http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.21, No.1, February 2016, pp.15-32
http://dx.doi.org/10.7838/jsebs.2016.21.1.015

온라인 뉴스 제목 분석을 통한 특정 장소 이벤트 성과 예측을 위한 형태소 분석 방법

A Morphological Analysis Method of Predicting Place-Event Performance by Online News Titles

최석재(Sukjae Choi)*, 이재웅(Jaewoong Lee)**, 권오병(Ohbyung Kwon)***

초 록

공개된 데이터인 온라인 뉴스 기사 중 상당수는 도시와 같은 특정 장소에서 발생하는 이벤트에 관련된 사실과 의견을 담고 있어 독자의 의사 결정에 영향을 끼친다. 따라서 대량의 인터넷 뉴스 기사를 분석하면 향후 사람들이 특정 이벤트에 대하여 어떠한 선택을 할지 예상할 수 있을 것이다. 이에 본 연구는 온라인 뉴스 기사 제목을 형태소 분석하여 특정 장소에서 이루어질 이벤트의 성과를 사전에 예측하는 방법을 제안하고자 한다. 기사 제목은 기사의 가장 핵심적인 내용을 담고 있어 본문보다 사실과 의견이 더 정확하게 발현될 뿐 아니라, 모바일 환경에서는 기사 본문보다 더 큰 영향력을 가지기 때문에 이벤트의 성과 예측에 효과적인 자료이다. 이에 인터넷 뉴스 기사의 제목을 수집하여 학습 데이터와 평가 데이터로 구분하고, 학습 데이터에서 유의한 극성을 보이는 형태소를 추출하여 전체 기사의 제목을 감성 분석하였다. 여기에 뉴스 기사가 갖는 특성이 반영될 수 있도록 기사 검색량과 기사 산출량 정보를 변인에 추가하여 이벤트 성과를 예측하는 알고리즘을 수립하였다. 그 결과 70.6%의 성공률로 성과를 예측하여 다른 비교 대상 분석 방법과 분명한 차이를 보였다. 도출된 이벤트 성과 예측 정보는 이벤트를 준비하는 기관 및 업체에서 예상 수요량을 결정할 때 도움을 줄 수 있을 것이다.

ABSTRACT

Online news on the Internet, as published open data, contain facts or opinions about a specific affair and hence influences considerably on the decisions of the general publics who are interested in a particular issue. Therefore, we can predict the people's choices related with the issue by analyzing a large number of related internet news. This study aims to propose a text analysis methodto predict the outcomes of events that take place in a specific place. We used topics of the news articles because the topics contains more essential text than the news articles. Moreover, when it comes to mobile environment, people tend to rely more on the news topics before clicking into the news articles. We collected the titles of news articles and divided them into the learning and evaluation data set. Morphemes

이 논문은2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2014S1A5B8060940).

* Humanitas BigData Research Center, Kyung Hee University(sjchoi@khu.ac.kr)
** Humanitas BigData Research Center, Kyung Hee University(jw_lee@khu.ac.kr)
*** Corresponding Author, School of Management, Kyung Hee University(obkwon@khu.ac.kr)
Received: 2015-12-08, Review completed: 2016-02-06, Accepted: 2016-02-13

are extracted and their polarity values are identified with the learning data. Then we analyzed the sensitivity of the entire articles. As a result, the prediction success rate was 70.6% and it showed a clear difference with other analytical methods to compare. Derived prediction information will be helpful in determining the expected demand of goods when preparing the event.

키워드 : 텍스트 마이닝, 비정형데이터, 장소 마케팅, 장소 이벤트, 예상 수요, 형태소 분석
Text Mining, Unstructured Data, Place Marketing, Place Event, Expected Demand, Morphological Analysis

1. 서 론

온라인에서는 도시나 건물과 같은 특정 장소를 언급한 뉴스 기사가 많이 등장한다. 그리고 기사의 내용으로는 그 장소에서 발생한 사건 사고와 같은 정보 전달적 내용뿐만 아니라, 이벤트(예: 축제, 체육대회) 업체나 주민 등 이해관계자들의 의견, 지역 특색, 동향 등 뉴스를 보다 입체적이고 심층적으로 만들어주는 다양한 요소들이 함께 사용된다. 온라인 뉴스 기사는 누구나 접근 가능하며, 신뢰성을 가지고 있고, 해당 이슈를 신속하게 전하므로 이들이 전하는 장소 및 이벤트에 대한 내용은 곧 지역에 대한 이미지와 브랜드를 형성하게 된다. 따라서 장소에 대한 긍정적인 이미지를 구현하려는 장소 브랜딩에는 필수적이고, 유용한 공유 데이터가 된다[7].

그러나 온라인 뉴스 기사라는 비정형 텍스트에서 기사가 의도하는 긍부정성(stance)의 정확도를 추론하는 것은 쉬운 일이 아니다. 주된 원인은 뉴스 기사가 사건을 심층적으로 다루는 경우에는 특정 사안에 대한 복수의 의견이 실릴 수 있어 텍스트 마이닝 추론 결과가 뉴스가 원래 의도하던 주된 긍부정성과 일치하지 않을 수 있기 때문이다. 즉, 하나의 뉴스 기사는 설득력을 갖추기 위하여 자신의 견해를 뒷받침하는 내용은 물론 다른 견해의 내용도 들어가며, 이해를 돕기 위해 핵심적인 내용이 아닌 주변적인 이야기도 함께 전달되는 것이다. 이는 극성 판단에 혼란을 일으키는 노이즈로 작용하여 문제가 된다. 앞 부분 혹은 마지막 부분에 핵심적인 내용이 담기기도 하지만[37], 모든 기사가 두괄식 혹은 미괄식 기술 방식을 따르지는 않는다. 결국 보통의 기사에는 핵심이 아닌 내용이 너무 많이 들어가 기계적인 분석으로는 정확한 분석이 어려운 것이다.

따라서 본 연구에서는 특정 장소의 이벤트에 대해서 사전에 게재된 비정형데이터로서의 인터넷 뉴스를 분석하여 그 장소에서 벌어질 이벤트의 성과(예: 예상 참가자 수)를 예측하는 방안을 마련하고자 한다. 이를 해결하기 위하여 본 연구에서는 뉴스 기사의 전문을 분석하기보다 기사의 가장 핵심적인 요소인 기사 제목만을 분석하는 접근법을 택하였다. 기사의 제목은 빠른 이해가 가능하도록 제한된 어휘를 사용하여 가장 중요한 정보만을 담는다[28]. 또한 인터넷 환경, 특히 모바일 환경에서는 특정 주제의 기사를 검색하면 작은 화면에 비슷한 논조의 기사 제목이 다수 등장하게 되는데 기사를 읽는 독자들은 본문까지 읽지 않고 제

목만으로 주제에 대한 전반적인 분위기를 짐작하려고 하므로 뉴스 기사에서 제목은 매우 중요하다. 해당 기사의 본문을 읽을 것인지를 상당수의 사람들은 단지 제목만을 보고 결정하기 때문이다[18].

또한 본 연구에서는 기존에 다루어온 동사, 형용사, 명사는 물론, 조사, 어미를 포함한 모든 품사 종류에 대하여 감성 분석을 시도하기 위해 정밀한 형태소 분석을 채택하였다. 수집된 텍스트를 통해서 칭찬, 기대, 의견, 비난, 의심 등과 같은 사람들의 견해를 도출해내기 위해서는 기사가 가지고 있는 주관적인 정보를 판단해야 하는데[31], 주로 감성워드넷(SentiWordNet)과 같은 감성 어휘 사전을 이용한다[4]. 그러나 감성워드넷은 동사와 형용사, 그리고 명사와 같은 실질형태소만을 다루고 있지, 관사나 전치사와 같은 형식형태소는 다루고 있지 않아 긍부정성 추론 정확도 제고에 한계가 있다. 특히 한국어의 경우에는 영어와는 달리 형식형태소가 발달한 언어로서 이들의 쓰임을 고려하지 않으면 의미를 제대로 파악해낼 수 없다. 예를 들어, '그는 꽃을 좋아한다'는 긍정적인 문장이지만 형식형태소만 바꾼 '그가 꽃을 좋아할까'는 부정적인 문장이 되며, '일반인에게도 공개하다'는 긍정적인 문장이지만, '일반인에게만 공개하다니'는 부정적인 문장이 된다. 이렇게 한국어 문장의 경우에는 실질형태소 뒤에 붙은 적은 양의 형식형태소가 전체 극성을 바꿀 수 있다. 따라서 본 연구에서는 정확한 분석을 위해 그 동안 감성워드넷 등에서 관심을 갖지 않았던 조사, 어미와 같은 형식형태소에 대해서도 감성 분석 대상에 포함하였다. 제안한 방법론의 성능 검증을 위해 2009년부터 2014년까지의 실제 지역 축제 이벤트 성과 자료를 수집하여 실증하였다.

2. 문헌 연구: 감성 단어의 수집과 분석

인터넷의 발달로 뉴스 기사는 플랫폼 자체에 변화가 찾아왔다. 기존의 종이 신문은 구독자가 감소하여 영향력이 줄어들었고, 대신 인터넷으로 제공되는 뉴스가 SNS의 전달 능력과 맞물려 파급 효과가 커졌다. 특히 스마트폰 보급의 증가로 인해 모바일 인터넷 사용자가 급증하였는데, 이미 스마트폰 사용자의 절반 이상은 모바일 기기를 이용하여 뉴스 기사에 접근하며, PC보다는 모바일 인터넷으로 뉴스에 접근하는 이용자가 많아졌다[28].

하지만 뉴스는 특정 이슈에 대해서 사실 위주로 단순하게 전달하는 경우도 있지만, 긍정적이거나 부정적인 관점을 가지고 기술하기도 한다[13, 40]. 따라서 많은 뉴스 기사는 이용자가 이슈를 긍정적 또는 부정적으로 판단할 수 있게 하는 많은 요소들을 가지고 있는데 가장 기본적인 요소는 감성 속성을 가진 단어 즉, 감성 단어이다. 기사에 감성 단어를 적절히 배치하여 해당 이슈를 기자가 원하는 관점으로 보게 하는 것이다. 따라서 감성분석을 실시하면 기사 내의 감성 단어가 추출될 것이 기대된다.

하지만 감성 분석이 탐색하려는 감성 및 감성 단어는 매우 추상적인 영역에 속하므로 정의하기가 어렵고[8], 연구에 사용할 수 있는 객관적인 목록도 확보하기가 어렵다. 영어권의 경우 2,800개 정도의 감성 어휘 목록을 제시한 연구도 있으나[3], 주관적 판단에 의하여 이루어져서 감성 단어로 보기 어려운 것이 상당히 존재하였다. 하지만 감성 단어 판정 기준을 너무 엄밀히 하다 보면 실용적으로 쓸 수 있는 충분한 양의 목록을 얻기 어렵다는 문제가 발생한다. 한편 보다 객관적이고 문헌에 기초하여 엄밀하게 감성 단어

목록을 만든 경우엔 최종 수집된 감성 단어의 수가 130개 정도에 불과하여 활용도 측면에서 많은 문제를 갖게 되었다[34]. 이에 자연언어처리(Natural Language Processing)를 기반으로 감성 단어를 대량으로 확보할 수 있는 알고리즘들이 개발되었는데[14] 초기의 방법은 주관적 선택에 의한 경우가 많아 기초 단어(seed words)가 너무 많이 선택된다는 문제가 있었다. 이에 Turney[35]은 기초단어 선정을 매우 제한하여 맥락과 상관없이 항상 일정한 극성을 보여주는 긍정 단어 7개와 부정 단어 7개만을 결정하고, 다른 단어들이 이 두 그룹의 단어 중 어떤 단어들과 더 많이 어울리는지를 보고 극성과 강도를 결정하였다. 비슷하게 Kamps[16]는 감성 단어를 단 두 개만 이용하되, 워드넷(WordNet)을 사용하여 감성 단어의 목록을 확장하고 강도를 결정하였다. 긍정과 부정 기초 어휘로 가장 대표되는 것 한 가지씩만을 두고(good, bad), 워드넷에서 동의어 관계로 있는 것이 이들과 어느 정도의 거리에 있는지를 측정하여 반비례 관계로 극성의 강도를 결정한 것이다. 또한 문형을 이용하여 단어의 극성을 파악한 경우도 있었는데[25, 36], 이들은 같은 수식어를 사용하는 단어는 같은 극성을 가질 것이라는 전제 하에 기초 감성 단어로부터 감성 단어의 목록을 확장하는 등 문장의 구조적 특징을 감성 분류에 활용하였다. 이렇게 파악된 감성 단어의 목록은 이제 다음 단계의 세부 문제를 평가하는 데 이용되고 있다[25].

이어 단어와 문장 차원을 넘어 문서 차원의 감성 분류도 시도되었다. 문서의 분류에는 기존에 사용되어 왔던 기계학습 기법 즉, 나이브 베이지안(Naïve Bayes), 최대 엔트로피(Maximum Entropy), 지지벡터기계(Support Vector Machines)와 같은 방법들을 문서의 감성 분류에 활용하였고, 그 결과를 서로 비교하였다[26, 12]. 이러한 알고리즘들은 오랜 기간 다양한 분야에 사용되어 기본적인 분류 알고리즘으로 자리를 잡고 있지만, 문제는 이들을 이용한 문서의 감성 분류 정확도가 대체로 70~80% 정도에 머무른다는 점이다.이는 단어만을 이용한 문서의 감성 분석은 한계가 있다는 것을 보여주는 것으로서, 기존에 다루어지지 않은 문서의 다른 부분에까지 분석 대상을 확장할 필요를 느끼게 한다. 이와 관련하여 Read[30]는 이모티콘을 이용하여 영화 리뷰의 극성을 판단하는 연구를 진행하였는데 비록 이 방법은 뉴스 기사와 같은 정형화된 문체에는 적용되기 어렵지만, 일반 단어가 아니라 기호도 감성 단어의 한 종류로 간주할 수 있음을 보여주는 사례로서 분석 대상을 확장시켜 준 면에서 의미가 있다.

위와 같은 연구들의 성과를 기반으로 주가 변동률을 뉴스 기사의 분석으로 예측하는 연구들이 나타나기 시작했다. 뉴스와 주가 변동성 사이에 상관 관계가 있음을 입증하는 연구로 시작하여[25], 뉴스와 주가 사이의 시계열 상관관계를 규명하는 연구[19], 그리고 이를 기반으로 뉴스 기사로 실제의 미국 증시를 예측하는 시스템 개발 연구[33] 등이 진행되었고, 국내에서도 뉴스 기사를 이용하여 국내 증시 예측 시스템의 모형 연구[1, 39]와 환율 정보 예측 시스템 연구[28]가 이루어졌다. 문서 분류와 감성 분류의 연구 성과를 실제 응용 단계로 이끌어냈을 뿐만 아니라, 뉴스 기사의 분석으로 특정 대상에 대한 가치 판단 예측이 가능하다는 것을 보여주었다는 점에서 큰 의미가 있다.

영어권에서는 위와 같이 감성 분류에 관한 연구가 비교적 활발히 이루어졌으나, 다른 언

어권에서는 연구가 아직 미진한 상태에 있다[5]. 따라서 다른 언어권에서는 영어의 감성 분석 성과를 이용하기 위하여 자동번역기의 번역 결과를 이용하거나[15] 병렬 코퍼스를 활용하고는 하는데[38], 이는 감성 분석과는 직접 상관이 없는 번역 층위를 여러 단계 거친 것으로서 해당 언어에서의 감성 분석 정확도는 떨어질 수밖에 없다.

이처럼 감성 단어의 수집과 감성 문서의 분석에는 언어적 특징과 문서의 특징을 고려해야 함에도 불구하고 기존의 연구는 이를 충분히 반영하지 못했다. 본 논문에서는 이러한 점을 보완 및 반영하여 연구를 진행한다.

3. 뉴스 제목의 긍부정성을 활용한 특정 이벤트 성과 예측

3.1 전체 진행 구조

<Figure 1>은 본 논문에서 제안하는 뉴스 기사 제목을 활용한 이벤트 성과 예측 방안의 전체 진행 구조이다. 먼저 사전에 웹 상의 오픈 데이터인 특정 장소의 이벤트 관련 뉴스 기사의 제목을 크롤링한다. 그리고 수집된 기사 제목에 대하여 복수의 전문가가 그 제목의 전체적인 긍부정성을 파악하였다. 이때 전원 일치하는 기사 제목만을 채택하여 일종의 학습데이터를 확보했다. 그 후 자체적으로 개발한 형태소 분석기로 기사 제목에 대해 실질형태소와 형식형태소를 정교하게 추출 및 분석한다. 이를 통해 기사를 각 단어 혹은 어근별로 분리하며 동일한 단어에 대해서도 품사 분류를 통해 단어의 종류를 정확히 식별할 수 있다. 그리고 기사별로 추출된 형태소가 전체 문장의 긍부정성에 미치는 영향 유무를 파악하며, 그 결과인 긍정성에 주로 영향을 미치는 형태소와 부정성에 영향을 미치는 형태소를 파악하여 데이터 셋에 저장한다. 다음으로는 이벤트 시작 직전에 게시된 뉴스를 수집하고, 확보된 긍부정성에 영향을 미치는 형태소 데이터 셋을 이용하여 기사들의 전반적인 긍부정성을 판단한다. 한편 이벤트에 대한 기사량의 증감분은 뉴스 기사가 노출되는 대표적 포털 사이트에서 파악한다. 본 연구에서는 네이버 트렌드를 이용하였다. 마지막으로 각 이벤트마다 누적 긍부정성의 증감율, 긍부정 사건의 유무, 그리고 이벤트 검색량의 증감율 정보를 토대로 다가오는 이벤트에 참가할 예상 고객수의 증감에 대한 정보를 예측한다.

〈Figure 1〉 Event Performance Prediction Process based on News Title

3.2 분석 대상 형태소

뉴스 제목의 극성은 <Table 1>과 같이 단어의 극성, 어미의 극성 그리고 단어 또는 어미와

〈Table 1〉 Source of Polarity of News Title

Polarity	Explanation	Example
Polarity of Words	Polarity of words themselves	Example of Positive Words: gaetong(opening), youmyung(famous), chinhwangyung (environment-friendly), choigo(best), sungjang(growth), salrangsalrang(smoothly), myomi(charm), harmony Example of Negative Words: doongap(transforming appearance), smulsmul(gradually), aggapge(regrettably)
Polarity of Ending	Ending polarity provides the whole word's polarity by combining its stem. It amplifies the polarity or gives the opposite polarity.	~hane, ~ine, ~seyo, ~yeora, ~haeyo, ~ja
Polarity of Punctuation marks	Punctuation marks can give the polarity by combining words or ending.	choigo!(the best!), asiwo…(regrettably…), ~ilgga?(is it?)

문장부호 결합의 극성 등 세 가지로 이루어진다. 극성의 발견을 위해서는 형태소 분석이 필요하다. 뉴스 제목에 포함되어 있는 형태소에 대한 각각의 극성이 결정되면 이것을 토대로 해당 뉴스 기사에 대한 긍부정성을 계산할 수 있다. 극성이 1과 0, 그리고 -1의 값을 가진다고 할 때, 긍부정성은 이들의 가중합이므로 임의의 실수가 될 수 있다.

3.3 극성 형태소의 추출

학습 데이터 셋은 다음과 같은 과정을 거쳐 구축되었다. 첫째, 먼저 온라인 인터넷 뉴스와 같은 공공 자료로부터 사례를 수집한다. 그리고 수집되는 n개의 사례 중에서 m개의 사례를 선택한다(단, $n > m$).

둘째, 선택된 사례에서 긍정적인 기사의 제목과 부정적인 기사의 제목을 구분하여 별개의 파일에 저장한다. 극성 판단은 복수의 연구원이 하여 공통된 것을 우선으로 하고, 서로 합의가 되지 않는 기사 제목은 사례 목록에서 제외한다.

셋째, 긍정적인 기사와 부정적인 기사로 분리된 파일에 대하여 형태소 분석을 실시하여 나타난 형태소의 그룹별 등장 빈도를 % 단위로 구한다. 각 그룹에 유의하게 많이 출현하는 형태소를 편향(skewed) 형태소, 한쪽 그룹에만 등장하면 완전편향(purely skewed) 형태소, 양쪽 다 등장하나 특히 어떤 한쪽에 많이 등장하면 부분편향(partially skewed) 형태소, 양쪽에 고루 등장하는 형태소는 혼재(confused) 형태소라고 한다. 만약 부분편향 형태소이거나 혼재 형태소로 나타나면 형태소 그룹 단위를 고려한다. 형태소 그룹이란 하나 이상의 형태소로 이루어진 것을 말한다. 편향 형태소인 경우에는 그 하나의 형태소가 형태소 그룹을 이루게 된다.

이상과 같이 하여 긍정적 형태소 그룹과 부정적 형태소 그룹을 인식한 후 학습 데이터 셋에 저장한다. 이번 실험에서는 완전편향 형태소를 구하여 사용하였다. 완전편향 형태소는 긍정과 부정 각각에 대하여 형태소, 빈도, 품사 정보를 주어 저장하였다.

3.4 이벤트 성과 예측 알고리즘 변수

뉴스 기사는 일반 문서와는 다른 특징이 있으므로 이것이 예측 알고리즘에 반영될 수 있는 변수를 설정한다. 첫째, 개최 직전 홍보성 기사의 증가는 예상 관람객 수의 하락 변수로 본다. 기사 중 홍보성 기사는 기사의 공급원이 이벤트를 개최하는 기관이다. 따라서 개최 직전 긍정적 홍보성 기사가 증가하는 것은 오히려 각 기관이 예상 관람객 수가 감소할 것을 우려하여 적극적인 홍보 활동을 한 결과로 볼 수 있으므로 이를 예상 관람객 수의 하락 변수로 볼 수 있다.

둘째, 반대로 개최 직전 긍정적인 홍보성 기사의 감소는 예상 관람객 수 증가의 결정요인으로 본다. 개최를 앞두고도 홍보성 기사가 감소한다는 것은 관람객 수의 감소를 우려하지 않는 것으로 해석할 수 있기 때문이다.

셋째, 통제 가능한 홍보성 기사와는 달리 환경적으로 발생하여 사회에 큰 영향을 미친 긍부정적 사건의 발생은 예상 관람객 수의 증가에 정비례하게 영향을 미치는 것으로 보았다. 예를 들어 세월호 사건과 같이 사회 전반에 부정적 영향을 끼치는 사건이 일정 기간 이내에 발생하면 관광사업은 크게 위축을 받는다. 따라서 이를 예상 관람객 수의 하락을 가져오는 변수로 보았다. 본 연구에서는 이 일정 기간을 한 달로 보았다.

넷째, 개최 직전 특정 이벤트에 대한 기사 및 문서의 검색량은 예상 관람객 수의 증감에 정비례하게 영향을 미치는 것으로 보았다. 검색량은 대중의 관심을 반영하기 때문이다.

위의 네 가지 요인이 동일한 결과를 예측하지 않을 때는 다음의 조정 규칙을 통해 해결한다.

[규칙 1] 위의 네 요인 중에서 이벤트에 영향을 미칠 긍부정적 사건의 요인을 가장 우선시한다.

[규칙 2] 특별한 긍부정적 사건이 존재하지 않는 경우 연도 t의 이벤트 개최 직전 기사 제목의 (1)긍부정성 증가율 ($\Delta n_t = \frac{n_t - n_{t-1}}{n_{t-1}}$)과 (2)이벤트 검색량의 증가율($\Delta q_t = \frac{q_t - q_{t-1}}{q_{t-1}}$)의 가중 평균과 특정 threshold인 θ과의 관계로 다음 식과 같은 판정을 한다.

(3) $\Delta\epsilon_t = \alpha\Delta n_t + (1-\alpha)\Delta q_t =$

$$\begin{cases} \Delta\epsilon_t \geq \theta, & \text{예상 관람객 수 증가} \\ -\theta < \Delta\epsilon_t < \theta, & \text{정체} \\ \Delta\epsilon_t \leq -\theta, & \text{예상 관람객 수} \end{cases}$$

이때 실수 $\alpha(0 \leq \alpha \leq 1)$는 기사 제목의 긍부정성을 중시하는 상대적 가중치이며, θ도 0과 1사이의 임의의 실수이다.

4. 실험

4.1 성과 자료 수집

본 논문에서 제안한 방법의 성능을 분석하기 위하여 고양국제꽃박람회, 담양대나무축제, 보령머드축제, 부여서동연꽃축제, 포항불빛축제의 다섯 가지 축제를 실험 대상 이벤트로 선정하였다. 위 다섯 가지 축제는 주요 포털 사이트에서 '축제'로 검색했을 때 가장 많이 등장하는 것일 뿐만 아니라, 각 지방 자치단체에 요청한 성과 자료의 내용이 비교적 충실하였기 때문이다. 후보로 예비 선정된 다른 축제들의 경우에는 보내온 성과 자료의 내용 중 상당 부분이

〈Table 2〉 Official Number of Visitors of Festival

Place	Year	Period	Number of visitors	Days	Average number of visitors per day
Goyang	2009	4.23~5.10	514,745	18	28,596.9
Goyang	2012	4.26~5.13	548,539	18	30,474.4
Goyang	2013	4.27~5.12	553,912	16	34,619.5
Goyang	2014	4.25~5.11	451,002	17	26,529.5
Damyang	2011	5.03~5.08	511,500	6	85,250.0
Damyang	2012	5.01~5.06	325,850	6	54,308.3
Damyang	2013	5.03~5.08	315,250	6	52,541.7
Damyang	2014	6.27~6.30	209,000	4	52,250.0
Boryung	2010	7.17~7.25	2,680,000	9	297,777.8
Boryung	2011	7.16~7.24	2,250,000	9	250,000
Boryung	2012	7.14~7.24	3,084,000	11	280,363.6
Boryung	2013	7.19~7.28	3,171,000	10	317,100.0
Boryung	2014	7.18~7.27	3,299,000	10	329,900.0
Buyeo	2011	7.21~7.24	100,000	4	25,000.0
Buyeo	2012	7.26~7.29	120,000	4	30,000.0
Buyeo	2013	7.18~7.21	180,000	4	45,000.0
Buyeo	2014	7.17~7.20	250,000	4	62,500.0
Pohang	2010	7.23~7.26	1,040,000	4	260,000.0
Pohang	2011	7.28~7.31	1,120,000	4	280,000.0
Pohang	2012	7.27~8.05	1,530,000	10	153,000.0
Pohang	2013	7.26~8.04	1,880,000	10	188,000.0
Pohang	2014	7.31~8.03	750,000	4	187,500.0

결여되어 있어 자료로 활용하기가 어려웠다. 성과 자료의 수집은 '정부3.0 대한민국정보공개'(https://www.open.go.kr/) 사이트를 통해 각 지방 자치단체에 성과자료를 요청하였다. 수집된 성과자료는 다음과 같이 정리되었다.

4.2 기사 및 기사 제목의 수집

각 지역별 축제를 보도하는 기사의 수집은 네이버의 기사 검색을 활용하였다. 네이버는 국내에서 점유율이 가장 높은 포털 사이트일 뿐만 아니라, 다른 사이트에서 검색되는 기사는 대부분 네이버 기사 검색에서도 나타난다.

검색어는 축제의 핵심어와 지역명을 조합하여 만들고, 검색 기간은 각 지자체에서 보내 준 성과 자료의 해당 년도 축제가 시작하기 전날까지로 하여 최신 순으로 수집하였다. 예를 들어 2014년 7월 18일 개막한 '보령머드축제'의 경우, 검색어는 '머드 보령'으로, 검색 기간은 2014년 4월 24일까지로서 24일에 가까운 최신 순으로 수집하였다.

수집된 네이버 기사의 파일은 HTML 형식이므로 태그 정보를 활용해 기사 제목을 추출하였다. HTML 태그인 "txt_inline" 내의 "title" 태그에서 기사 제목을 찾고, 이와 함께 "yyyy.mm.dd." 형식으로 된 날짜 정보도 함께 추출하여 제목의

기간별 정리가 가능하게 하였다. 추출된 정보는 CSV 파일에 “도시이름, 기사제목, 날짜”의 형태로 저장하였다.

이 과정을 통해 3,070개의 기사 제목이 수집되었고, 이중 20%에 해당하는 614개의 뉴스 기사 제목을 임의 추출하여 학습용 데이터로 정하였다. 이 학습용 데이터에 대하여 세 사람의 연구원이 기사 제목만으로 긍정, 부정, 중립의 극성을 판단하게 하였고, 그 결과 73.8%의 일치도를 보였다. 불일치한 26.2%의 기사에 대해서는 세 연구원에게 보인 후 필요하면 자신의 결정을 변경할 수 있도록 했다. 이렇게 두 차례에 걸친 coding의 결과 98.7%의 일치도(614건 중 606건 일치)로 합의할 수 있었다. 그리고 끝까지 일치하지 않은 기사 제목에 대해서는 다수결의 원칙으로 처리했고, 완전히 split이 난 경우에는 사례에서 제거하였다. 그 결과로 뉴스 기사의 긍부정성을 계산하였고, 이를 각 지자체, 각 연도별로 합산하였다.

4.3 형태소 분석 및 극성 파악

최종 학습용 데이터는 본 연구팀에서 자체 개발한 한글 형태소 분석기 RHINO 2.0을 이용하여 문장부호를 포함한 모든 품사에 대하여 형태소 분석하였다. 그 결과 극성 형태소가 <Figure 2>와 같은 방식으로 정리되었다.

<Figure 2>에서 보면, 긍정 극성을 보이는 형태소에는 일반명사 ‘꽃, 홍보, 대나무, 개최’와 감탄의 의미를 전달할 수 있는 느낌표(!), 부사격조사 ‘로’, 보조사 ‘요’ 등이 사용되었고, 부정 극성을 보이는 형태소에는 ‘취소, 세월호, 애도, 참사’ 등이 사용된 것을 볼 수 있다. 형태소의 극성 강도는 이들의 개별 빈도를 총빈도로 나누어 결정한다.

4.4 기사 검색량 확보

실험에 사용된 축제 관련 데이터는 정부 혹은 지자체의 공식적인 자료 협조가 이루어진 고양의 꽃 축제(2012~2014년), 담양의 대나무축제(2012~2014년), 보령의 머드축제(2011~2014년), 부여의 연꽃축제(2012~2014년), 그리고 포항의 불빛축제(2011~2014년)관련 성과 자료이다. 또한 특정 축제가 언급된 기사 및 문서의 검색량은 네이버 트렌드(http://trend.naver.com)를 통하여 확보하였다. 네이버 트렌드는 특정 검색어가 특정 기간 동안 얼마나 검색되었는지를 그래프와 수치로 제시해 준다. 검색 기간은 축제가 시작되기 직전 일까지로 하였다.

Polarity	Morpheme	Frequency	POS	Polarity	Morpheme	Frequency	POS
Positive	kkot(flower)	35	Noun	Negative	cuiso(cancel)	5	Noun
Positive	hongbo(advertising)	32	Noun	Negative	sewolho(ship's name)	3	Pronoun
Positive	daenamu(bamboo)	28	Noun	Negative	kyunggi(game)	2	Noun
Positive	!	25	Punctuation	Negative	aedo(sympathy)	2	Noun
Positive	ro(by)	25	Postpostion	Negative	iyong(use)	2	Noun
Positive	bulbit(light)	24	Noun	Negative	junmyun(entire)	2	Noun
Positive	seo(from)	24	Postpostion	Negative	chamsa(disaster)	2	Noun
Positive	yo(honorific)	24	Postpostion	Negative	MB	1	Pronoun
Positive	kwa(with)	22	Postpostion	Negative	STOP	1	Noun
Positive	pakramhoi(exhibition	21	Noun	Negative	ganhaeng(enforcing)	1	Noun
Positive	euro(by)	20	Noun	Negative	gulin(beggar)	1	Noun
Positive	gaechoi(opening)	20	Noun	Negative	gulyok(indignity)	1	Noun

〈Figure 2〉 Example of Polarity Morphs

4.5 비교 평가

본 제안 방법을 평가하기 위하여 다음과 같은 방법들과 비교 평가하였다.

방법 1: 기사의 긍부정성의 증감(Δn_t) 여부로 하루 평균 관람객 수의 증감 여부를 예측하는 방법

방법 2: 네이버 트렌드 상의 검색량 증감(Δq_t)여부로 하루 평균 관람객 수의 증감 여부를 예측하는 방법

방법 3: 본 연구에서 제안한 방법으로서 $\Delta \epsilon_t$과 환경적 부정기사의 존재 유무로 하루 평균 관람객 수의 증감 여부를 예측하는 방법

성과 측정치는 증감 예측 정확도(%)로서 전체 연도 중 예측이 적중한 연도의 수의 비율이다. 먼저 선정된 각 지자체 축제별로 뉴스 제목의 긍부정 점수와 실제 이벤트에 참여한 고객의 수를 비교한 결과는 <Figure 3>과 같다. 그림 전체적으로 보면 대체로 긍부정 점수와 실제 이벤트 참여 실적 사이에 상관관계가 있는 것으로 보이지만, 이를 더욱 정확히 분석할 필요가 있었다.

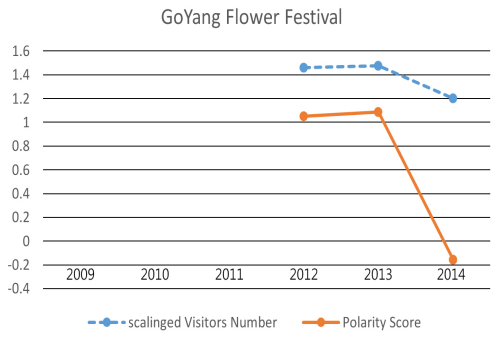
이에 다음 비교하는 세 가지 방법으로 특정 지자체의 특정 연도 축제의 하루 평균 관람객 수의 증감을 예측한 결과 <Table 3>과 같은 결과를 얻었다. <Table 3>은 실험에 사용된 실례가 고양, 담양, 보령, 부여, 포항 등에서 개최된 다섯 축제임을 보여주고 있으며, 자료의

〈Table 3〉 Result of Experiment

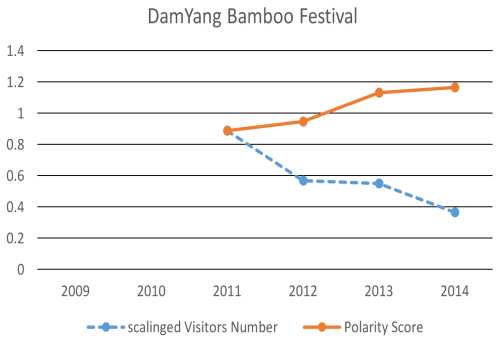
Year	City	Increasing rate of search volume (YoY)	Increasing rate of polarity (YoY)	Increasing rate of average daily visitors (YoY)	Method 1	Result 1	Method 2	Result 2	Method 3	Result 3
2012	Goyang	-0.778	0.363	-0.454	-1	Match	-1	Match	-1	Match
2013	Goyang	-0.490	0.064	-0.363	0	Mismatch	-1	Match	-1	Match
2014	Goyang	-0.128	-1.141	-0.234	1	Mismatch	-1	Match	-1	Match
2012	Damyang	-0.021	0.273	-0.160	-1	Match	0	Mismatch	-1	Match
2013	Damyang	-0.235	0.195	-0.033	-1	Mismatch	-1	Mismatch	-1	Mismatch
2014	Damyang	-0.128	0.029	-0.006	0	Match	-1	Mismatch	0	Match
2011	Boryung	0.075	-1.144	-0.003	1	Mismatch	0	Match	1	Mismatch
2012	Boryung	0.121	-0.078	0.022	0	Match	1	Mismatch	0	Match
2013	Boryung	-0.320	0.094	0.040	0	Match	-1	Mismatch	-1	Mismatch
2014	Boryung	0.080	0.351	0.077	-1	Mismatch	0	Match	-1	Mismatch
2012	Buyeo	-0.174	-0.036	0.121	0	Mismatch	-1	Mismatch	0	Mismatch
2013	Buyeo	1.632	-0.114	0.131	1	Match	1	Match	1	Match
2014	Buyeo	1.600	0.037	0.136	0	Mismatch	1	Match	1	Match
2011	Pohang	1.000	0.370	0.200	-1	Mismatch	1	Match	1	Match
2012	Pohang	14.500	-0.280	0.229	1	Match	1	Match	1	Match
2013	Pohang	4.882	-0.192	0.389	1	Match	1	Match	1	Match
2014	Pohang	1.833	0.037	0.500	0	Mismatch	1	Match	1	Match
				Prediction Accuracy		0.471		0.647		0.706

확보 정도에 따라 2011년 또는 2012년부터 2014년까지의 실제 데이터로 검증한 것이다. 이때 네이버 트렌드로 본 전년대비 뉴스 검색량 증감과 뉴스제목에서 나타난 글의 극성값의 전년대비 증감 정보를 가지고 방법 1~방법 3에 의하여 전년대비 1일 평균 관광객 수 증감에 대해 예측한 결과가 실제 관광객 증감과 일치하는 지를 보여준 표이다. 단, 전년 대비 이벤트 시작 하루 전 뉴스 기사의 긍부정성이 증감율과 이벤트 참여 관객 수의 증감률을 반비례 관계로 본 이유는, 보통 하루 전 홍보성 기사는 이벤트 참여자의 수가 많지 않을 것으로 예상될 때 더 많이 올라가고 그 기사 내용도 홍보성이므로 긍정적 기사를 올리기 때문이다. 즉, 하루 전날 이벤트에 대한 사실적 기사보다 긍정성을 담은 홍보성 기사는 그 지자체가 전년대비 이벤트 참여인원의 감소를 예상하고 있음을 나타내는 증거이다.

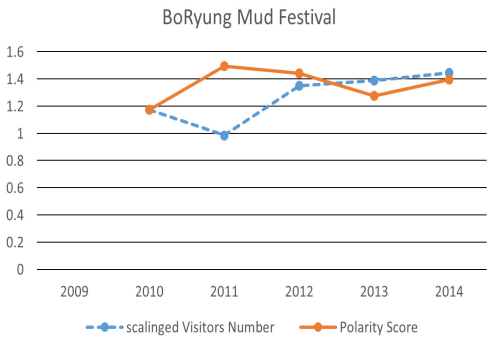
<Table 3>을 근거로 볼 때 본 논문에서 제안하는 방법인 방법 3이 정확도 측면에서 다른 두 가지 방법보다 더욱 우수한 것으로 나타났다. 단, $\theta = 0.1$로 정했기 때문에 하루 관람객 증가율이 -0.100부터 0.100까지의 값을 가질 경우에는 증감이 없는 0으로 판정하였다. 그리고 0.100 이상과 -0.10 이하의 경우에는 각각 증가와 감소라고 판정하였다. 따라서 무작위로 예측을 할 경우 기대되는 예측 정확도는 33.3%가 된다. 그 결과 방법 1의 경우에는 47.1%, 방법 2의 경우에는 64.7%의 예측 정확도를 보였는데, 제안 방법인 방법 3의 경우에는 70.6%의 예측 정확도를 보였다. 이는 무작위 예측 정확도가 33.3%임을 고려할 때 두 배 이상의 판정 정확도이며 상대적으로도 높은 수준의 예측을 한 것으로 보인다.



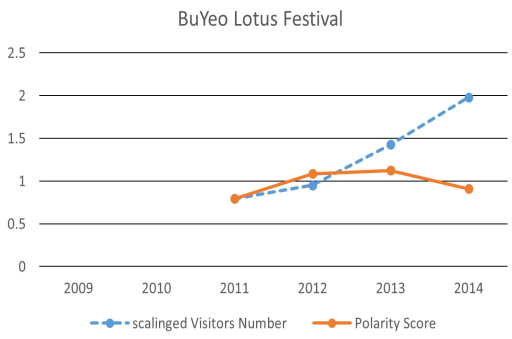
〈Figure 3〉 Comparison of Visitors' Numbers and Polarity Score of Goyang Festival



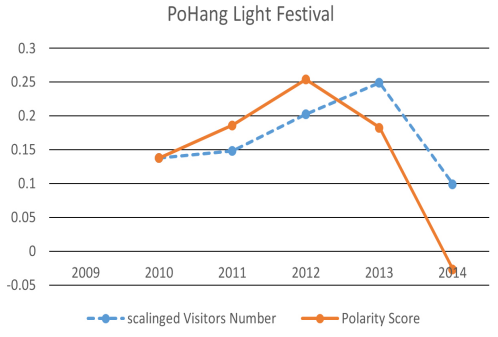
〈Figure 4〉 Comparison of Visitors' Numbers and Polarity Score of DamYang Festival



〈Figure 5〉 Comparison of Visitors' Numbers and Polarity Score of Boryung Festival



〈Figure 6〉 Comparison of Visitors' Numbers and Polarity Score of Buyeo Festival



〈Figure 7〉 Comparison of Visitors' Numbers and Polarity Score of Pohang Festival

5. 토의 및 결론

5.1 시사점

본 연구는 다음과 같은 시사점을 갖는다. 첫째, 본고에서 제시한 방법은 인터넷 상에 올려진 데이터를 분석하여 특정 대상에 대한 선호도를 파악해낸 것으로서 즉응적이며 비용효율적인 분석 방법이라는 점이다. 기존에는 브랜드 이미지나 이벤트의 효과를 측정하려면 설문조사와 같은 인적 노력에 의한 별도의 조사를 실시해야 했다[25]. 그러나 이는 조사에 많은 시간을 필요로 하므로 원하는 시점마다 즉응적으로 결과를 파악할 수 없다. 더구나 장소 이벤트의 경우에는 특성상 전국적인 조사를 실시하여야 하므로 많은 비용이 들어 자주 사용되지 못한다. 반면 본 연구가 제시한 방법은 저비용, 적시적일뿐만 아니라, 객관적인 데이터에 의한 자동 분석이므로 높은 신뢰도를 갖는다.

둘째, 영향력이 크면서 예측력이 있는 빅데이터 자원을 발굴하여 실제 응용 분야에 적용했다는 점이다. 뉴스 기사는 작성자의 특정 대상에 대한 선호도가 반영된 자료로서 사회적 선호도를 파악할 수 있는 가치 있는 공개된 자료이다. 신문 기사는 오피니언 리더로서 일반 대중에게 미치는 영향력이 크며[6], 특히 많은 뉴스가 비슷한 논조로 특정 대상을 언급할 때 일반 대중은 많은 영향을 받는다. 따라서 뉴스 기사가 특정 대상에 대하여 갖고 있는 일반적인 경향을 파악한다는 것은 대중의 의사 결정의 움직임을 예측할 수 있다는 의미를 갖는다. 그리고 일반적인 빅데이터 자원인 소셜미디어는 대부분 이벤트가 발생한 다음에 산출되는 반면, 뉴스 기사는 이벤트 발생 전에도 자주 작성된다는 점에서도 뉴스 기사의 자료는 더욱 가치가 있다. 일반 대중은 정보가 제한적이기 때문에 이벤트 발생 전에 글을 쓰기는 어렵고, 경험하고 난 뒤에야 리뷰 형식의 글을 작성하는 것이 대부분이다. 하지만 언론 매체는 보다 많은 정보 자원을 갖고 있어서 일어날 이벤트에 대한 예상이 가능하다. 본 연구는 뉴스 기사의 영향력과 예측력을 인지하고 활용한 점에서 좋은 사례가 되어준다.

셋째, 데이터의 성격을 고려한 데이터 마이닝이라는 점이다. 기존의 빅데이터 연구에서는 데이터의 성격을 고려하지 않고 데이터 내에 존재

하는 형태소의 빈도를 단순 계산하는 방식을 많이 사용하였다. 그러나 본 연구에서는 뉴스라는 데이터의 성격을 고려하여 이에 영향을 줄 수 있는 변인들을 고려해 이벤트를 예측하였다. 인터넷 시대로 접어들며 특정 이벤트의 홍보를 원하는 조직은 이벤트를 앞두어 놓고 속보성 기사 형태로 관련 자료를 다수의 언론사에 이메일로 발송한다[9]. 따라서 이벤트를 며칠 앞두어 두고서는 그에 대한 긍정적인 홍보성 기사가 급증할 수밖에 없다. 그러나 해당 이벤트의 인지도가 이미 높아 성공에 대한 자신감이 있다면 홍보성 기사의 양은 그렇게 높아지지 않을 것이다. 본 연구에서는 개최 직전 홍보성 기사의 증감을 예상 관람객 수에 반비례하는 결정요인으로 본 반면, 홍보성 기사와 무관한 사회적으로 발생한 부정적 사건은 예상 관람객 수에 비례하게 부정적으로 영향을 미치는 결정요인으로 보았다. 이와 같은 뉴스 기사의 특징을 고려한 본 연구의 예측 함수는 긍정적 기사의 양을 단순히 긍정적으로 보는 단순 예측 함수에 비하여 높은 성과를 내었다. 본 연구는 데이터 마이닝을 통한 성과 예측 시에는 데이터의 긍부정성의 증감을 데이터의 성격에 비추어 해석해야 한다는 시사점을 제시한다.

넷째, 한국어에 맞는 긍부정 단어의 목록을 확보하는 새로운 방법을 제시하였다는 점이다. 제2장에서 제시된 전산적 방법에 의한 감성 단어 추출 방법은 영어에만 적용되는 것이어서 조사와 어미가 발달한 한국어의 경우에는 적용하기가 어렵다. 본고에서는 문서를 모든 품사의 형태소 단위로 분리하여 긍정적인 문서에서 배타적으로 사용되는 형태소와 부정적인 문서에서 배타적으로 사용되는 형태소를 추출하여 극성 형태소 목록을 수집하였다. 기존 연구에서는 명사와 동사 등 실질 어휘만을 대상으로 긍부정 어휘 목록을 구축한 것에 비하여 본 연구에서는 형식 형태소에 대해서도 목록 가능성을 살핌으로써 한국어의 특성에 맞는 긍부정 단어의 목록을 확보하였다. 영어의 경우에도 전치사, 접속사, 관사 등의 형식형태소가 적지 않은 만큼 적용이 가능할 것이다.

다섯째, 그 동안 관심을 갖지 않았던 언어 단위의 가치를 발견했다는 점이다. 그 동안 자연언어처리에서는 형식형태소와 아울러 '., !, ?'와 같은 문장부호는 stopwords로 처리하고 분석 단계에서 아예 배제하였다[32, 10]. 그러나 뉴스 기사와 같이 특정 대상을 심층 분석하거나 홍보하는 문서에서는 감탄의 의미를 갖는 언어 단위는 그것이 문장부호라고 하더라도 대개 긍정적인 의미를 갖고, 의문의 의미를 갖는 언어 단위는 대개 부정적인 의미를 갖는다. 어떤 면에서는 이들은 문장 전체의 의미를 가늠하게 하는 주요 기능어(function words)일 수도 있다. 본고는 형태소 분석 단계에서 그 동안 관심을 받지 못하던 언어 단위까지 포괄하여 분석함으로써 감성 분석은 물론, 자연언어처리 수준에서의 분석 범위를 확대하였다.

5.2 결 론

본 연구는 데이터 마이닝 기법을 이용하여 이벤트의 성과를 예측하는 방안을 제안하였다. 제안한 방안은 저비용이면서 객관적이고 적시적인 정보를 줄 수 있다는 점에서 성과 예측을 위하여 많은 비용을 들이기 어려운 지방 자치단체 및 관련 기관에 큰 도움을 줄 수 있을 것으로 생각한다. 특히 본 연구의 접근 방법은 IT 정보 기술이 문화 산업 융성에 기여를 할 수

있다는 것을 보여주었다. 해당 주제에 대하여 많은 속성 정보를 가지고 있는 뉴스 기사를 활용함으로써 지역의 문화적 특징을 잘 포착하고, 경제 발전에 기여할 수 있는 정보를 제공할 수 있도록 하였다.

본 연구의 성과는 향후 지역의 특징을 찾아 그 아이템을 중심으로 각종 이벤트를 발굴하고, 지역을 브랜딩하는 데도 사용될 수 있다. 예를 들어, '보령'은 머드를 지역 특산품으로 내세울 수 있었는데 단순히 머드팩과 같은 화장품을 파는 데 그치지 않고, 머드로 체험할 수 있는 각종 상품을 개발하고, 이를 한 자리에 모아 축제 수준으로 끌어올렸다. 그리고 웰빙과 체험이라는 두 키워드로 관광객들을 불러들였고, 지금은 지역 경제를 활성화시키는 데 큰 몫을 담당하고 있다. 보령은 문화적 자산을 새로 만들어 낸 것이 아니라, 기존의 자산에 시의적절한 문화적 코드를 덧입혀 큰 경제적 효과를 창출한 것이다. 지역의 문화적 특징과 보편적 문화적 코드는 뉴스 기사에 잘 드러난다. 본 연구에서 사용한 기법을 발전시키면 미처 인식되지 못한 중요한 문화적 자산을 발굴하고, 이를 소비자가 원하는 형태로 발전시킬 수 있을 것이다.

하지만 본 연구의 결과를 더욱 정밀하게 하고 확장하기 위해서는 향후 다음과 같은 부분에 대한 연구가 필요하다. 먼저, 긍부정 목록을 더욱 많이 확보하는 방안이다. 현재 알려진 감성 단어는 그 종류가 많지 않아 대량의 문서에 대하여 정밀한 작업을 하기가 어렵다. 기계적이고 자동화된 방법으로 대량의 감성 단어를 추출할 필요가 있다. 앞으로 점수가 부여된 영화 리뷰와 같이 긍부정성이 명확한 문서를 대량으로 수집하여 감성 단어를 확보하는 방안에 대하여 연구하고자 한다.

둘째, 본 연구에서는 문장부호까지도 긍부정 목록에 포함시켰으나 이들의 사용에 제한을 둘 필요가 있다. 긍부정성 판단에 문장부호까지 두는 것은 판별 상황을 크게 확대하고, 적절한 상황에서 사용되면 기존의 방법에 비하여 정확성을 크게 높일 가능성이 있다. 하지만 문장부호는 기본적으로 긍부정성이 없는 것이므로 어떠한 상황에서 사용되는 것이 적절한지 그 조건을 다각도로 연구할 필요가 있다.

셋째, 지역 뉴스를 분석하여 잠재적 가능성이 있는 이벤트 아이템을 발굴하는 방안이다. 지역간 뉴스를 비교하여 해당 지역에만 특징적으로 나타나는 키워드를 분석하는 방법[21]을 통해 그 지역의 차별화된 강점을 찾고, 이를 시대적 문화 코드에 접목시켜 경제적 효과를 낳을 수 있는 방안을 마련할 필요가 있다. 객관적 사실을 시의적절하게 제공하는 뉴스의 특성을 이용하면 해당 지역만이 갖는 특징을 판별하여 새로운 브랜딩 및 이벤트 발굴이 가능할 것이다. 보다 나아가서는 차별화된 특성을 필요로 하는 연예, 엔터테인먼트 산업 전반에도 활용할 수 있을 것이다.

본 연구는 이벤트 시작 직전 일의 기사를 바탕으로 이루어졌다. 의미 있는 결과를 얻을 수 있었으나, 차후 전 기간에 걸친 이벤트 관련 기사의 수집을 통한 분석이 이루어진다면 추가적으로 시사성 있는 결과를 확보할 것으로 보인다.

References

[1] Ahn, S. and Cho, S., "Stock Prediction Using News Text Mining and Time Series

Analysis," Korea Computer Congress, Vol. 37, No. 1, pp. 364-369, 2010.

[2] Ahn, S. H., Lee, S. H., and Kwon, O. S., "Activation Dimension: A Mirage in the Affective Space?," Korean Psychology Association, Vol. 7, No. 1, pp. 107-123, 1993.

[3] Allport, G. W. and Odbert, H. S., "Trait-names: A psycho-lexical study," Psychological Monographs, Vol. 47, No. 1, 1936.

[4] Baccianella, S., Esuli, A., and Sebastiani, F., "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," In Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10), pp. 2200-2204, 2010.

[5] Bautin, M., Vijayarenu L., and Skiena, S., "International Sentiment Analysis for News and Blogs," ICWSM, 2008.

[6] Entman, R. M., "How the Media Affect What People Think: An Information Processing Approach," The Journal of Politics, Vol. 51, No. 2, pp. 347-370, 1989.

[7] Falkheimer, J., "When Place Images Collides: Place Branding and News Journalism," InGeographies of Communication: the Spatial Turn in Media Studies, Nordicom, 2006.

[8] Fehr, B. and Russell, J. A., "Concept of emotion viewed from a prototype perspective," Journal of experimental psychology: General, Vol. 113, No. 3, pp. 464-486, 1984.

[9] Fenton, N., "New Media, Old News, Journalism and Democracy in the Digital Age," English language edition Published by SAGE Publications, 2009.

[10] Fox, C., "A Stop List for General Text," SIGIR forum, Vol. 24, No. 1-2, pp. 19-35, 1990.

[11] Gim, E., "A Study on the Korean Emotion Verbs," Ph.D. Thesis, Chonnam National University, 2004.

[12] Go, A., Huang, L., and Bhayani, R., "Twitter sentiment analysis," Entropy, p. 17, 2009.

[13] Godbole, N., Srinivasaiah, M., and Skiena, S., "Large-Scale Sentiment Analysis for News and Blogs," ICWSM, pp. 7-21, 2007.

[14] Hatzivassiloglou, V. and McKeown, K. R., "Predicting the semantic orientation of adjectives," Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics, Association for Computational Linguistics, pp. 174-181, 1997.

[15] Hiroshi, K., Tetsuya, N., and Hideo, W., "Deeper sentiment analysis using machine translation technology," Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, p. 494, 2004.

[16] Kamps, J., Marx, M., Mokken, R. J., and Rijke, M. De., "Using WordNet to Measure Semantic Orientations of Adjectives,"

LREC, Vol. 4, pp. 1115-1118, 2004.

[17] Kanhabua, N., Balnco, R., Matthews, M., "Ranking related news predictions," SIGIR 2011 Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 755-764, 2011.

[18] Korea Press Foundation, "2014 Media Audience Opinion Survey," 19th User Behavior Survey of the media environment changes 2014-5, 2014.

[19] Lee, G., "Economic News and Stock Market Correlation: A Study of the UK Market," Conference on Terminology and Knowledge Engineering, 2002.

[20] Lee, S. J. and Kim, H. J., "Keyword Extraction from News Corpus using Modified TF-IDF," The Journal of Society for e-Business Studies, Vol. 14, No. 4, pp. 59-73.

[21] Lee, W. and Lim, N., "A Study on the Elements of City Brand Image and Influences," Journal of Korea Planners Association, Vol. 40, No. 6, pp. 177-192, 2005.

[22] Leon, J. A., "The effects of headlines and summaries on news comprehension and recall," Reading and Writing: An Interdisciplinary Journal, Vol. 9, pp. 85-106, 1997.

[23] Liu, B., Hu, M., and Cheng, J., "Opinion observer: analyzing and comparing opinions on the web," Proceedings of the 14th international conference on World Wide Web, ACM, 2005.

[24] Mitchell, M. L. and Mulherin, J. H., "The impact of public information on the stock market," Journal of Finance, pp. 923-950, 1994.

[25] Nasukawa, T. and Yi, J., "Sentiment analysis: Capturing favorability using natural language processing," Proceedings of the 2nd international conference on Knowledge capture, ACM, pp. 70-77, 2003.

[26] Pang, B., Lee, L., and Vaithyanathan, S., "Thumbs up?: sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp. 79-86, 2002.

[27] Park, I. J., "The analysis of Korean affective terms: listing affective terms and exploring dimensions in the affective terms," Master thesis, Seoul National University, 2001.

[28] Peramunetilleke, D. and Wong, R. K., "Currency Exchange Rate Forecasting from News Headlines," ADC '02 Proceedings of the 13th Australasian database conference, Vol. 5, pp. 131-139, 2002.

[29] Pew Research Center, The State of the News Media 2012: An Annual Report on American Journalism, Retrieved from http://www.journalism.org/2012/10/01/future-mobile-news/>, 2012.

[30] Read, J., "Using emoticons to reduce dependency in machine learning techniques

for sentiment classification," Proceedings of the ACL student research workshop, Association for Computational Linguistics, pp. 43-48, 2005.

[31] Riloff, E., Wiebe, J., and Wilson, T., "Learning subjective nouns using extraction pattern bootstrapping," Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Association for Computational Linguistics, Vol. 4, pp. 25-32, 2003.

[32] Salton, G., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of," Reading: Addison-Wesley, 1989.

[33] Schumaker, R. P. and Chen, H., "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," ACM Transactions on Information Systems(TOIS), Vol. 27, No. 2, p. 12, 2009.

[34] Shaver, P., Schwartz, J., Kirson, D., and O'connor, C., "Emotion knowledge: further exploration of a prototype approach," Journal of personality and social psychology, Vol. 52, No. 6, pp. 1061-1086, 1987.

[35] Turney, P. D. and Littman, M. L., "Measuring praise and criticism: Inference of semantic orientation from association," ACM Transactions on Information Systems (TOIS), Vol. 21, No. 4, pp. 315-346, 2003.

[36] Wilson, T., Wiebe, J., and Hoffmann, P., "Recognizing contextual polarity in phrase-level sentiment analysis," Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, pp. 347-354, 2005.

[37] Yang, C., Lin, K. H. Y., and Chen, H. H., "Emotion classification using web blog corpora," Web Intelligence, IEEE/WIC/ACM International Conference on, IEEE, pp. 275-278, 2007.

[38] Yao, J., Wu, G., Liu J., and Zheng, Y., "Using bilingual lexicon to judge sentiment orientation of Chinese words," Computer and Information Technology, 2006. CIT '06. The Sixth IEEE International Conference on, IEEE, p. 38, 2006.

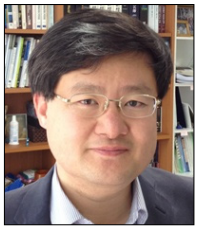
[39] Yu, E., Kim, Y., Kim, N., Jeong, S. R., "Predictiong the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," Journal of Intelligence and Information Systems, Vol. 19, No. 1, pp. 95-110, 2013.

[40] Yu, H. and Hatzivassiloglou, V., "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics, pp. 129-136, 2003.

저 자 소 개

최석재 (E-mail: sjchoi@khu.ac.kr)
1999년 고려대학교 국어국문학과 (학사)
2001년 고려대학교 대학원 국어국문학과 (석사)
2008년 고려대학교 대학원 국어국문학과 (박사)
2001년~2002년 카네기멜론 대학 전산학부 방문연구원
2003년~2005년 연변과학기술대학 언어공학연구소 실장
2008년~2010년 고려대학교 BK21 연구교수
2011년~2014년 성신여자대학교 국어국문학과 초빙교수
2014년~현재 경희대학교 경영대학 학술연구교수
관심분야 한국어 정보화, 빅데이터분석, 감성분석

이재웅 (E-mail: jw_lee@khu.ac.kr)
2015년 한성대학교 경제학과 (경제학사)
2015년~현재 경희대학교 경영학과 석사과정
관심분야 데이터마이닝, WOM, 감성분석, 재무분석



권오병 (E-mail: obkwon@khu.ac.kr)
1988년 서울대학교 경영학과 (경영학사)
1990년 한국과학기술원 경영과학과 (공학석사)
1995년 한국과학기술원 경영과학과 (공학박사)
2001년~2002년 카네기멜론대학 전산학부 방문연구원
2009년~2011년 샌디에고주립대학 경영정보학과방문교수
2004년~현재 경희대학교 경영대학 교수
관심분야 빅데이터분석, 유비쿼터스 컴퓨팅, 의사결정지원시스템