The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 16, No. 2, pp.55-62, Apr. 30, 2016. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2016.16.2.55
JIIBC 2016-2-7

기계학습기법을 이용한 광고 외식 블로그의 자동분류

Automatic Classification of Advertising Restaurant Blogs Using Machine Learning Techniques

장재영*, 이병준**, 조세진**, 한다혜**, 이규홍**

Jae-Young Chang*, Byung-Jun Lee, Se-Jin Cho**, Da-Hye Han**, Kyu-Hong Lee****

요 약 최근 들어 블로그가 제공하는 정보를 활용하여 외식업소를 선택하는 사용자가 크게 늘고 있다. 그러나 국내의 외식관련 블로그들은 파워 블로거에 의한 광고 블로그들이 다수를 차지하고 있어 신뢰성을 잃은 지 오래다. 따라서 블로그의 신뢰성을 확보하기 위해서는 허위 또는 과장되게 작성된 광고 블로그들을 필터링하는 기술이 필수적이다. 본 논문에서는 자동분류 기술을 이용하여 광고 블로그들을 판별하는 기법을 제안한다. 제안된 기법에서는 우선 외식 블로그들 중에서 광고 블로그로 판명된 블로그들을 수집하고 이들에 공통적으로 나타나는 특징들을 분석하였다. 이렇게 추출된 특징들을 이용하여 데이터 마이닝의 자동 분류 알고리즘을 적용하여 광고 블로그 여부를 판단하였다. 또한 다양한 실험을 통해 최적의 알고리즘과 특징들을 선별하였다.

Abstract Recently, users choosing a restaurant basedon information provided by blogs are increasing significantly. However, those of most blogs are unreliable since domestic restaurant blogs are occupied by advertising postings written by 'power bloggers'. Thus, in order to ensure the reliability of blogs, it is necessary to filter the advertising blogs which are sometimes false or exaggerated. In this paper, we propose the method of distinguishing the advertising blogs utilizing an automatic classification technique. In the proposed technique, we first manually collected advertising restaurant blogs, and then analyzed features which are commonly found in those blogs. Using the extracted features, we determined whether a given blog is advertising one applying automatic classification algorithms. Additionally, we select the features and the algorithm which guarantee optimal classification performance through comparative experiments.

Key Words : advertising blog, review, filtering, machine learning, classification

Ⅰ. 서 론

Web 2.0 시대의 도래로 인한 SNS의 빠른 확산은 대중들이 사회 전반에 대한 그들의 주관적 의견(subject opinion)들을 피력할 수 있는 다양한 장의 토대가 되고 있다. 특히 포털 사이트에서 제공되는 블로그(blog)는 작성자의 지식이나 경험들을 독자들과 손쉽게 공유할 수 있어 많은 이용자들을 확보하고 있다. 블로그에는 다양한 분야에 대한 지식들이 공유되고 있는데, 그중에서 외식 정보에 관련된 블로그가 큰 비중을 차지하고 있다. 심지어는 외식 블로그를 통한 마케팅 연구도 활발히 진행되고 있는 실정이다[1][2]. 그 만큼 블로그가 외식업소 선택에 있어서 큰 비중을 차지하고 있다.

외식 블로그는 작성자가 외식업소에 직접 방문하여

*정회원, 한성대학교 컴퓨터공학과
**준회원, 한성대학교 컴퓨터공학과
접수일자 : 2016년 2월 18일, 수정완료 : 2016년 3월 18일
게재확정일자 : 2016년 4월 8일

Received: 18 February, 2016 / Revised: 18 March, 2016 / Accepted: 8 April, 2016
*Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

체험한 주관적 혹은 객관적 정보를 독자에게 전달하는 역할을 한다. 독자들은 외식업소를 선택하기 전에 블로그에서 추천하는 업소들에 대한 평가를 참고한다. 그러나 인터넷상에 범람하는 각종 광고 블로그들은 독자에게 객관적 정보를 제공하기 보다는 광고를 의뢰한 업체의 이익을 대변하는 것에 초점을 맞춤으로써 독자들에게 왜곡된 정보를 제공하여 신뢰성을 훼손시키고 있다. 물론 외식업소 정보를 제공하는 블로그도 마케팅을 일종으로 인식되고 있어서, 블로그를 운영하는 포털 사이트에서는 홍보용 블로그임을 명시하는 조건으로 게시를 허용하고 있다. 그러나 대부분의 광고 블로그들은 직접 체험한 리뷰를 가장한 허위 또는 과장된 내용들이다. 따라서 이러한 블로그들을 순수한 리뷰를 작성한 블로그들로부터 필터링하는 것이 사용자들에게 올바른 정보를 제공하기 위한 중요한 요소이다.

광고 블로그 필터링과 유사한 기술로는 스팸메일 필터링(spam mail filtering) 기술이 있다[3][4]. 이 기술은 이미 많은 연구가 이루어져서 대부분의 메일서버에서 활용되고 있다. 하지만 광고 블로그 필터링 기술은 스팸메일 필터링과는 달리 쉽게 구현되기 어려운 기술이다. 그 이유는 스팸 메일의 경우 메일에 포함된 패턴이나 단어 분포만으로도 스팸인지 아닌지가 대부분 쉽게 판별 가능하나, 광고 블로그는 광고가 아닌 것으로 위장하여 작성되므로 필터링 난이도가 매우 높다. 또한 최근에 이에 대한 필요성이 대두됐음에도 불구하고 아직까지 많은 연구가 이루어지지 않고 있는 실정이다[5][6][7].

본 논문에서는 외식 블로그 중에서 순수한 리뷰(비광고)로 위장된 광고 블로그를 필터링하는 기술을 제안한다. 필터링은 데이터 마이닝(data mining) 기술 중 하나인 자동분류(automatic classification) 알고리즘[8]을 활용하였다. 자동분류를 적용하기 위해서는 양질의 학습 데이터가 필수적인데 본 논문에서는 광고 블로그들의 전형적인 특징들을 수작업으로 조사한 후, 이를 이용하여 학습 데이터로 활용할 블로그들을 수집하였다. 또한 광고 블로그의 특징(feature)들을 정량적으로 표현하기 위해 다양한 전처리 기술을 적용하였는데, 적용된 기술로는 불용어(stopword) 제거, 맞춤법 검사, 원형 단어로의 변경 등이 있다. 특히 리뷰 성격을 갖는 블로그의 특성상 게시글에 주관적 표현이 다수 포함되어 있어 감성분석(sentiment analysis)[9] 결과를 추가로 실시하여 그 결과를 광고 블로그를 판별할 특징으로 활용하였다. 광고 블로그들의 특징들은 학습 데이터를 이용한 자동분류 알고리즘에서 활용되는데, 어떠한 특징들을 활용할지를 결정하기 위해 상관분석(correlation analysis)을 실시하여 그 결과를 바탕으로 다양한 특징의 조합을 구성하였다. 이렇게 구성된 특징 조합으로 자동분류 알고리즘을 이용하여 분류 정확도를 실험하였고, 최적의 알고리즘과 특징 조합을 탐색하였다. 본 논문에서는 자동분류를 위해 나이브 베이즈(Naive bayes) 분류 알고리즘과 신경망(neural network) 분류 알고리즘을 활용하였다[8].

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해서 설명하고 3장에서는 데이터 수집 절차를 설명한다. 4장에서는 특징선택 과정과 상관분석 결과를 보여주고, 5장에서는 자동분류 결과를 보여준다. 마지막으로 6장에서는 결론을 맺는다.

II. 관련연구

광고 블로그들은 대부분 순수한 리뷰 블로그로 위장하기 때문에 이를 필터링하는 것은 매우 어려운 문제이다. 몇몇 국내외 사이트[10][11]에서 이러한 광고성 허위 게시글을 판별하기 위한 기술을 사용한다고 알려져 있으나 실제로 어떤 기술을 사용하고 있고 효과가 어느 정도인지는 알려져 있지 않다. 대표적으로 미국의 대표적인 외식정보제공 업체인 Yelp는 광고성 리뷰들을 적절히 필터링하여 외식업체 정보의 신뢰성을 높인 것으로 알려져 있다. 하지만 필터링 방법을 공개할 경우 다시 이를 회피하는 기술이 역으로 사용될 것을 예상하여 비공개로 운영하고 있다.

Yelp의 필터링 기술은 공개되지 않았지만 다행히 2013년에 적용기술을 추정하는 논문이 발표되었다[12]. 이 논문에서는 광고 게시글의 주요 탐지 기술을 크게 언어적 특성(linguistic features)과 행동적 특징(behavioral features)으로 나누어 연구하였다. 논문은 먼저 Yelp에서의 분류 모델이 언어적 특성을 주된 기준으로 사용하는지 알아보기 위해 광고 게시글과 비광고(순수한 리뷰) 게시글의 단어분포를 비교하였다. 그 결과 광고 게시글의 분포와 비광고 게시글의 분포 사이에 큰 차이가 없었다. 따라서 언어적 특성을 이용한 필터링 방식은 Yelp 분류 모델의 주된 방식이라 할 수 없다고 결론을 내렸으며, 광고 게시글을 작성하는 사용자의 행동적 특징이 주된 필

터링 방식이라고 추정하였다. 이 논문에서 추정한 광고 게시글의 행동적 특징은 다음과 같다.

1. 하루에 작성한 게시글 수: 광고 작성자의 25%는 하루 평균 5개, 75%는 6개 이상의 게시글을 작성하며, 그렇지 않은 작성자의 90%는 3개 이하로 작성한다.
2. 긍정 댓글 비율: 광고 작성자의 85%는 작성한 게시글의 80% 이상이 긍정적인 게시글이다.
3. 댓글 글자 수: 광고 작성자의 80%는 평균 댓글 글자 수 135자이며, 그렇지 않은 작성자의 92%는 200자 이상이다.
4. 내용의 유사성: 광고 게시글들은 이전의 게시글들과 내용이 유사하다.

Yelp와 같은 외식정보 사이트에서의 댓글과 같은 게시글은 비교적 글자 수가 많지 않기 때문에 위와 같은 판단 기준은 적절하다고 볼 수 있으며, [12]에서는 이를 실험적으로도 증명하였다. 하지만 본 논문에서는 블로그를 대상으로 하므로 위의 일부 기준들은 적절하지 않다. 예를 들어 블로그는 댓글 형식의 게시글과는 달리 비교적 길기 때문에 하루에 다수의 글을 작성하기 어려우며, 블로그의 특성상 부정적인 글들도 많지 않다. 따라서 블로그를 대상으로 할 경우에는 새로운 특징들을 고려해야한다.

[7]에서는 광고 상품평을 자동분류 기술을 이용하여 필터링하는 방법을 제안했다. 특히 이 논문에서 가장 중점을 둔 것은 학습 데이터를 생성하기 위해 광고 상품평인지를 판단하는 기준을 제시했는데, 이 방법이 중요한 이유는 광고 상품평인지를 수작업으로 판단하는 것은 지극히 주관적이어서 양질의 학습 데이터를 확보하기 매우 어렵기 때문이다.

[5]와 [6]에서는 국내의 온라인 쇼핑몰에서 작성된 한글 광고 상품평들을 필터링하기 위한 방법을 제안하였다. 이 연구들에서는 단어 분포와 같은 언어적 특성을 이용한 문서 분류 방법으로 이용하였고 좋은 성능을 얻었다고 밝혔다. 그러나 학습 데이터를 수집할 때 광고 상품평에 대한 특징들을 미리 규정지어 수작업으로 수집하였고, 그 특징에 따라 분류 모델을 생성했으므로 객관성이 떨어진다고 볼 수 있다. 또한 이 연구들에서도 댓글 형식의 상품평을 대상으로 하였으므로, 본 연구의 대상이 되는 블로그와 차이가 있다.

Ⅲ. 데이터 수집

[7]에서 지적한 바와 같이 리뷰 블로그를 직접 읽고, 이것들이 광고 블로그인지를 그렇지 않은지를 수작업으로 판단하는 것은 매우 어려운 일이다. 다행히 블로그에서는 협찬을 받은 공식적인 광고 블로그의 경우 그 사실을 본문에 명시하도록 되어 있다. 따라서 협찬이 명시된 블로그들을 수작업으로 수집하여 이들의 공통적 특징들을 도출하였다. 그 결과 상당수의 블로그들이 본문에 상호명이 반복되어 언급되었고, '맛집'이란 단어도 자주 언급되었으며, 업소의 상세 '주소'가 포함되어 있었다. 또한 블로그 제목에 '맛집'이란 단어가 포함된 경우도 상당수 존재한다는 사실을 밝혀내었다. 따라서 이러한 사실들에 기초하여 광고 블로그임이 확실시되는 판단 기준을 표 1의 두 번째 열과 같이 정의하였다.

반면에 비광고(순수한 리뷰) 블로그는 전적으로 주관적인 판단으로 수집되었다. 수집 기준의 왜곡을 줄이기 위해서 다수의 실험자들로부터 비광고로 확실하게 판단되는 블로그들을 수집하게 하였다. 이렇게 수집된 비광고 블로그에 대해서 광고 블로그들의 판단기준을 그대로 적용하였고, 그 결과 표 1의 3번째 열과 같은 특징들을 도출할 수 있었다. 즉, 제목에는 대부분 '맛집'이란 단어가 포함되어 있지 않았고, 본문에 '맛집'과 '상호명'이 1회 이하로 언급되었으며, 대부분 상세주소에 대한 언급도 없었다.

표 1. 데이터 수집 기준
Table 1. Data Collection Criteria

수집기준	광고 블로그의 특징	비광고 블로그의 특징
'맛집' 언급[제목]	존재	없음
'맛집' 언급[본문]	3회 이상	1회 이하
'상호명' 언급[본문]	4회 이상	1회 이하
주소' 언급[본문]	존재	없음

학습 데이터로 활용할 블로그 수집을 위해서 주어진 블로그가 외식업소 블로그인지를 판단하는 기준도 필요한데 이를 위해 우선 공공데이터포털(https://data.go.kr)에서 외식업소명들을 확보하였다. 이중에서 무작위로 업소명을 선택하였고, 선택된 업소명과 더불어 '맛집'이란 단어로 블로그를 검색하여 수집하였다. 이러한 방법으로 총 66,000개의 외식업소관련 블로그들을 수집하였다.

다음 단계로 위와 같이 수집된 블로그들 중에서 일부를 학습 데이터로 추출하였는데, 학습 데이터는 수집된 블로그 중에서 표 1을 기준으로 3개 이상 일치하는 것을 광고 블로그와 비광고 블로그로 각각 추출하였다. 이렇게 수집된 학습 데이터는 광고 블로그 1,132개, 비광고 블로그 2.236개로 총 3,368개이다.

Ⅳ. 특징탐색 및 분류

본 장에서는 자동분류 알고리즘의 독립변수를 정의하기 위해 블로그들의 특징들을 탐색하는 과정을 설명한다. 우선 후보 특징들을 정의하고, 이들과 목적변수(광고 혹은 비광고)간의 상관분석을 통하여 상관정도 정도에 따라 특징들을 분류한다.

1. 특징 정의

자동분류 알고리즘을 이용하여 광고 블로그를 추출하기 위해서는 우선 분류에 활용될 특징들을 정의해야한다. 본 논문에서는 광고와 비광고 블로그 사이에 차이가 있을 것으로 예상되는 특징들을 정의하였는데, 크게 블로그 구성에 대한 특징과 감성 표현에 대한 특징으로 나눌 수 있다. 블로그 구성에 대한 특징은 블로그에 나타나는 특정 단어나 표현, 형식들에 관한 것들이며, 감성 표현에 대한 특징은 블로그에 출현하는 감성 단어들의 종류나 분포에 따라 정의된 것들을 의미한다. 우선 블로그의 구성으로 정의된 특징 리스트는 표 2와 같다.

표 2. 블로그 구성에 따라 정의된 특징들
Table 2. Features Defined by Blog Formats

변수	변수 설명
Tmatzip	제목에 '맛집' 키워드 언급 유무
juso	'주소' 키워드 언급 유무
keyword_count	키워드(상호명)의 언급 수
matzip_count	'맛집' 키워드 언급 수
word_count	블로그 본문의 단어의 수
day	블로그 작성 요일
content_length	본문의 길이
map	지도의 유무
image_count	이미지의 수
right	저작권 표시 유무
phone	전화번호 표시 유무

표 2에서 가장 상위 4개의 특징들(Tmatzip, juso, keyword_count, matzip_count)는 표 1의 학습 데이터 수집기준과 동일하다. 이들은 광고, 비광고 블로그를 구별하기 위해 본 논문이 정의한 가장 중요한 기준이므로 당연히 분류모델의 특징들에 포함시켜야한다. 그 이외에도 본문의 단어 수, 블로그 작성 요일, 본문의 길이, 업소 위치를 알려주는 지도포함 여부, 이미지의 수, 저작권 표시 유무, 업소 전화번호 포함 유무 등을 특징으로 정의하였다.

다음으로 감성표현에 따른 특징들을 정의하였다. 외식업소에 대한 블로그는 리뷰형식으로 작성하므로 대부분 감성 표현들이 많이 포함된다. 따라서 광고/비광고 블로그를 분류하는데 있어서 감성표현의 차이가 이들을 분류하는데 중요한 역할을 할 것으로 기대하였다. 블로그에 출현하는 단어들이 긍정 혹은 부정적인 표현인지를 판단하는 감성분석(sentiment analysis)은 오픈한글(http://api.openhangul.com/)의 API를 이용하였다. 여기서는 주어진 단어에 대해서 긍정/부정/중립 단어 중 어디에 속하는지를 판단함과 동시에, 그렇게 될 확률을 정량적으로 알려준다. 확률은 0부터 100까지의 스코어로 나타내며 100에 가까울수록 그럴 확률이 높다는 것을 의미한다. 이를 이용하여 본 논문에서 정의한 감성표현에 따른 특징들은 표 3과 같다.

표 3에서 POScore는 블로그의 내용이 전반적으로 긍정적인지를 판단하는 것으로, 긍정으로 판명된 단어들의 모든 스코어 합이 부정으로 판명된 단어들의 스코어의 합보다 크면 1, 그렇지 않으면 0을 부여한다. NAScore는 그 반대로 점수를 부여한다. 이와 같이 POScore와 NAScore는 블로그의 전반적인 감성에 대한 극성(polrity)을 판단하여 0혹은 1의 값만을 갖는다. 반면에 CNOScore, CPOScore, CNAScore는 각각 중립, 긍정, 부정 단어들에 대한 스코어의 합으로 POScore, NAScore와는 달리 블로그 마다 그 정도에 따라 다양한 정량적인 값을 가질 수 있다.

표 3. 감성표현에 따라 정의된 특징
Table 3. Features Defined by Blog Formats

변수	변수 설명
POScore	긍정 스코어 값이 부정 스코어 값보다 큰 경우 1 작은 경우 0
NAScore	부정 스코어 값이 부정 스코어 값보다 큰 경우 1 작은 경우 0
CNOScore	중립단어에 대한 스코어 합
CPOScore	긍정단어에 대한 스코어 합
CNAScore	부정단어에 대한 스코어 합

2. 상관분석

본 논문에서는 4.1절에서 명시한 특징들이 광고나 비광고 블로그를 분류하는데 어느 정도 관련성이 있는지를 파악하기 위해 상관분석(correlation analysis)을 실시하였다. 분석결과는 자동분류에 이용할 독립변수들의 조합을 결정하기 위해 사용된다. 상관분석 방법으로는 피어슨 상관계수(Pearson Correlation Coefficient)[13]를 이용하였으며, 이를 위해 비광고 블로그는 0, 광고 블로그는 1의 값을 부여하였다. 상관분석 결과는 표 4와 같다. 이 표에서 상관정도는 다음과 같이 상관계수(r)의 값에 따라 결정하였는데, 이 구분은 한글 위키피디아의 상관분석 페이지을 참고하였다.

표 4. 상관분석 결과 및 분류 및 상관 정도에 따른 특징분류
Table 4. Results of Correlation Analysis and Feature Classification by Degrees of Correlation

변수	상관계수	상관정도
Tmatzip	1	SP
juso	1	SP
keyword_count	0.818	SP
matzip_count	0.696	CP
word_count	0.500	CP
day	0.004	IGN
content_length	0.345	CP
map	0.435	CP
image_count	0.209	WP
right	0.134	WP
phone	0.770	SP
POScore	0.203	WP
NAScore	-0.118	WN
CNOScore	0.537	CP
CPOScore	0.479	CP
CNAScore	0.287	WP

- r이 -1.0과 -0.7 사이이면, 강한 음의 상관관계(SN)
- r이 -0.7과 -0.3 사이이면, 뚜렷한 음의 상관관계(CN)
- r이 -0.3과 -0.1 사이이면, 약한 음의 상관관계(WN)
- r이 -0.1과 +0.1 사이이면, 거의 무시될 수 있는 상관관계(IGN)
- r이 +0.1과 +0.3 사이이면, 약한 양의 상관관계(WP)
- r이 +0.3과 +0.7 사이이면, 뚜렷한 양의 상관관계(CP)
- r이 +0.7과 +1.0 사이이면, 강한 양의 상관관계(SP)

이 표에서 보는 바와 같이 학습 데이터 수집기준인 상단의 4개 변수 모두 강한 또는 뚜렷한 양의 상관관계를 보였다. 이들은 광고 블로그와 비광고 블로그를 구분하는 최초의 기준이었으므로 당연한 결과로 보인다. 나머지 특징들을 보면 전화번호의 언급(phone)이 강한 상관관계를 갖는 것으로 나타났으며, 문서길이(content_length), 지도 포함 여부(map), 중립 혹은 긍정 단어에 대한 감성 스코어의 합(CNOScore, CPOScore)도 뚜렷한 상관관계를 갖는 것으로 나타났다. 반면에 블로그의 전체적인 감성에 대해 단순한 극성(0 또는 1)만을 표현한 POScore나 NAScore는 약한 상관관계를 갖는 것으로 나타났다. 이는 외식업소에 대해서 전체적으로 부정적인 감성을 표현하는 블로그는 거의 없기 때문인 것으로 분석된다.

V. 자동분류와 평가

본 논문에서는 비광고 블로그를 추출하기 위한 방법으로 자동분류 알고리즘을 이용하였다. 자동분류에는 다양한 알고리즘이 있으나 본 논문에서는 나이브 베이즈 분류(Naïve Bayes classification)와 신경망(neural network)[8] 알고리즘을 이용하였다. 독립변수들은 표 5와 같이 7가지의 조합으로 나누어서 실시하였다. 여기서 독립변수의 조합은 표 4의 상관정도에 따라서 결정하였다. 표 4를 보면 음의 상관관계를 갖는 변수는 거의 없었다. 따라서 양이나 음의 상관관계를 별도로 고려하지 않고 상관정도만을 고려하였다.

표 5. 독립변수 정의
Table 5. Definition of Independence Variables

독립변수 조합	설명
All	표 4의 모든 변수
Basis	학습 데이터 수집 기준(4개)
C	뚜렷한 상관관계
S	강한 상관관계
W+C	약한 상관관계와 뚜렷한 상관관계
S+C	강한 상관관계와 뚜렷한 상관관계
W+S	약한 상관관계와 강한 상관관계

테스트 데이터로는 표 6과 같이 광고와 비광고 블로그를 각각 200개씩 사용하였다. 특히 광고/비광고 각각에 대해서 학습을 위해 자동으로 수집된 블로그들 중에서 100개씩 무작위로 추출하여 테스트 데이터로 활용하였으며, 이와 별도로 다수의 실험자에 의한 주관적인 판단 하에 수동으로 각각 100개씩 수집하였다.

표 6. 테스트 데이터 구성
Table 6. Test Data Construction

광고	수동	100개
	자동	100개
비광고	수동	100개
	자동	100개

실험 결과는 그림 1과 같다. 우선 나이브 베이즈 분류에서는 S와 W+S에 대한 정확도가 각각 84.9%, 86.6%으로 가장 좋았다. 즉, 강한 상관관계를 갖는 변수가 포함된 경우가 좋은 분류 성능을 보였다. 반면에 W+C의 경우에는 43.7%로 가장 좋지 않은 성능을 보였다. 특히 데이터 수집기준의 조합인 Basis를 포함한 나머지 조합의 경우에는 60%대의 정확도를 보였다. Basis의 경우는 데이터 수집기준으로 학습 데이터를 구성한 가장 중요한 변수들이며, C나 S+C도 높은 상관계수를 보이는 변수들로 조합되었는데, 이들이 약한 상관계수가 포함된 W+S보다도 정확도가 낮은 의외의 결과를 보였다.

신경망 분류는 전반적으로 나이브 베이즈 분류에 비해 상당히 좋은 성능을 보이고 있다. 또한 어떠한 특징으로 조합을 하여도 70%~80%대의 비교적 고른 정확도를 보였다. 특히 Basis로 구축한 모델에서는 90.7%로 가장 높은 정확도를 보였다. 이러한 결과로 볼 때 광고 블로그 추출을 위한 분류 기법에서는 신경망 알고리즘이 나이브 베이즈 분류기법 보다 더 적합 것으로 볼 수 있으며, 강한 또는 뚜렷한 상관관계를 갖는 특징들로 독립변수를 구성하는 것이 더 좋은 성능을 보인다는 것을 확인할 수 있다.

Ⅵ. 결론

본 논문에서는 외식업소에 대한 평가를 가장한 광고 블로그를 필터링하는 방법을 제안하였다. 제안된 방법에서는 자동분류 기법을 활용하였으며, 이를 위해 블로그 수집, 전처리, 상관분석, 특징 선택, 자동분류 및 평가의 과정으로 진행하였다. 광고와 비광고 블로그 수집하기 위해서 광고 블로그들을 수작업으로 선별하였고, 이들의 특징들을 비광고 블로그와 비교하여 추출하였다. 또한 감성분석을 실시하여 그 결과를 특징들에 추가하였다. 결정된 특징들에 대해서는 상관분석을 통해서 광고 블로그의 전형적인 특징들을 정량적으로 분석하였으며. 그 결과를 이용한 다양한 조합으로 자동분류 기법을 적용하였다. 실험 결과 상관계수가 높은 특징들의 조합이 더 높은 분류 정확도를 보였으며, 나이브 베이즈 분류보다 신경망을 이용한 분류 기법이 더 높은 정확도를 보였다. 신경망의 경우 최고 분류정확도가 90%가 넘을 만큼 좋은 성능을 보였다.

[7]과 [12]에서 지적한 대로 광고 블로그는 실제 후기 블로그를 가장하므로 이를 필터링하는 것은 매우 어려운 작업이다. 더구나 실제 블로그를 운용할 때 광고 블로그

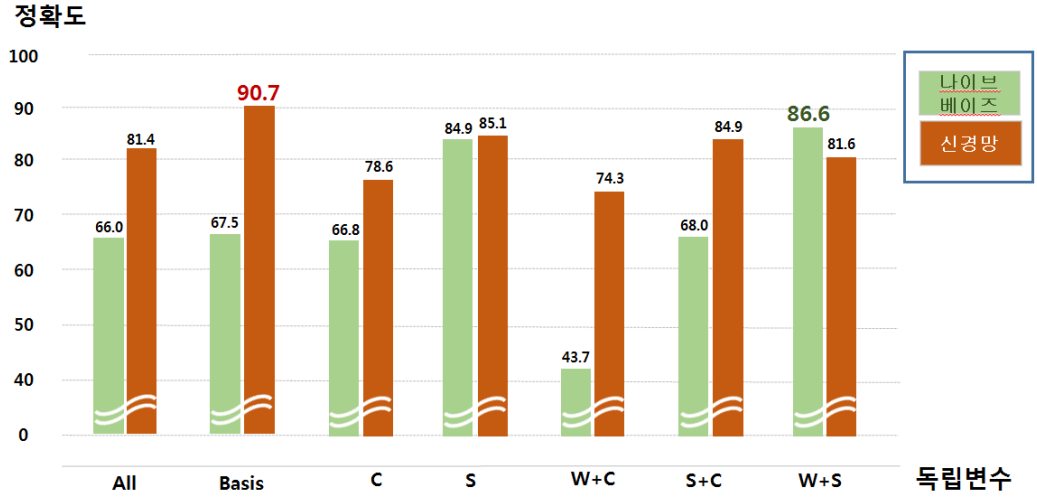


그림 1. 분류 정확도 비교
Fig. 1. Comparison of Classification Accuracy

를 필터링한다는 것이 공개적으로 알려지면 점차 이 기술을 회피하도록 작성할 가능성 있어 더더욱 교묘해질 가능성도 있다. 따라서 실질적인 효과를 발휘하기 위해서는 광고 블로그 필터링 기술은 고정된 특징들이 아닌 상황에 따라 적응 또는 진화되도록 설계되어야 하며, 본 논문의 후속 연구도 이 부분에 초점을 두어 진행할 계획이다.

References

[1] J. Kim and Y. Kim, How the characteristics of the food-blog marketing effect to purchasing intension with the mediation effect of trust, tourism review, Vol. 30, No. 5, pp. 85-105, 2015.
[2] J. Kim, H. Kim, S. Park, Study on Blog users' Response to Blog Marketing, information Systems Review, Vol. 11, No. 3, pp.1-17, 2009.
[3] E. Blanzieri and A. Bryl, A survey of learning-based techniques of email spam filtering, Artificial Intelligence Review, vol. 29, no. 1, pp. 63-92, 2008.
[4] G. Cormack, Email Spam Filtering: A Systematic Review, Foundations and Trends in Information Retrieval, vol. 1, no. 4, pp. 335-455, 2007.
[5] I. Park, H. Kang, S. Yoo, Classification of Advertising Spam Reviews, Proceedings of the 22th Annual Conference on Human and Cognitive Language Technology, 2010.
[6] H. An and B. Park, Extracting similar advertising review for Opinion Mining, IEEK Conference 2014, pp.1593-1596, 2014.
[7] N. Jindal and B. Liu, Opinion Spam and Analysis, Proceedings of WSDM, pp. 219-229, 2008.
[8] I. Oh, Pattern Recognition, KyoboBooks, 2008.
[9] J. Chang, and I. Kim, An Experimental Evaluation of Short Opinion Document Classification Using A Word Pattern Frequency, Journal of the Institute of Internet, Broadcasting and Communication, Vol. 12, No. 5, 2012.
[10] http://www.yelp.com/
[11] http://www.diningcode.com/
[12] A. Mukherjee, V. Venkataraman, B Liu and NS Glance, What Yelp Fake Review Filter Might Be Doing?, Proceedings of International AAAI Conference on Web and Social Media, 2013.
[13] M. Seo. Practical Data Processing and Analysis Using R, GilBut, 2014.
[14] J. Shim, and H. C. Lee, The Development of Automatic Ontology Generation System Using Extended Search Keywords, Journal of the Korea Academia-Industrial cooperation Society, Vol. 11, No. 6, 2009.

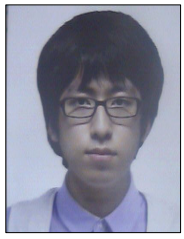
저자 소개

장 재 영(정회원)

- 1992년 : 서울대학교 계산통계학과 (이학사)
- 1994년 : 서울대학교 계산통계학과 (이학석사)
- 1999년 : 서울대학교 계산통계학과 (이학박사)
- 2000년~현재 : 한성대학교 컴퓨터공학과 교수

<주관심분야 : 데이터베이스, 정보검색, 데이터마이닝>

이 병 준(준회원)



- 2016년 : 한성대학교 컴퓨터공학과 (공학사)

<주관심분야 : 빅데이터분석, 데이터마이닝>

조 세 진(준회원)

- 2016년 : 한성대학교 컴퓨터공학과 (공학사)

<주관심분야 : 빅데이터분석, 데이터마이닝>

한 다 혜(준회원)

• 2016년 : 한성대학교 컴퓨터공학과 (공학사)
<주관심분야 : 빅데이터분석, 데이터 마이닝>

이 규 홍(준회원)

• 2016년 : 한성대학교 컴퓨터공학과 (공학사)
<주관심분야 : 빅데이터분석, 데이터 마이닝>

※ 본 연구는 한성대학교 교내학술연구비 지원과제임.