어는 특정 주제 범주를 대표할 만한 충분한 정보가 되지 못하고 전체적인 성능에도 큰 영향을 미치지 못한다는 것에 있다. DF는 매우 간단한 방법이며 계산량 또한 적으나, 정보 검색의 경우 전통적으로 문서 빈도 값이 낮을수록 색인어로서의 가중치를 높게 할당하는 것과는 대치된다.

③ MI(Mutual Information) 이용: 두 용어 중 한 용어가 다른 용어에 대해 갖고 있는 정보량을 이용한다. 즉, 두 용어 중 한 용어가 출현했다는 사건이 다른 용어의 출현 여부를 예측하는 데 기여하는 정도를 수치적으로 나타내는 것이다. A가 범주 c 에 속한 문서 중 용어 t를 가진 문서의 개수, B가 범주 c에 속하지 않는 문서 중 용어 t를 가진 문서의 개수, 그리고 C가 범주 c에 속한 문서 중 용어 t가 없는 문서의 개수라고 한다면, MI는 아래와 같은 수식으로 계산된다.(여기서 Pr은 확률을 의미하고 N은 전체 문서의 개수를 나타낸다)

$$MI(t,c) = \log\frac{Pr(t \wedge c)}{Pr(t) \times Pr(c)}$$

$$MI(t,c) \approx log\frac{A \times N}{(A+C) \times (A+B)}$$

$$MI_{avg}(t) = \sum_{i=1}^{m} Pr(c_i) I(t,c_i)$$

$$MI_{\max}(t) = \max_{i=1}^{m} \{I(t,c_i)\}$$

④ IG(Information Gain) 이용: 특정 단어의 출현 여부가 문서 분류에 기여하는 정도를 계산하여 기여도가 높은 자질만을 선택하는 것으로, 모든 용어들의 정보 획득량을 계산하여 일정 임계값 이상의 값을 갖는 용어들만을 자질로 선택하게 된다. 문서에서의 출현 빈도뿐 아니라 출현하지 않은 빈도까지 고려하여 각 범주에서의 용어 정보량을 계산한다. 용어의 출현 빈도를 고려한 MI의 평균값과 용어가 출현하지 않은 빈도의 MI의 평균값의 합으로 계산되므로, MI보다 문서 분류에서 더 좋은 성능을 보이는 경우가 많다. 범주 집합을 $\{c_1, c_2, \cdots, c_m\}$라 하면, 수식은 다음과 같다.

$$IG(t) = -\sum_{i=1}^{m} Pr(c_i) log Pr(c_i) + Pr(t)\sum_{i=1}^{m} Pr(c_i|t) log Pr(c_i|t) + Pr(\bar{t})\sum_{i=1}^{m} Pr(c_i|\bar{t}) log Pr(c_i|\bar{t})$$

⑤ CHI(□□² statistics) 이용: 용어 t와 범주 c 간의 의존성을 측정해 용어의 중요도를 구하는 방법으로, t와 c 두 값의 차가 클수록 용어 t가 자질로 선정될 확률이 높아진다. 문서 빈도를 사용해 범주별 발생분포가 일반적인 단어들의 발생분포와 다른 정도를 계산하고,

$$\square\square^2(t,c) = \frac{N \times (AD-CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

$$\square\square^2_{avg}(t) = \sum_{i=1}^{m} Pr(c_i) x^2(t,c_i)$$

$$\square\square^2_{\max}(t) = \max_{i=1}^{m} \{x^2(t,c_i)\}$$

그 차이가 특정 값 이상인 단어를 자질로 선정하게 된다. 저 빈도 용어에 대해서는 신뢰할 수 없다고 알려져 있다. 각 용어 및 범주에 대해 통계값을 계산 후 각 용어마다 계산된 카이제곱 통계량의 평균이나 최대값을 구한다.

각 학습 대상 문서를 학습에 적합한 형태로 표현하기 위해서는 선택된 자질에 가중치를 부여해야 한다. 각 문서들은 전 단계에서 선택된 자질들의 값으로 표현된다. 전 단계에서 선택된 내용어(자질어)들은 자질이 되고 문서 내에서의 빈도수 등을 이용한 단어 가중치가 값이 된다. 즉, 〈자질:값〉 형태의 표현법이 사용되며, 가장 일반적으로 사용되는 문서 표현 방법은 벡터 공간 모델이다. 이것은 문서 전체에 나타난 각 자질의 출현 빈도(TF)를 이용하여 문서를 하나의 벡터로 표현하는 것이다. 고빈도 용어는 대부분 기능어로서 많이 등장하지만 그 문서의 내용을 나타내지는 못하므로, TF와 역문헌빈도(Inverse Document Frequency)를 함께 고려하여 가중치를 부여하는 방법이 주로 사용된다. 각 자질의 가중치는 해당 문서에서 각 자질의 빈도(TF)와 역문헌빈도(IDF)의 곱으로 나타내어진다.[2] 즉, 문서 i에서 자질 k의 가중치 a_{ik}는 다음과 같이 표현된다. 여기서 f_{ik}는 문서 i에서 출현한 자질 k의 빈도이고, N은 전체 문서의 수, 그리고 n_k는 자질 k가 출현한 문서의 수이다. TF-IDF 가중치 방법은 여러 가지 변형이 존재한다.

$$a_{ik} = f_{ik} \times log(\frac{N}{n_k})$$

문서 간 분리도 대신 범주 간 분리도가 높은 용어에 높은 가중치를 주는 방법으로 TF-ICF (Inverse Category Frequency) 가중치 부여 방법이 있다. 범주 분리 능력이 우수한 색인어에 높은 가중치 부여. 즉, 소수 범주에 많이 나온 용어에 높은 가중치를 주고, 여러 범주에서 고르게 나오는 용어에 대해 낮은 가중치를 부여한다. 문서 분류의 경우 범주 간 구분에 도움이 되는 색인어가 중요도가 높다고 할 수 있으므로 역 카테고리빈도가 역 문헌빈도보다 의미 있는 계산 방법이다. 한편, 가중치 부여시 사용되는 TF는 단순 TF일 수도 있으나, 이진TF, 로그TF, 더블로그TF, 루트TF, 보정TF, 더블로그2TF, 루트직선TF 등 여러 가지 변형된 형태가 사용될 수 있다.

2) 구문 분석 기반 자질 추출

구문 분석을 통하여 특정 기능을 가진 단어나 구를 식별하고 이를 자질어로 사용하는 기법이다. 구문 분석은 대체로 자연어 처리에서 연구되고 있는 기술로서, 구문 분석 기술은 질의응답 시스템이나 기계 번역 등 대부분의 응용 분야에서 필요로 하고 있지만, 자연 언어 처리에서 완벽한 구문 분석은 매우 어려운 일이다. 자질어 추출에서는 구문 분석의 어려움에 비해 그 효과가 크지 않기 때문에 실제로 의미 처리를 포함하여 구문 분석에 의한 자질 추출 시스템을 구현하기가 불가능하다.

3. 텍스트 분류

많은 양의 문서를 관리하고 이를 효율적으로 검색하기 위한 문서 분류 모델에는 기계 학습 분야에서 사용되는 알고리즘들이 사용되는데, 크게 규칙 기반 모델(Rule based

Model)과 학습 기반 모델(Inductive Learning Model), 그리고 유사도 기반 모델(Similarity based Model)로 분류된다. 먼저 규칙 기반 모델은 학습 문서들에서 나타나는 범주간의 구별된 규칙을 전문가가 찾아내고 그 규칙을 이용하여 문서를 분류하는 모델이다. 학습 모델은 학습 문서에서 자질을 추출하여 이를 확률적인 접근 방법으로 사용한 베이지언 확률 모델과 트리 구조로 표현하여 자질의 유무로 범주를 결정하는 결정 트리 모델, 학습 문서를 통해 생성된 양성 자질(positive feature)과 음성 자질(negative feature)을 벡터 공간으로 표현하고 이들의 차이를 극명하게 하는 지지 벡터(support vector)를 찾는 지지 벡터 기계 등이 있다.[3][4] 또한, 정보 검색 관점에서 분류할 대상 문서를 질의로 보고 이와 유사한 문서를 찾는 방법인 k -최근린법과 선형 분리기 등이 있다.

1) 규칙 기반 분류

학습 문서들에서 나타나는 범주간의 구별된 규칙을 전문가가 찾아내고 그 규칙을 이용하여 문서를 분류하는 모델이다. 작성된 규칙이 단어들의 불리안(boolean) 조합이라면, 입력 문서가 특정 범주의 불리안 조합을 포함할 때 해당 범주로 분류하게 된다. 규칙이 잘 정의되고 오랜 기간 동안 전문가가 그 규칙을 다듬는다면 매우 높은 분류 성능을 보일 수 있지만, 규칙을 정의하고 유지하는데 많은 시간과 비용이 필요하다는 단점이 있다.

2) 학습 기반 분류

① 지지 벡터 기계 (SVM: Support Vector Machine): 두개의 범주를 구분하는 문제를 해결하기 위해 1995년에 소개된 비교적 최근의 학습 기법으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리할 수 있는 결정면(decision surface)을 찾는 모델이다. 지지 벡터 기계는 직선으로 나눌 수 있는 문제에 사용되는 알고리즘이지만, 다차원의 부드러운 곡선을 이용하여 경계면을 설정하거나, 실제 데이터 벡터를 새로운 자질을 포함한 새로운 벡터 공간에 매핑하는 방법을 통해서 직선으로 나눌 수 없는 문제도 해결할 수 있다. Joachims는 최근에 지지 벡터 기계 모델을 문서 범주화에 적용하여 좋은 성능을 보였다.[5]

② 베이지언 확률 모델 (Bayesian Probability Model): 주로 기계 학습(machine learning)에서 연구되어 왔다. 이 모델은 주어진 입력 문서의 각 범주에 할당될 확률을 구하기 위해서 문장에 속해 있는 용어들과 범주와의 결합 확률값을 사용하는 방법이다.

③ 결정 트리 모델 (Decision Tree Model): 결정 트리는 학습 결과를 이용해 규칙을 생성하고, 어떤 자질이 어떤 효과를 발휘하는지 분석이 용이하게 해준다. 그리고 과잉 학습을 방지하기 위해 가지치기(pruning) 기능을 포함한다. 하나의 범주에 하나의 결정 트리 학습 모델이 생기며, 결정 트리의 각 노드는 특정 자질이 문서에 나타나는지 여부 또는 빈도를 표현한다.

3) 유사도 기반 분류

① k-최근린법 (k-Nearest Neighbor): 40년 동안 패턴 인식에서 연구가 이루어진 잘 알려진 예제 기반의 학습 방법으로서 문서

범주화 연구가 시작된 초기부터 적용되어 왔으며 가장 좋은 성능을 가진 분류기중의 하나이다. 최근린법의 기본 알고리즘은 매우 간단하다. 즉, 실험 문서가 주어졌을 때 학습 문서 중에서 실험 문서와의 유사도가 가장 높은 k개의 문서를 추출하고 그들을 사용하여 각 후보 범주의 순위를 매기는 방법이다. k개의 추출된 학습 문서는 미리 정해진 범주가 있으므로, 각 범주의 실험 문서와의 유사도는 각 범주별로 추출된 k개의 문서와 실험 문서와의 유사도의 합으로 계산된다. k-최근린법은 모든 범주의 문서를 하나의 공간에 표시하면서 범주의 영역을 나누려는 노력을 할 필요가 없으므로, 선형분리 문제로 고생하지 않으며, 지지 벡터 기계 모델과 더불어 문서 범주화 분야에서 가장 좋은 성능을 보이는 것으로 알려져 있다. 그러나, 각 실험 문서에 대해서 모든 학습 문서를 비교해야 하기 때문에 수행 속도가 느리다는 치명적인 단점을 가지고 있으며, 이를 극복하기 위한 연구가 진행 중이다. 최근린법은 정보검색 분야에서 주로 사용되고 있는 기술이다.

② 선형 분류 모델 (Linear Classification Model): 범주 c의 가중치 벡터 Wc와 문서 자질 벡터 d의 내적(inner product) 값을 문서 d에 대한 범주 c의 할당 여부 결정에 사용하는 기법이다. 선형 분류 모델은 일반적으로 정보검색 모델에서의 검색 결과에 대한 순위 결정에 사용되나, 문서 범주화에서는 문서에 대한 범주의 할당 여부를 결정하기 위한 선형 분류기로 사용된다. 선형 분류기의 가중치 벡터를 학습시키기 위하여 여러 가지 알고리즘이 사용되는데, 전체 학습 문서에 대해 한번의 계산으로 가중치 벡터를 생성해 내는 Rocchio 알고리즘과 개별 학습 문서들을 가중치 벡터 조정에 참여시키는 Widrow-Hoff 알고리즘이 대표적인 것이다.[6][7]

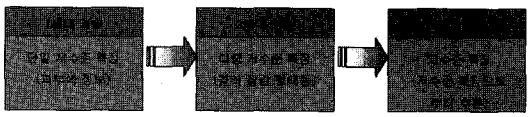
III. 유해 이미지 분류 기술

유해 이미지 분류 기술이란, 입력 이미지가 주어졌을 때 이미지 처리 기술을 사용하여 이미지의 유해성을 판단하고 분류하는 기술을 의미한다. 유해 이미지 분류 기술은 1996년 발표된 "Finding naked people"이라는 논문을 시작으로 하여 계속 발전 되어오고 있다.[8] 초기에는 내용기반 이미지 검색(CBIR: Content Based Image Retrieval)에 사용되는 이미지의 특징과 분류 기술을 가지고 유해 이미지를 구분하려는 시도가 많이 있었다. 하지만 최근에는 유해 이미지에 특화된 특징을 추출하고, 추출된 특징을 학습 알고리즘의 입력 값으로 넣어서 유해 이미지를 판단하려는 시도가 많이 진행 되고 있다. 초기 유해 이미지 분류 소프트웨어는 초기에는 이미지의 피부색 정보만을 추출하여 유해성을 판단하였기 때문에, 판단의 정확성이 떨어졌다. 이와 같은 문제점을 개선하기 위해서 이미지의 피부색 정보 뿐 만 아니라, 이미지의 다른 정보(컬러, 형태, 질감)들을 같이 사용하는 유해 이미지 분류 소프트웨어들이 최근에 개발 되고 있다. 현재, 유해 이미지 분류 소프트웨어는 스팸 메일 차단 소프트웨어, 유해 사이트 차단 소프트웨어, 등급 분류 서버 소프트웨어 등에 탑재 될 수 있는 컴포넌트 형태로 많이 개발 되고 있다. 유해 이미지 분류 기술의 구성요소로는 유해이미지 전처리기술, 유해이미지 특징추출 기술,

유해 이미지 분류 기술, 유해이미지 분류 시스템 통합 기술, 유해이미지 분류시스템 평가기술 등이 있다. 그 중에서 유해이미지 특징추출 기술과 유해 이미지 분류 기술이 주로 연구되고 있는 반면에, 다른 요소에 대한 연구는 부족한 실정이다. 본 장에서는 주로 연구되고 있는 유해이미지 특징추출 기술과 분류 기술의 현황에 대해 살펴보고자 한다.

1. 유해 이미지 특징 추출 기술

유해 이미지를 분류하기 위해서는 유해 이미지를 무해 이미지와 구별할 수 있는 특징을 선택하고 추출하는 기술이 필요하며, 이런 기술을 유해 이미지 특징 추출 기술이라고 한다. 유해 이미지 특징 추출 기술은 이미지로부터 추출하는 정보에 따라서 다음과 같이 구분 할 수 있다.

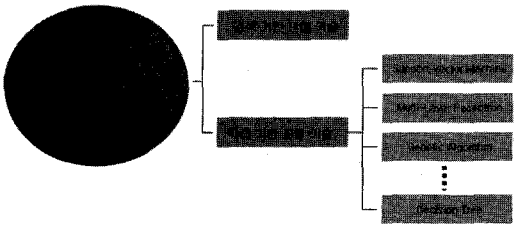


〈그림 2〉 유해 이미지 특징 추출 기술 발전도

1세대 유해 이미지 특징 추출 기술은 이미지에서 피부색 영역을 찾은 후, 피부색 영역의 크기, 이미지 크기에 대한 상대적 피부색 영역의 크기, 피부색 영역의 개수 등을 피부색에 관한 몇 가지 정보만을 추출하는 기술을 의미한다. 2세대 유해 이미지 특징 추출 기술은 이미지에서 피부색 영역의 정보뿐만 아니라, 이미지의 컬러, 형태, 질감 등과 같은 여러 가지 저수준 특징(low level feature)을 추출하는 기술을 의미한다. 이 때 이미지로부터 컬러, 형태, 질감 등을 정보를 추출하는 방법으로는 MPEG-7 기반 방법[9]과 Non MPEG-7 기반 방법이 있다. 3세대 유해 이미지 특징 추출 기술은 이미지의 컬러, 형태, 질감 등의 저수준 특징들로부터 새로운 정보를 추론하여 의미를 갖는 고수준 특징을 추출하는 기술을 의미한다. 이미지에서 피부색 영역, 에지(edge) 등의 저수준 특징을 추출한 후, 이를 사용하여 이미지 속에 존재하는 사람의 포즈(pose)를 분석하고, 포즈의 의미를 분석하여 유해성 여부를 분류하는 방법이 3세대 유해 이미지 특징 추출 기술의 한 예라고 할 수 있다.

2. 유해 이미지 분류 기술

유해 이미지 분류 기술이란, 유해 이미지 특징 추출 기술로부터 추출된 이미지의 특징을 입력 값으로 사용하여 유해성 여부를 결정하는 기술을 의미한다. 유해 이미지 분류 기술은 검색 기반 분류 기술과 학습 기반 분류 기술이 있다.



〈그림 3〉 유해 이미지 분류 기술

검색 기반 분류 기술은 유해 기준 이미지와 무해 기준 이미지를 데이터베이스에 저장하고 있다가, 입력 이미지가 들어오면 모든 기준 데이터와 유사도를 측정 한다. 그리고 기준 이미지 중에서 입력 이미지와 유사도가

높은 순으로 정렬하여 K개의 기준 이미지를 선택하고, 그 중에서 유해 이미지가 N개 많으면 유해 이미지로 판정하는 방법이다. 검색 기반 분류 기술은 유사도를 측정할 때 가중치(weights)를 주는 방법에 다양한 방법이 존재한다. 검색 기반 분류 기술은 모든 기준 데이터와 유사도를 측정해야 하기 때문에 처리 시간이 오래 소요 되며, 메모리 공간도 많이 차지하고, 기준 이미지에 따라서 성능이 좌우되는 단점을 가진다. 학습 기반 분류 기술은 유해 및 무해 이미지 학습 샘플들로부터 범주간의 특징을 추출하고 이를 사용하여 분류 모델을 생성한 후, 입력 이미지에 대해서 생성된 분류 모델을 사용하여 이미지의 유해성을 판단하게 방법을 의미한다. 학습 기반 분류 기술은 검색 기반 분류 기술에 비해서 처리 속도가 빠르며, 메모리 공간도 효율적으로 사용 할 수 있고, 사용자가 임의로 모델을 정하지 않고 학습 하는 과정동안 최적화된 값을 찾기 때문에 사용자의 개입을 최소로 할 수 있는 장점이 있다. (학습 기반 분류 기술은 2장 참조)

Ⅳ. 향후 발전 방향

1. 유해 텍스트 분류 기술의 발전 방향

구문 분석 기반 자질 추출 기술은 구문 분석을 통하여 특정 기능을 가진 단어나 구를 식별하고 이를 자질어로 사용하는 기법으로 대체로 자연어 처리에서 사용하는 구문 분석 기법들을 이용한다. 구문 분석 기술은 질의응답 시스템이나 기계 번역 등의 응용 분야에서 절실히 필요로 하고 있지만, 자연 언어 처리에서 완벽한 구문 분석은 매우 어려운 일이다. 또한 자질어 추출에서는 구문 분석의 어려움에 비해 그 효과가 크지 않기 때문에 실제로 형태소 기반 자질 추출 기술을 주로 사용한다. 자질어 선별 기법의 경우 TF, DF, MI, CHI, IG 등 여러 가지 기법들이 제안되었고, 가중치 부여 방법 역시 TF-IDF, TF-ICF 등 여러 가지 방법들이 제안되어 왔다. 이들 형태소 분석 기반 자질 추출 기술들은 자질 추출의 대상과 적용 분야에 따라 적절한 기법들이 채택되어 사용되고 있다. 한편, 학습 문서들에서 나타나는 범주간의 구별된 규칙을 전문가가 찾아내고 그 규칙을 이용하여 문서를 분류하는 규칙 기반 텍스트 분류 기술은 규칙이 잘 정의되고 오랜 기간 동안 전문가가 그 규칙을 다듬는다면 매우 높은 분류 성능을 보일 수 있지만, 규칙을 정의하고 유지하는데 많은 시간과 비용이 필요하므로 최근에는 학습 기반 텍스트 분류 방법이나 유사도 기반 텍스트 분류 방식이 주로 사용된다. 유사도 기반 분류 방법의 경우 정보 검색 관점에서 분류 대상 문서를 질의로 보고 이와 유사한 문서를 찾는 방법으로 주로 텍스트 검색 분야에서 많이 이용되고 있으나 텍스트 분류 분야에서는 주로 학습 기반 기법에 관한 연구가 주를 이루고 있다.

2. 유해 이미지 분류 기술의 발전 방향

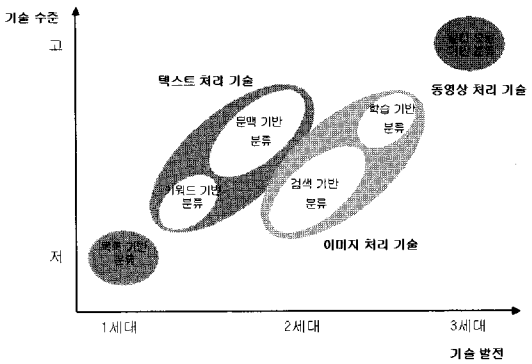
고수준 특징 추출 기술의 경우, 현재 유해 이미지 분류에 사용되고 있는 특징들은 원본 이미지의 다른 형태, 또는 간략화된 형태이기 때문에 저수준의 특징이라고 하며, 저수준의 특징은 이미지에 따라서 많이 달라지

며, 이것은 유해 이미지 분류에 가장 큰 어려움인 부정형성을 더 크게 만든다. 고수준의 특징을 사용하여 유해 이미지를 분류하게 되면 유해 이미지의 부정형성을 줄일 수 있고, 이미지의 변화에 영향을 많이 받지 않기 때문에 유해 이미지 분류의 정확성을 높일 수 있다. 따라서 앞으로는 유해 이미지 분류에 고수준 특징 추출 기술을 적용하려는 시도가 많이 있을 것이다. 한편, 학습 기반 분류 기술의 경우, 유해 이미지 분류 초기에는 검색 기반 또는 규칙 기반 분류 기술을 사용하였다. 그러나, 규칙 기반, 검색 기반 분류 기술의 경우 관리자의 개입이 많이 필요하여 관리자에 의해 성능이 좌우되는 문제점이 존재하였다. 규칙 기반 분류 기술의 경우에는 규칙에 의해서만 판단하므로, 처리 시간이 적게 소요되는 장점이 있지만, 무해 이미지를 유해 이미지로 판정하거나, 유해 이미지를 무해 이미지로 오판하는 경우가 많이 발생하는 단점이 있다. 검색 기반의 분류 기술은 분류에 기준이 되는 데이터가 증가함에 따라서 처리 시간 및 메모리 공간이 계속 증가하는 문제점 있다. 이런 문제점을 개선하기 위해서 최근에는 학습 기반의 분류 기술이 등장하고 있다. 학습 기반의 분류 기술의 경우에는 학습 샘플만 주어지만 학습 하는 동안 최적화된 분류 모델을 자동으로 생성해주기 때문에 관리자의 개입이 거의 필요하지 않다. 또한 기준 데이터를 계속 가지고 있지 않고, 학습 샘플들로부터 공통된 특징을 추출하여 분류 모델을 생성하기 때문에 메모리 공간을 검색 기반 분류 기술보다 효율적으로 사용할 수 있고, 처리 시간도 더 적게 소요된다. 이런 장점들로 인하여 앞으로는 학습 기반 분류 기술이 유해 이미지 분류에 더 많이 사용 될 것 이며, 특히 하나의 분류기를 사용하지 않고, 병렬 또는 직렬로 연결 하여 사용하는 분류기 조합을 통하여 분류 성능을 높이는 기술에 관한 연구가 활발히 진행될 것이다.

데이터 마이닝은 데이터들간의 상호관계를 분석하지 이전에 존재하지 않았던 새로운 정보를 얻는 것을 의미한다. 최근 멀티미디어 데이터들이 폭발적으로 증가함에 따라서 멀티미디어 데이터들의 상호관계를 분석하려는 멀티미디어 마이닝에 관한 연구가 활발히 진행되고 있으며, 이를 멀티미디어 검색에 사용하려는 연구 또한 많이 이루어지고 있다. 이미지들의 상호관계를 분석하고 이를 통하여 이미지를 모델링 할 경우, 유해 이미지 분류의 정확도를 높일 수 있는 부가 정보를 얻을 수 있기 때문에, 앞으로 멀티미디어 마이닝 기술이 유해 이미지 분류에 많이 사용 될 것으로 예상 된다. 한편, 유해 이미지 분류 기술은 이미지 내에 존재하는 정보만을 사용하여 유해성을 판단하는 것에 비해, 문서 분류 기술이나 동영상 분류 기술에서는 특징이 다른 여러 가지 정보(텍스트, 이미지, 동영상, 오디오)들을 결합해서 문서나 동영상의 최종 범주를 결정하는 멀티 모달 방법을 사용하고 있다. 현재 유해 이미지 특징 추출 기술에서 대부분의 정보를 추출 하여 사용하고 있기 때문에, 분류 성능 향상을 위해서는 지금까지 사용하지 않았던 이미지의 생성 정보(텍스트), 파일명 등의 정보들을 결합하여 유해 이미지를 분류 하려는 방향으로 기술이 발전 할 것으로 예상 된다.

V. 결론

지금까지 유해정보의 대표적인 형태인 텍스트와 이미지에 대한 분류 기술과 각각의 향후 추세에 대하여 살펴보았다. 유해 텍스트는 야설과 같이 소설의 형태, 유해 사이트 초기 화면이나 메뉴와 같은 형태, 게시판 등의 다양한 형태로 존재하기 때문에 모든 형태의 유해 정보를 분류할 수 있는 기술 개발에 제약이 있다. 즉, 유해 사이트 입구 페이지, 이미지/동영상 리뷰 문서는 단문 또는 단어만으로 구성되고 문서의 길이가 짧아서 분석 대상이 되는 자료가 부족하며, 야설의 경우 문장 단위나 단어 단위로 볼 때 유해하지 않은 부분이 상당히 많이 포함되어 있어서 무해한 문서로 분류할 가능성이 매우 높다. 또한, 유해어의 특성상 표준어가 아닌 단어를 많이 사용하고 동일한 단어도 다양한 형태로 표현하며, 계속해서 새로운 유해어가 나타나기 때문에 최신 유해어 정보를 유지하기가 매우 어렵다. 즉, 유해한 뜻을 지닌 은어나 비속어가 다양하고, 정형화되어 관리되지 않으므로 수집이 어렵고, 새로운 유해어는 어느 정도 널리 퍼지기 전에는 알기 어려울 뿐만 아니라, 유해어를 사용할 때 맞춤법이나 띄어쓰기, 기타 규칙을 잘 따르지 않기 때문에 약간의 변형이 있을 경우 자동으로 파악하기 어렵다. 유해 이미지 분류 기술의 경우, 이미지의 부정형성이 분류 성능을 향상시키는데 방해 요인이 된다. 자동 분류를 위해서는 컴퓨터에 유해 이미지의 특징을 학습 시켜야 하는데, 유해 이미지로 분류될 수 있는 이미지의 특징을 정의하기가 어려운 실정이다. 다양한 이미지의 품질 수준 또한 분류 성능 향상에 방해 요인이 된다. 모든 유해 이미지들이 분석하기에 적합한 해상도의 데이터를 포함하고 있는 것이 아니며 또한 이미지의 사이즈, 사진을 찍었을 때의 조명, 사진의 구도 등의 관점에서 유해 이미지들이 다양한 품질 수준으로 분포하고 있기 때문에 이들 모두를 요구 수준에 맞게 판정하는 분류 기술을 개발하는데 어려움이 있다. 또한, 현재 유통되고 있는 이미지가 고용량화됨에 따라 이미지 처리에 소요되는 시간도 같이 증가할 수밖에 없으며, 이는 실시간 유해 이미지 분류 기술 개발 시에 이미지 처리 속도 향상에 방해 요인이 될 수 있다. 위와 같은 유해정보 방지기술 개발에 관한 문제점들을 해결할 수 있는 새로운 방안들이 요구된다.



〈그림 4〉 유해 컨텐츠 분류 기술의 발전 예상도

〈그림 4〉와 같이 유해정보 분류 기술은 텍스트 처리 기술에서 이미지 및 동영상 처리 기술로 발전하고 있으며, 향후 텍스트와 이미지 등 멀티미디어 정보를 이용한 멀티 모달 기반의 분류 기술로 발전할 것으로 예상된다. 텍스트 처리 기술은 목록기반 분류 기술에서 키워드 기반과 문맥 기반 분류 기술로 발전하고 있으며, 이미지 처리 기술은 검

색기반 분류 기술에서 학습 기반 분류 기술로 발전하고 있다. 그리고 본고에서는 다루지 않았으나, 유해정보의 중요한 형태 중 하나인 동영상에 대한 처리 기술의 경우 기존의 텍스트 및 이미지 처리 기술을 포함한 멀티모달기반의 분류 기술이 핵심 기술로 부상할 것으로 예상된다.

참고 문헌

[1] Y.Yang and J.O.Pederson, A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning (ICML' 97), 412-420, 1997

[2] W.Frakes and R.Baeza-Yates, Information Retrieval: Data Structures and Algorithms, Chapter7, Prentice-Hall, 1992

[3] F.Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol43 No1, 1-47, 2002

[4] C.Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, Vol2, 121-167, 1998

[5] T.Joachims, Estimating the Generalization Performance of a SVM Efficiently, Proceedings of the International Conference on Machine Learning, 2000

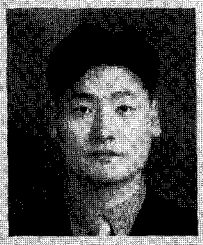
[6] T.Joachims, A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Proceedings of ICML-97, 14th International Conference on Machine Learning, 1996

[7] H.Abdi, D.Valentin, B.Edelman and A.O' Toole, A Widrow-Hoff learning-rule for a generalization of the linear auto-associator, Journal of Mathematical Psychology, Vol 40, 175-182, 1996

[8] M.Fleck, D.Forsyth and C.Bregler, Finding Naked People, European Conference on Computer Vision, Vol2, 592-602, 1996

[9] B.S.Manjunath, P.Salembier and T.Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Wiley, 2002

저자소개

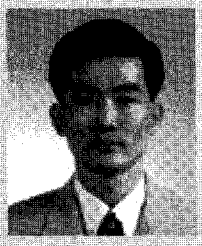


김 영 수

1998년 02월 성균관대학교 정보공학과 공학사
2000년 02월 성균관대학교 컴퓨터공학과 석사
2000년 02월 – 현재 한국전자통신연구원 선임연구원
주관심 분야 암호학, 네트워크보안, 개인정보보호

남 택 용

1987년 – 현재 한국전자통신연구원 개인정보보호연구팀 팀장(책임연구원)
2004년 – 현재 과학기술연합대학원대학교(UST) 정보보호공학 교수
주관심 분야 개인정보보호, 네트워크보안, 인터넷기술, 차세대네트워크구조



장 종 수

1984년 02월 경북대학교 전자공학과 공학사
1986년 02월 경북대학교 전자공학과 공학석사
2000년 02월 충북대학교 컴퓨터공학과 공학박사
1989년 – 현재 한국전자통신연구원 네트워크보안그룹 그룹장(책임연구원)
주관심 분야 네트워크보안, 웹서비스보안, Secure OS, IDS/IPS, 트래픽관리