텍스트 마이닝을 이용한 웹 포럼 불량글 탐지 모델

우지영*

요약

웹포럼상의 불량글은 사용자의 불편함을 증가시키고 사용자 의견 수렴의 원천으로서의 웹포럼의 가치를 하락시킨다. 웹포럼상의 글의 중요도는 일반적으로 해당글에 참여한 사용자 수나 댓글수로 측정하므로, 이러한 불량글은 웹포럼상의 의견수집 시 많은 불필요한 데이터를 포함하게 되고, 이렇게 분석된 결과는 왜곡된 결과를 내기도 한다. 본 연구에서는 웹포럼상의 불량글을 자동 분류하는 탐지 모델을 제안한다. 웹포럼상의 글에 대해 텍스트마이닝 기술을 적용하여 글의 구조적, 의미적, 형식적 텍스트 특성을 추출하고, 정상과 불량글을 두 개의 클래스로 한 분류 모델을 학습시킨다. 불량글을 자동으로 분류하는 모델을 학습시키기 위해서는 불량글과 정상글로 사전에 분류된 학습데이타가 필요하기 때문에, 평가자가 글의 불량성 여부를 평가하도록 한다. 자동화된 학습 모델을 구축하기 위해 데이터마이닝 및 텍스트마이닝에서 그 성능이 입증된, Naive Bayesian, Support Vector Machine (SVM), Decision Tree를 이용하였다. 이를 통해 불량글과 정상글을 구분하는 특성을 파악할 수 있으며, 학습된 분류 모델을 이용하여 새로운 글을 자동으로 분류할 수 있다. 본 연구에서는 제시하는 모델을 월마트(Walmart) 관련 최대 포럼인 YahooFinace내에 존재하는 Walmart-forum에 적용하였다.

The Spam Detection Model for Web Forums using Text Mining Techniques

Jiyoung Woo*

ABSTRACT

The spam in the discussion web forum causes user inconvenience and lowers the value of the web forum as the open source of user opinion. The importance of postings is evaluated in terms of the number of involved authors, so the spam distorts the analysis result by adding the unnecessary data in the opinion analysis. We propose the automatic detection model of spam postings in the web forum. We extract text features of posting contents using text mining techniques from the perspective of linguistics and then perform supervised learning to recognize spam from normal postings. Significant features are derived through the learning process and the automatic detection model is built based on those features. To build the automatic detection model of normal postings and spam, four evaluators are asked to recognize the spam posting in prior. We adopted the Naive Bayesian, Support Vector Machine (SVM), decision tree, which are known to perform well in data and text mining tasks. We can extract the text features to recognize the spam and detect automatically the newly posted spam. We apply the proposed model to the YahooFinace-Walmart forum, which is the world largest Walmart-related web forum.

Key Words : Web forum, Social media, Spam, Posting quality, Text mining

* 고려대학교 정보보호대학원(✉jywoo@korea.ac.kr)
· 제1저자(First Author) : 우지영 · 교신저자(Correspondent Author) : 우지영
· 접수일(2012년 1월 17일), 수정일(1차 : 2012년 2월 15일), 게재확정일(2012년 2월 17일)

Ⅰ. 서 론

블로그(blog), 위키(wiki), 웹 포럼(web forum)등의 웹 기반 소셜미디어(social media, 이하 소셜미디어)의 등장은 사람들의 의사소통 현상에 큰 변화를 가져왔다. 소셜미디어는 컨텐츠의 생성비용과 물리적 장벽을 낮춰, 정보의 생산과 소비를 가속화시키고 있다. 사람들은 제공된 컨텐츠만을 받아들이는 것이 아니라, 직접 컨텐츠를 생성하고 이를 웹상에서 공유함으로써, 다른 사람들의 정보 습득 및 사고에 영향을 미치고 있다.

기업, 정부, 기관 등에서도 최근 소셜미디어의 중요성을 인식하고, 전략적으로 활용하고자 하는 움직임을 보이고 있다. 학계에서도 사회과학, 경영학, 경제학, 물리학 등의 다양한 분야에서 소셜미디어에 대한 연구가 새로운 주제로 급부상하고 있다.

웹 포럼상의 특정 주제에 대한 풍부한 정보와 이에 대한 집중적인 논의는 웹 포럼을 사람들의 의견 형성에 가장 중요한 역할을 하는 미디어로 만들고 있다. 하지만, 일부 사용자의 부적절한 사용으로 인해 그 활용가치가 감소하고 있다. 보통 일반 사용자에게 열려있는 웹포럼은 누구나 글을 쓸 수 있기 때문에 주제와 관련없는 광고성 글도 상당수 나타난다. 또한 실질적인 토론이 아닌 비생산적인 서로에 대한 인신공격으로 토론이 장기화되기도 한다. 웹포럼상의 글의 중요도는 일반적으로 해당글에 참여한 사용자 수나 댓글수로 측정하므로, 이러한 불량글은 웹 포럼상의 의견수집 시 많은 불필요한 데이터를 포함하게 되고, 이렇게 분석된 결과는 왜곡된 결과를 내기도 한다. 웹 포럼상의 사용자 의견을 정확히 파악하기 위해서는 불량글을 사전에 제거하는 전처리 과정을 거쳐야 한다. 사용자 입장에서도 불필요한 글들은 웹 포럼 활동을 저해하는 요인으로 작용한다. 웹 포럼 운영 입장에서도 불필요한 글들, 사용자간의 불필요한 다툼 등은 사용자의 이탈로 이어질 수 있기 때문에 해결해야할 문제이다.

본 연구에서는 이러한 문제를 해결하고자, 웹 포럼상의 불량글 탐지 모델을 제시한다. 글의 텍스트 특성을 추출하여 불량글과 일반글을 분류할 수 있는 변수집합을 추출하고 이를 토대로한 자동화된 학습 모델을 구축한다.

Ⅱ. 문헌연구

텍스트마이닝(Text Mining)은 웹상의 글처럼 비정형화된 데이터에 대하여 자연어처리(Natural Language Processing) 기술과 문서처리 기술을 적용하여 유용한 정보를 추출하고 이를 가공하는 것을 목적으로 하는 기술이다.

기존의 설문조사나 전화 조사 등의 방법에서 벗어나 웹 기반의 시장조사(고객 조사 및 여론 조사 포함)에 대한 연구는 이메일을 통한 고객의 응답을 분석하는 것으로 시작된다. Sampson [1]은 웹의 등장으로 온라인상의 사용자 커뮤니티가 시장조사의 정보원으로 사용될 것을 예측하였으며, Gillin [2]은 온라인상에서 고객이 서로에게 미치는 영향 및 이것이 시장 경쟁에 미치는 영향이 점점 더 커지고 있음을 언급하였다. 온라인 상에서의 의견 분석에 대한 연구는 회사 홈페이지나 온라인 상점의 상품평을 기초로 주로 특정 상품에 대한 의견을 분석하고, 이를 상품개발이나 시장전략 수립에 활용하는 연구가 진행되어 왔다. Morinaga [3]은 온라인 서베이를 통해 상품의 인지도를 조사하는 연구를 수행하였다. Liu [4]은 상품의 특성에 대한 온라인상의 고객 평가가 다른 고객의 해당 상품 구매에 미치는 영향을 시각화하는 시스템을 개발하였다. Glance [5]은 웹 블로그나 웹 포럼으로부터 상품평에 대한 정보를 파악하고 분석할 수 있는 모델을 제시하였다.

최근 소셜미디어 중 특히 블로그부터 의견을 파악하고, 영화 박스오피스 순위나 상품 판매량과 같은 기

업의 성과지표와의 관련성을 찾는 연구가 진행되고 있다. 최근 10년간 사용자들이 제품을 직접 평가한 자료를 활용한 연구는 활발히 진행되어 왔다. 하지만 자유 형식으로 작성한 글의 내용을 직접 분석한 연구는 아직 초기 단계이다. Gruhl [6]는 블로그에서 해당상품에 대해 언급한 수와 블로그간의 링크 구조가 해당 상품의 판매 급증과 관련이 있음을 보여주었다. Liu [4]은 PLSA(Probabilistic Latent Semantic Analysis)를 블로그에 맞게 적용한 SPLSA(Sentiment PLSA) 모델을 통해 블로그의 컨텐츠로부터 잠재되어있는 센티멘트를 도출하고, ARSA(Autoregressive Sentiment-Aware) 모델을 통해 상품 판매량과 상품 관련 센티멘트와의 관계를 분석하였다. Wenger [7]는 여행자 블로그로부터 여행지에 대한 의견을 분석하는 시스템을 개발하였다.

소셜미디어 상의 스팸에 대한 기존의 연구는 다음과 같다. Wanas [8]는 포럼의 글을 자동으로 평가하는 모델을 개발하였다. 주제 관련성, 원본여부, 포럼상의 글이라서 갖는 특성들(원글 인용수,댓글수), 글의 표면상 나타나는 특성(글의 길이, 답글에 대한 저자의 반응속도, 글의 포맷 품질), 글 컨텐츠 특성(웹링크수, 질문수)등의 특성을 도출하고, SVM(Support Vector Machine)을 이용하여 자동 평가 모델을 구축하였다. Niu [9]는 웹포럼의 스팸 문제를 검색 엔진의 품질을 떨어뜨리는 요소로 지적하며, 웹포럼의 스팸과 블로그의 스팸을 비교 분석하였다. 더 나아가 웹포럼 스팸을 자동으로 탐지할 수 있는 컨텍스트 기반(redirection과 cloaking) 모델을 제시하였다. Hayati와 Potdar [10]의 연구에서는 Web2.0에서의 스팸 문제를 언급하고, 스팸을 탐지하거나 방지하는 현재의 기법을 비교 분석하였다. Lin [11]은 컨텐츠 특성과 시간의 규칙성을 이용한 스팸 블로그 탐지 모델을 제시하였다. Mishne [12]은 블로그상의 댓글 중 스팸을 탐지하는 모델을 제시하였고, Han [13]은 댓글과 트랙백(trackback)을 포함하여 블로그상에서의 광고성 글을 탐지하는 모델을 제시하였다. Jindal과 Liu [14]는 웹상의 의견, 특히 상품 후기 스팸을 탐지하고자, 리뷰 컨텐츠, 리뷰어 특성 및 상품 특성을 반영한 모델을 제시하였다. Zinmna [15]의 연구에서는 소셜 네트워킹 사이트에서의 스팸 문제를 제기하였다. 불법 사용자의 프로필 조작으로 인해 발생하는 스팸 문제를 탐지하는 모델을 제시하였다. Benevenuto [16]은 Youtube에서의 컨텐츠 스팸 즉, 제목과 다른 광고성 동영상 등, 에 대해 언급하고 이를 추출하는 모델을 제시하였다.

기존의 웹상의 스팸을 탐지하는 연구는 대체로 블로그에 치중되어 있고, 일부 연구가 웹포럼을 다루고 있다. 또한 기존의 웹상의 스팸을 탐지하는 모델 중 일부가 텍스트 특성을 사용하였지만, 글의 길이, 글에 포함된 의문문 수나 링크 수 등 한정된 특성만을 사용하였다. 이에 본 연구는 포럼 상 글의 텍스트적 특성을 다양하게 고려하여 불량글을 탐지할 수 있는 방법에 대해 연구하고자 한다.

III. 불량글 탐지모델

본 연구에서 제시하는 모델은 웹포럼 상의 불량글을 자동으로 분류하는 학습 모델이다. 불량글 탐지를 위해서 텍스트마이닝 기술을 이용하여 글의 특성을 추출하고, 이들 변수를 토대로 불량글과 정상글을 구분하는 분류모델을 구축한다. 불량글이란 광고성글, 악성코드나 불법 사이트로의 링크를 포함한 글, 주제를 이탈한 글, 사용자간의 욕설이나 비방으로 장기화된 토론 등을 말한다.

본 연구에서는 글의 구조적 특성(syntax), 의미적(semantic), 어휘적 특성(lexical)의 변수를 학습모델의 특성으로 사용한다. 이는 언어학에서 글을 분석할 때 일반적으로 사용하는 분석 관점이다. 과거 텍스트 분석에 관련한 연구에서는 의미적 특성 즉, 단어나 단어의 조합을 주로 이용하였다. 본 연구에서는 이외에도

문장의 구조적 특성 및 어휘적 특성을 추가하여 다양한 관점에서의 텍스트 특성을 파악하도록 한다. 본 연구 모형에서는 의미적 특성으로 unigram와 bigram [17]을, 구조적 특성으로 part of speech [18]를, 어휘적 특성으로는 punctuation, line length, contains non-stop words, stemming, rare words [19]을 사용하였다.

표 1. 텍스트 특성 변수
Table 1. Text features

분류	텍스트 변수
구조적(syntax)	part of speech
의미적 (Semantics)	unigram, bigram
어휘적(Lexical)	punctuation, line length, contains non-stop words, stemming rare words

문장 구두점(punctuation)은 문장의 끝에 붙는 부호로 마침표, 쉼표, 의문표, 따옴표 종류에 따라 글을 분류하는 변수이다. Unigram은 하나의 단어를, bigram은 두 개의 연속된 단어의 조합을 의미한다. 예들 들어, 텍스트는 unigram, 텍스트 마이닝은 bigram이다.

Part of speech (POS)는 bigram과 유사하나 문법적으로 구분되는 단어의 조합이라는 점에서 다르다. 이는 글의 구문론적 특성을 나타낸다. 구문론적 구분은 보통 동사, 명사, 형용사, 부사 등이다. 따라서 POS는 동사+부사, 형용사+명사, 명사+명사 등의 조합인지를 나타내는 특성이다. 글의 길이(line length) 또한 글의 깊이와 상세함을 나타내는 평가 기준이다. Contains non-stop word는 제외어를 포함하고 있는지를 나타내는 변수이다. 여기서 제외어란 전치사, 관계대명사 등 의미가 확연히 구분되지 않아 텍스트 분석에서 제외하는 단어를 말한다. 이 변수는 글이 얼마나 많은 정보를 포함하고 있는지를 나타내게 된다. 어근(stemming)은 형태는 다르지만 같은 어근을 갖는 단어가 문장 내 나타나는 빈도를 측정하는 변수이다. Rare words는 글 내에서 반복되지 않는 단어의 비율을 측정한다.

추출한 특성치를 토대로 불량글과 정상글을 구분하는 특성을 추출하도록 분류모델을 구축한다. 제안하는 모델의 전체적인 프로세스는 <그림 1>과 같다.

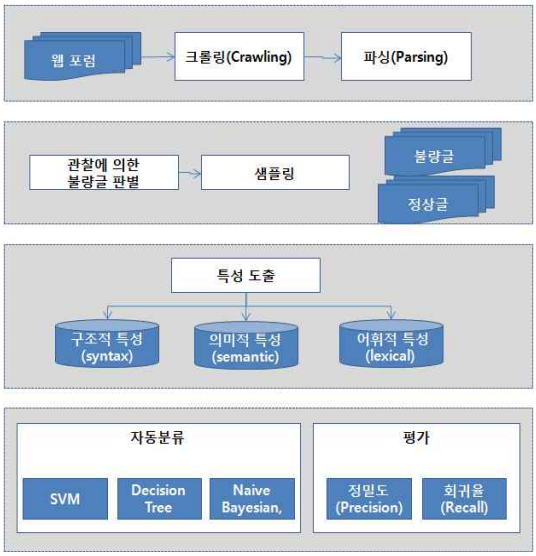

그림 1. 불량글 탐지 모델
Fig. 1. The framework of web-forum spam detection

첫 번째로 Yahoo 사이트의 API(Application Program Interface)를 이용하여 웹포럼의 글을 수집하는 크롤링(crawling) 과정을 거친다. 그 다음은 파싱과정(parsing)으로 Html 형태의 글로부터 필요한 변수를 추출하고 이를 데이터베이스화한다. 불량글을 자동으로 분류하는 모델을 학습시키기 위해서는 불량글과 정상글로 사전에 분류된 학습 데이타가 필요하다. 이를 위해서 평가자가 글의 불량성 여부를 평가하도록 한다. 전체 데이터를 평가할 수 없으므로, 데이터 중 일부를 샘플링하여 이에 대해 평가가 이루어지게 한다. 불량글과 정상글로 분류된 샘플에 대해서 텍스트 특성 변수를 추출한다. 앞서 언급한 대로 제안하는 모델에서 사용하는 변수는 구조적 특성, 의미적 특성, 그리고 어휘적 특성이다.

자동화된 학습 모델을 구축하기 위해 데이터마이닝 및 비정형화된 데이터를 다루는 텍스트마이닝에서 그 성능이 입증된, Naive Bayesian [20], Support Vector Machine [21], 의사결정나무(decision tree) [22]를 이용하였다. Naive Bayesian은 변수간 독립성을 가정한 확률 모형인데, 많은 연구에서 복잡한 시스템에서의 우월한 성능이 입증되었다. 해당 방법의 장점은 적은 수의 데이터로도 학습이 가능하다는 점이다. 이는 변수가 독립성을 가정하여 변수간 공분산 계산이 필요 없기 때문이다. 서포트 벡터 머신(support vector machine)은 비확률적 모형으로 다차원 공간에서 클래스를 구분할 수 있는 다차원 평면을 구성하는 알고리즘이다. 과거 많은 연구에서 클래스 분류 학습에서 성능이 뛰어난 학습 모델로 알려져 있다. 의사결정 나무(decision tree)는 엔트로피를 감소시키는 방향으로 변수의 트리를 구성하는 알고리즘이다. 위의 두 모델과는 달리 모든 변수를 사용하는 것이 아니라 엔트로피 감소량을 기준으로 트리를 구성하는 변수가 선택된다.

본 연구에서는 웹포럼상의 불량글 탐지에 대한 자동 분류 모델의 성능을 비교하고, 가장 좋은 성능을 내는 모델을 실험을 통해 찾아내고자 한다. 제한된 데이터로 학습 및 테스트를 병행하기 위해서 cross-validation 방법을 택한다.

학습된 자동분류 모델의 성능은 정밀도(precision), 회귀율(recall)로 측정한다. 정밀도는 학습모델이 불량글로 분류한 글 중 실제로 불량글로 판명된 글의 비율을 의미한다. 회귀율은 실제 불량글 중 학습모델을 통해 불량글로 판명되는 비율을 나타낸다. 이 두 값은 서로 상충되는 경향이 있어 정밀도 값이 높아지면 회귀율이 떨어지거나 정밀도 값이 낮아지면 회귀율이 높아지는 경향이 있다. 따라서 이 둘의 값을 종합하여 모델의 성능을 비교하는 것이 필요한데, 이들 두 측정치를 비교하기 위해 이를 평균한 F-value 값을 이용한다 [23].

$$Precision = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FP}$$

$$F-value = \frac{2\mathrm{P}recision \times Reacll}{(\mathrm{P}recision + Recall)}$$

TP: $True\ Positive$
(제대로 분류된 불량글)
FP: $False\ Positive$
(잘못 분류된 불량글)
FN: $False\ Negative$
(잘못 분류된 정상글)

(식 1). Precision, Recall, F-value

IV. 실험 결과

본 연구는 월마트 포럼 데이터를 대상으로 진행되었다. 월마트 최대 포럼인 Yahoo! Finance 내의 WalMart Message Board -WMT (http://messages.finance.yahoo.com/mb/WMT/)을 이용하였다. 해당 포럼은 1999년 1월에 생성되어 2008년 6월까지 139,062건의 원글이 포스팅 되었으며, 댓글까지 포함한 글수는 441,954건, 참여한 사용자 수는 25,500명으로 그 크기가 매우 큰 포럼이다. 주로 투자자들이 의사소통을 하는 포럼이지만, 직원이나 소비자도 사용하면서 다양한 주제가 논의되고 있다.

해당 포럼의 2008년 3월 시점의 데이터를 크롤링하여 수집하고, 파싱과정을 거쳐 필요한 변수를 데이터베이스화 하였다. 그 중 232개의 글을 무작위 추출하여 4명의 평가자가 글의 불량 여부를 평가하도록 하였다. 1명의 평가자로부터 불량글로 판정된 글은 전체의 48.3%로 높게 나타났다. 본 연구에서는 2명 이상의 평가자로부터 불량글로 평가를 받은 글을 불량글로, 나머지는 정상글로 분류하였다. 불량글은 전체의 27.6%로 나타났다. 평가자 간의 평가 신뢰도를 측정하기 위해 Kappa [24] 값을 계산한 결과 0.657로 평가자간의 일치도는 신뢰할 만한 수준으로 나타났다.

k=(평가자의 일치도-우연에 의해 평가가 일치할 확률)/

사용자의 글별로 3장에서 언급한 텍스트기반 특성 변수 513개를 도출하였다. 전제 232개를 학습데이타로 하고, 10번의 cross-validation을 실행하였다.

세 개의 자동분류 모델로 학습시킨 결과는 표 2와 같다. 제시하는 모델의 주요 목적은 불량글을 탐지하는 것이므로 불량글 클래스에 대한 정밀도와 회귀율로 모델의 성능을 평가하였다. 정밀도는 SVM의 정밀도가 0.46로 가장 높게 나타났으며, 회귀율은 Naive Bayesian이 높게 나타났다. 회귀율까지 고려한다면 SVM보다는 Naive Bayesian이 높은 성능을 나타냈다. Decision tree의 경우는 다른 두 개의 모델보다 성능이 현저히 떨어지는 것으로 나타났다.

표 2. 불량글 탐지 결과
Table 2. The results of web-forum spam detection

분류	정밀도 (precision)	회귀율(recall)	F-value
Decision tree	0.28	0.14	0.19
Naive Bayesian	0.44	0.71	0.54
SVM	0.46	0.48	0.47

Decision tree의 경우는 -, :, part of speech와 4개의 unigram으로 분류 규칙을 생성하였다. 사용된 part of speech는 전치사_구분자, 기본 동사형_전치사, 과거동사_전치사이고, 사용된 단어는 law, wmt, cost등으로, 총 9개의 변수를 사용하였다. 모든 변수를 다 사용하는 다른 두 개의 학습 모델에 비해 상대적으로 적은 수의 변수를 포함한 학습 모델을 도출하기 때문에 글의 텍스트 특성을 사용한 분류에서는 그 성능이 떨어지는 것으로 추측된다. Naive Bayesian은 변수별 평균값이 높은 순으로 중요변수를 살펴보면 상위 10개의 중요 변수 중 contain_non_stopword, 마침표, 쉼표, 물음표를 제외하고 나머지는 part of speech로 나타났다. 즉, 해당 모델에서는 불량글을 판별하는데 중요한 변수로 어휘와 구조적 특성이 중요한 역할을 하는 것으로 나타났다. SVM 모델은 선형커널 함수를 사용하였다. 선형모델에서 계수값에 따라 정상/불량을 판별하는 중요 변수 중 상위 10개 살펴보면, 7개가 part of speech이고, 나머지 세 개가 unigram으로 나타났다. SVM은 주로 구조적 특성과 일부 의미적 특성을 이용한 모델을 구축하였다.

세 가지의 학습모델로부터 생성된 분류 모델은 part of speech가 가장 중요한 구분자로, 그 다음이 의미를 나타내는 단어, 컨텐츠의 풍부함(contain_non_stopword), 문장 부호 등의 순으로 중요한 변수로 나타났다.

탐지 모델의 정확도를 높이기 위해 변수 추출 단계를 거친후 학습 모델을 구축하였다. 여러 개의 변수 중 글의 판별하는데 결정적인 역할을 하는 변수만을 이용하여 학습 모델을 구축하는 것은 일반적으로 학습 속도를 향상시키고, 모델의 정확도를 높이는 것으로 기존의 연구에서 나타났다. 그러나 본 연구의 실험 결과 Decisiosn tree와 SVM의 경우에는 변수 추출 과정이 모델의 성능에 악영향을 주는 것으로 나타났다. 그에 비해 Naive Bayesian은 그 성능이 약간 향상되었다. 일반적으로 알려져 있는 것에 반해 다수의 텍스트 변수를 사용한 불량글 탐지 모델 수립에서는 그 성능이 향상되지 않는 것으로 나타났다.

표 3. 변수 추출 후 불량글 탐지 결과
Table 3. The detection performance after feature selection

분류	정밀도 (precision)	회귀율(recall)	F-value
Decision tree	0	0	0
Naive Bayesian	0.48	0.65	0.55
SVM	0.44	0.49	0.46

본 연구는 웹포럼 상의 불량글을 자동으로 판별하는 모델을 제시한다는데 의의를 가진다. 웹 상의 스팸 중 특히 블로그 스팸에 대한 연구가 활발히 진행되었

는데 웹 포럼의 스팸에 대한 연구는 부족하였다. 본 연구는 웹 포럼 상의 불량글 탐지에 대한 중요성을 환기하고, 글의 텍스트적 특성을 이용하여 불량글을 탐지할 수 있음을 실험적으로 증명하였다. 또한 텍스트 분석에 효율적인 학습 알고리즘을 찾기 위하여 가장 성능이 좋은 것으로 알려진 세 개의 알고리즘을 비교 분석하였으며, 탐지 모델의 성능을 높이고자 특성 추출(feature selection)을 실시하였으나 성능이 향상되지 않는다는 점을 발견하였다. 본 연구에서 제시하는 모델을 이용해 웹포럼 관리자는 불량글을 자동으로 탐지할 수 있고, 불량글을 자주 포스팅하는 불량 사용자를 효율적으로 관리할 수 있다. 본 연구는 웹 포럼상의 글의 품질을 높이는데 기여할 것이다.

V. 결 론

웹포럼 상의 글의 중요도는 일반적으로 해당글에 참여한 사용자 수나 댓글수로 측정하므로, 이러한 불량글은 웹포럼상의 의견수집 시 많은 불필요한 데이터를 포함하게 하고, 이렇게 분석된 결과는 왜곡된 결과를 내기도 한다. 본 연구에서는 웹포럼상의 불량글을 자동화 분류하는 탐지 모델을 제안한다. 웹포럼상의 글에 대해 텍스트마이닝 기술을 적용하여 글의 텍스트적 특성을 추출하고, 정상과 불량글을 두 개의 클래스로 한 분류모델을 학습시킨다. 이를 통해 불량글과 정상글을 구분하는 특성을 파악할 수 있으며, 학습된 분류모델은 새로운 글을 자동으로 분류할 수 있다. 본 연구에서는 제시하는 모델을 월마트 관련 최대 포럼인 YahooFinace내에 존재하는 Walmart 포럼에 적용하여, 성능을 평가하였다.

현재의 연구에서는 글의 텍스트 특성을 이용하여 불량글을 탐지하는 모델을 제시하였다. 향후 연구에서는 웹 포럼상에서 나타나는 글의 구조를 활용하여 불량글을 탐지하고자 한다. 예를 들어 두 명의 사용자가 같은 글에서 반복적으로 대화를 주고 받는 형태가 나타나는 정도를 변수로 활용할 수 있다. 또한 글 작성자의 특성을 또 다른 변수로 추가할 수 있다. 작성자의 과거 포럼에서의 행동 패턴, 예를 들어 원글 수, 답글 수, 글의 품질 등을 추가적인 변수로 이용하여 모델의 정확도를 높이는 연구를 수행할 계획이다. 변수를 추가하여 모델을 상세화하는 것 이외에도 제시하는 모델을 다른 포럼으로 확대 적용해보는 연구도 진행할 계획이다. 또한 스팸을 세분화하는 방향으로 본 연구를 확장할 수 있다. 현재는 불량글/정상글로만 구분하였는데, 불량글을 광고글, 상대방 인신공격으로 세분화하여 탐지한다면 두 개의 성격이 다른 불량글을 한 번에 탐지하는 모델에 비해 그 성능이 향상될 것으로 기대된다.

참고문헌

[1] Sampson S, “Gathering customer feedback via the Internet: instruments and prospects”, *Industrial Management & Data Systems*. Vol. 98, No. 1-2, pp.71-+,1998.

[2] Gillin P, "T*he New Influencers, A Marketer’s Guide to the New Social Media*", Quill Driver Books\Word Dancer Press, CA, USA, 2007.

[3] Morinaga S, Yamanishi K, Tateishi K, and Fukushima T, "Mining product reputations on the Web", *The eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 341, 2002.

[4] Liu Y, Huang X, An A, and Yu X, "ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs", *SIGIR*, 2007.

[5] Glance N, Hurst M., Nigam K, and Siegler M, "Deriving Marketing Intelligence from Online Discussion", *KDD'05,* Chicago, Illinois, USA, 2005.

[6] Gruhl D, Guha R, Kumar R, and Novak J, "The predictive power of online chatter", *KDD ’05*, pp. 78-87, 2005.

[7] Wenger A, “Analysis of travel bloggers' characteristics and their communication about Austria as a tourism

destination", *Journal of Vacation Marketing*. Vol. 14, No. 169.2008,

[8] Wanas N, El-Saban M., Ashour H, and Ammar W , "Automatic Scoring of Online Discussion Posts", *WICOW'08,* Napa Valley, California, USA, 2008.

[9] Niu Y, Wang Y, Chen H, Ma M, and Hsu F, "A Quantitative Study of Forum Spamming Using Context-based Analysis", *Network & Distributed System Security (NDSS) Symposium*, 2007.

[10] Hayati P and Potdar V , "Toward spam 2.0: An evaluation of Web 2.0 anti-spam methods Industrial Informatics", *INDIN 2009. 7th IEEE International Conference on,* pp. 875-880, Cardiff, Wales, 2009.

[11] Lin Y, Sundaram H, Chi Y, Tatemura J, and Tseng BL, "Splog detection using self-similarity analysis on blog temporal dynamics", *AIRWeb '07 Proceedings of the 3rd international workshop on Adversarial information retrieval on the web,* 2007.

[12] Mishne G, "Blocking Blog Spam with Language Model Disagreement", *AIRWeb*, 2005.

[13] Han S, "Collaborative blog spam filtering using adaptive percolation search", *WWW,* 2006

[14] Jindal N and Liu B, "Opinion Spam and Analysis", *WSDM'08*, Palo Alto, California, USA, 2008.

[15] Zinman A, "Is Britney Spears spam", *Fourth Con-ference on Email and Anti-Spam Mountain View*, 2007.

[16] Benevenuto F, "Identifying Video Spammers in Online Social Networks", *AIRWeb '08,* 2008.

[17] Dunning T, "*Statistical Identification of Language*", New Mexico State University. New Mexico State University. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.1958. Technical Report MCCS pp. 94-273, 1994.

[18] Paul K, "*Analyzing Grammar: An Introduction*", Cambridge: Cambridge University Press. pp. 35, 2005.

[19] Robert F and Lasnik H (eds.), "*Syntax. Critical Concepts in Linguistics*", New York: Routledge, 2006.

[20] Lewis D, "*Naive (Bayes) at forty: The independence assumption in information retrieval*", Machine Learning, pp. 4 – 15, 1998.

[21] Vapnik VN, "*The nature of statistical learning theory*", New York: Springer-Verlag, 1995.

[22] Quinlan JR, "*Induction of decision trees. In Machine Learning*", pp. 81 – 106, 1986.

[23] Buckland M and Gey F, "The relationship between Recall and Precision", *Journal of the American Society for Information Science*, Vol. 45, No. 1, pp. 12 – 19, 1999.

[24] Gwet K, "*Handbook of Inter-Rater Reliability (Second Edition)*" ISBN 978-0-9708062-2-2, 2010.

감사의 글

본 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. [NRF-2009-352-D00329]

저자소개

우지영(Jlyoung Woo)

2000 KAIST 산업공학과 학사
2002 KAIST 산업공학과 석사
2006 KAIST 산업공학과 박사
2006 삼성화재 고객관계관리 부서
2008 미국 아리조나대학 인공지능연구실
2011년~현재 고려대학교 정보보호대학원 연구교수

※ 관심분야 : 온라인 컨텐츠 보안, 소셜미디어 정보확산