ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 June: 22(2): 97~113 http://dx.doi.org/10.13088/jiis.2016.22.2.097

영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석*

이상훈
한양대학교 자연과학대학 수학과
(*ilazzurro@naver.com*)

최정
한양대학교 일반대학원 경영학과
(*choij127@naver.com*)

김종우
한양대학교 경영대학 경영학부
(*kjw@hanyang.ac.kr*)

인터넷상의 데이터가 급속하게 증가함에 따라 막대한 양의 데이터를 목적에 맞게 적절히 활용하는 빅데이터 분석이 활발하게 진행되고 있다. 최근에는 기존의 정형 데이터분석이 가진 한계점을 보완하는 방법으로 비정형 데이터 분석 분야 중 하나인 텍스트마이닝 기법에 대한 연구들이 다수 이루어지고 있으며, 특히 텍스트를 기반으로 문장의 긍정, 부정을 판별하고 분류하는 감성분석과 관련된 연구들이 활발하게 이루어지고 있다. 이러한 연구의 연장선 상에서, 본 연구는 감성분석에 사용되는 감성사전을 데이터의 특성에 맞게 적절하게 변형하여 구축하는 방법을 시도하였다. 데이터가 속한 영역의 특성을 고려하지 않은 기존의 범용 감성사전을 감성분석에 사용할 경우, 해당 영역에서 쓰이는 단어 또는 감정 표현을 반영하지 못하므로 감성분석의 정확성이 떨어질 수 있다. 따라서 감성분석에 있어서 영역 맞춤형 감성사전의 사용 시 데이터 영역의 특성을 정확하게 반영해 분석의 정확성을 높여줄 것으로 기대할 수 있다. 본 연구에서는 영화 리뷰 데이터를 분석 대상으로 선정하였으며, 대표적 영화정보 사이트 IMDb에서 발생된 약 2년간의 영화리뷰 데이터를 수집·분석하였다. 분석에 앞서 영화 장르별 사용되는 단어의 의미가 각각 다를 것을 고려하여 영화를 '액션', '애니메이션', '코메디', '드라마', '공포', '과학공상' 6개 장르로 분류했다. 맞춤형 감성사전 구축을 위한 핵심 기법으로 SO-PMI(Semantic Orientation from Point-wise Mutual Information)를 활용하였으며, 어휘 간 극성이 뚜렷하게 구분되는 형용사에 한정하여 연구를 진행했다. 분석결과 맞춤형사전을 활용한 감성분석 예측정확도는 영화 장르별로 상이했다. '애니메이션'을 제외한 5개 장르에서 기존의 범용 감성사전대비 맞춤형 감성사전의 예측정확도가 통계적으로 유의한 수준의 성능 향상을 보였다. 본 연구에서는 데이터 영역의 특성에 맞는 맞춤형 사전 구축을 통한 감성분석의 예측의 성능 향상을 확인하였다. 향후 감성사전 구축 시 동사, 부사 등 다양한 품사의 어휘를 추가하여 감성분석 예측정확도를 높이는 방안을 모색할 수 있을 것이다.

주제어 : 감성 분석, 감성 사전, PMI, SO-PMI

논문접수일 : 2016년 3월 11일 논문수정일 : 2016년 4월 7일 게재확정일 : 2016년 4월 11일
원고유형 : 일반논문 교신저자 : 김종우

1. 개요

SNS와 온라인 쇼핑몰, 온라인 커뮤니티의 발전으로 소비자들은 상품에 대한 평가에 적극적으로 참여하고 있으며, 이렇게 축적된 상품 평가 정보는 소비자들의 구매행위에 영향을 미치고 있다. 소비자들은 상품구매 전 온라인을 통해 상품에 대한 평가를 검색하고 이를 구매의사에 반영하고 있으며 이에 대응해 기업들은 소비자들에게 리뷰를 조건으로 테스트 제품을 제공하는

* 이 논문은 한양대학교 교내연구지원사업으로 연구되었음(HY-2016년도).

한편, 악의적인 비방에는 강경하게 대응하는 등 판매상품에 대한 온라인상의 평가에 대해 적극적으로 대응하는 추세다. 이에 따라, 소비자들의 평가를 마케팅 데이터로 활용하기 위해 감성분석을 활용한 상품평 분류에 관한 연구가 지속되고 있다. 감성분석은 기업에게 상품에 대한 피드백을 제공하여 상품 개발 또는 판매전략 수립에 정보를 제공하고 있으며 소비자에게는 상품선택에 대한 의사결정을 보조하고 있다(Chang, 2009).

감성분석을 진행하기 위해서는 문장의 긍정, 부정 기준이 되는 감성사전을 사용하게 된다. 그러나 기존에 사용되는 범용 감성사전을 서로 다른 분야의 감성분석에 적용하는 것은 효과적이지 못하다. 예를 들어, 영화리뷰에서 사용가능한 ‘scary’의 경우 공포영화 장르에서는 긍정의 의미로 사용이 가능하지만 그 밖의 드라마, 애니메이션 등의 장르에서는 부정적인 단어로 사용된다. 이렇듯 분석 대상 영역의 특성에 따라 동일한 단어가 서로 다른 의미로 사용될 수 있기 때문에 데이터 영역의 특성에 따라 서로 다른 맞춤형 감성사전이 구축되어야 한다.

본 연구는 감성분석의 예측능력을 향상시키기 위해 데이터 특성에 맞는 맞춤형 감성사전 구축 방법을 제시한다. 본 연구에서 장르에 따라서 감성 어휘들이 차이가 나는 영화 장르별 리뷰 데이터를 대상으로 하였다. 여러 품사 중 어휘 간 극성이 명확하게 판별되는 형용사에 제한된 감성사전을 만들었다. 이후 영화 장르별 감성사전을 감성분석 예측에 활용하여 기존의 범용 감성사전과 성능차이를 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 감성분석과 감성사전 구축에 대한 기존의 선행연구들을 검토하고, 본 연구에 사용한 PMI(Pointwise Mutual Information) 분석방법을 소개한다. 3장에서는 영화 정보 사이트 IMDb에 등록된 영화리뷰를 활용해 영화 장르별 사전구축방법을 제시하며, 4장에서는 맞춤형 감성사전과 기존의 범용 감성사전과의 성능평가를 통해 맞춤형 사전구축 방법의 유용성에 대해 검증한다. 마지막 5장에서는 본 연구의 결론과 한계점, 추후 연구방향에 대하여 제시한다.

2. 관련 연구

2.1 감성분석(Sentiment Analysis) 연구

감성분석(sentiment analysis)은 텍스트마이닝 분석의 한 분야로 특정 문서의 긍정, 부정에 대한 감정을 추측하고 분류하는 방법이다. 감성분석은 각 문서의 최소단위인 단어의 감성극성(sentiment polarity)에 기반을 두어 이루어진다. 즉, 단어의 감성극성이 미리 정의된 감성사전을 구축한 후, 새로 주어진 문서에 출현한 단어의 감성극성에 따라 문서 전체의 감성을 분류하게 된다(Kim and Kim, 2014). 따라서 감성분석은 단어의 감성극성을 정확하게 반영한 감성사전을 사용하는 것이 중요하다. 감성사전 구축에 대한 대표적인 예는 PMI를 활용해 단어의 긍정/중립/부정을 평가한 SentiWordNet 연구가 있다. 하지만 언어의 사용과정에서 해당 문서의 주제에 따른 다양한 문맥적 의미 변이, 동음이의어 사용 등의 이유로 인해 문서의 특성에 맞지 않는 범용감성사전 사용 시 감성분석의 정확성이 낮아지는 문제가 발생한다. 따라서 각 문서의 주제에 맞는 맞춤형감성사전을 구축하고, 이를 통해 감성분석의 정확성을 높이려는 연구들이 다수 수행되고 있다.

감성사전 구축에 관한 연구로는 단어별 극성

판별에 집단지성을 활용한 연구(An et al., 2015), 단어의 출현 빈도수를 활용해 주제별 감성사전을 구축한 연구(Yu et al., 2013), 주식관련 콘텐츠에 OAR(Opinion Antonym Rule) 알고리즘을 사용하여 감성사전을 구축한 연구(Jo et al., 2015), LP(Label Propagation) 방법을 이용해 단어간의 인접도를 통해 감성사전을 구출한 연구(Kim et al., 2015) 등이 있으며 이밖에 TF-IDF (Term Frequency-Inverse Document Frequency) 등의 수치에 기반 하거나 PMI 등의 통계적 수치를 활용하여 단어간의 극성을 정의하는 방법도 있다(Turney and Littman, 2002; Wei Jin et al., 2009; Christopher Scaffidi et al, 2007). 특히 Song and Lee 연구(2013)는 범용감성사전을 활용해 감성분석을 수행하는 방식 보다는 주제에 특화된 감성사전 사용 시 감성분석의 정확성이 향상됨을 확인하였다.

선행연구를 통해 기존의 범용사전을 활용하여 감성분석을 진행하는 것에 비해 해당 문서의 주제별 맞춤형 감성사전을 구축하여 감성분석에 활용하는 방안이 더 효과적임을 알 수 있다. 그러나 기존 다수의 연구에서 사용된 단어의 출현 빈도수를 기준으로 한 감성극성 판별은 단어의 극성강도를 정확히 판별하기 어려우며 PMI방법은 기준단어 선정에 따라 단어극성 판별의 편차가 크게 달라지는 문제점이 발생한다. 또한 기존의 연구는 맞춤형 감성사전의 성능 향상이 다양한 분야에서 검증이 가능한지 대한 평가가 미비한 편이다. 이러한 문제점을 보완하기 위해 본 연구에서는 맞춤형 감성사전 구축에 있어 단어의 극성을 보다 정확히 판별하기 위한 방법인 SO-PMI(Semantic Orientation from Point-wise Mutual Information)를 사용한다. SO-PMI는 기존의 PMI를 보완한 방법으로 기준단어 선정에서 발생할 수 있는 오차를 줄여 단어의 극성 판별 정확성을 높여준다. 또한 본 논문에서는 기존 연구들이 특정 분야에 한정하여 맞춤형 감성사전의 성능평가를 진행한 것을 보완하기 위해 6개의 서로 다른 영화 장르에 대한 성능을 검증한다. 이에 따라 본 연구가 갖는 기존 연구와의 차이점은 다음과 같다. 첫째, 단어의 극성 판별에 SO-PMI를 활용해 장르별 맞춤형 감성사전을 구축하여 보다 정확한 감성사전 구축을 도모하였다. 범용으로 사용되는 감성사전의 단어는 연구대상과 목적에 따라 다른 의미로 사용될 수 있기 때문에 이를 보완하기 위해 영화 장르별 맞춤형 감성사전을 구축하였다. 둘째, 맞춤형 감성사전의 성능 향상을 보다 정확히 확인하기 위해 서로 다른 6개 장르에 적용하여 평가하였다. 마지막으로 맞춤형 감성사전을 구축하는 일반적인 방법을 제시해 다양한 주제의 연구 대상으로 확장 가능한 방법을 꾀하였다.

2.2 PMI(Pointwise Mutual Information) 연구

PMI는 두 확률변수의 연관성을 표현하는 방법으로 단어 간의 유사성을 분석하는 지표로 활용되며, 분석하고자 하는 두 단어의 의미극성이 비슷할 경우 같은 문서 내에서 사용될 확률이 높다는 가정을 통해 계산된다(Song et al., 2010). 두 단어간의 연관성을 계산하는 PMI의 계산식은 다음 식 (1)과 같다.

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \quad (1)$$

여기서 w_1과 w_2는 분석하고자 하는 두 단어를 나타내며 $PMI(w_1, w_2)$ 는 한 문서 내에서 w_1과 w_2가 함께 출현할 확률을 w_1과 w_2가 각각 출현할 확률의 곱으로 나눈 값이다. $PMI(w_1, w_2)$ 는 두 단어가 같은 문서에서 나타날 확률을 나타내며

PMI가 클수록 두 단어 사이의 유사성이 높아진다. 따라서 두 단어의 극성이 비슷한 것으로 이해할 수 있다. 반대로 PMI값이 작을수록 두 단어 사이의 유사성이 낮아지며 두 단어가 서로 다른 극성을 가짐으로 해석할 수 있다.

본 논문에서는 PMI를 보완한 SO-PMI를 활용한다. SO-PMI를 계산하는 방법은 다음 식 (2)과 같다.

$$SO-PMI(w) = \sum_{pw \in PW} PMI(w,pw) - \sum_{nw \in NW} PMI(w,nw) \quad (2)$$

여기서 *PW*는 긍정기준단어의 집합을, *NW*는 부정기준단어의 집합을 의미한다. $SO-PMI(w)$ 는 단어 *w*와 긍정단어집합과의 PMI 합에서 부정단어집합과의 PMI 합을 뺀 결과 값이다. PMI는 기준단어의 선택에 따라 분석되는 결과 값이 크게 변한다. 이를 보완하기 위한 SO-PMI는 다수의 기준단어를 설정해 PMI에서 발생될 수 있는 개별 단어에 대한 편향을 줄여주어 단어의 극성을 판별하는 기준으로 활용될 수 있다. SO-PMI점수가 높으면 긍정기준단어 집합과 유사성이 높으며 부정기준단어 집합과 유사성이 낮으므로 긍정단어로 해석할 수 있다. 마찬가지로 SO-PMI점수가 낮으면 부정단어로 해석할 수 있다(Song et al., 2010; Peter D et al., 2003). 본 연구에서는 감성사전 구축을 위해 단어의 극성을 판별하는 SO-PMI를 사용한다.

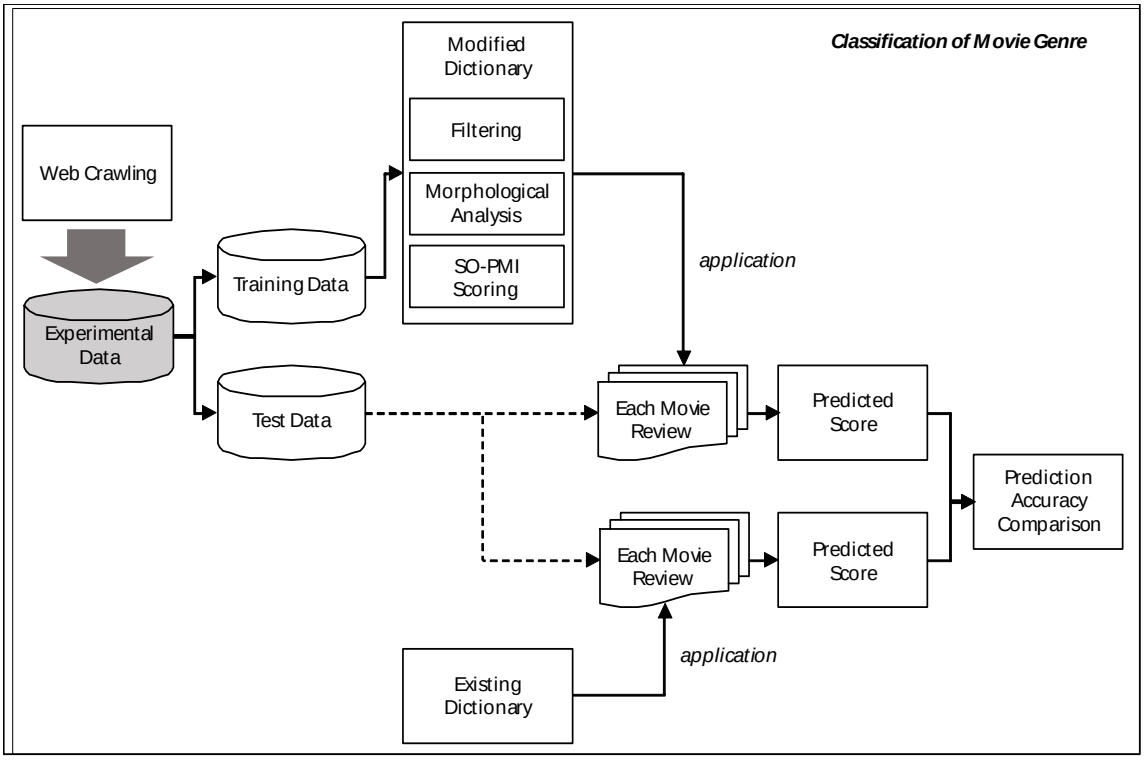
3. 연구 방안

3.1 연구 데이터

본 연구에서는 영화에 관한 정보를 제공하는 가장 대표적인 사이트인 인터넷 영화 데이터베이스(Internet Movie Database 약칭 IMDb)에 등록되어 있는 장르별 영화 리뷰 데이터를 활용하였다. 연구에서 활용한 영화의 장르는 '액션', '애니메이션', '코메디', '드라마', '공포', '과학공상' 6가지이며, 6가지 장르 이외에 영화 전체 장르에 대한 리뷰를 추가해 총 7개의 데이터 집합을 분석하였다. 훈련데이터는 각 장르별 별점(Rating) 기준 상위 5개 영화의 리뷰, 하위 5개 영화의 리뷰, 총 47,196개의 데이터를 사용하였으며, 테스트데이터는 2012년 9월부터 2014년 6월까지 기간에 각 6개 장르별 상위 100개 영화의 리뷰를 수집한 총 42,083개의 데이터를 사용하였다(<표 1> 참조). 훈련데이터는 사전 구축을 위해서 사용되는 데이터로 긍정, 부정 단어의 추출이 용이하도록 선정하였다. 즉, 긍정 위주의 리뷰가 주를 이루는 장르별 상위 5개 영화와 부정리뷰를 다수 포함하는 하위 5개 영화를 선정하여 사전 구축 시 다양한 긍정, 부정 어휘를 반영할 수 있도록 했다. 테스트데이터는 훈련데이터를 이용하여 구축된 사전들의 성능을 평가하기 위한 데이터로 일정기간 동안 개봉된 영화 중 리뷰의 개수가 어느 정도 이상 되는 상위 100개 영화를 기준으로 선정하였다. 훈련데이터 사용 시 여러 품사 중 단어 간의 극성이 뚜렷하게 구분되는 형용사를 기준으로 분석을 진행했다.

〈Table1〉 The number of movies in experimental data set

Genre	Training data set	Test data set
Action	15,724	13,894
Animation	4,109	2,870
Comedy	3,617	9,126
Drama	7,549	9,318
Horror	7,600	2,449
Sci-fi	8,597	4,426
Total	47,196	42,083



〈Figure 1〉 Movie rating prediction by using modified sentiment dictionary of genre

3.2 연구 절차

본 연구는 다음과 같은 연구절차를 통하여 수행되었다(<그림 1> 참조). 실험에 필요한 데이터는 영화관련 정보를 제공하는 IMDb에서 웹 크롤링(web crawling)으로 수집하였으며 각 영화별 리뷰 데이터를 활용했다. 맞춤형 사전구축용 훈련데이터는 각 장르별 전체 순위 상위 5개, 하위 5개 영화에 대한 리뷰를 사용하였으며 훈련용 데이터와 테스트용 데이터 분석은 각각 R에서 패키지로 제공되는 'tm'[1]과 'OpenNLP'[2], 'SentR'[3]을 활용하여 텍스트 분석을 수행하였다. 연구의 결과 비교는 'SentR' 패키지에서 기본으로 제공하는 범용 감성사전과, 훈련용 데이터를 통해 새로 정의한 형용사를 기존의 범용 감성사전 업데이트한 장르별 맞춤형 감성사전을 테스트용 데이터에 적용하여, 테스트용 데이터의 별점(Rating)에 대한 예측 정확도를 비교하였다.

1) Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software, Inc. [ctb, cph] (2014). tm: Text Mining Package. R Package Version 0.6. http://CRAN.R-project.org/package=tm
2) Kurt Hornik [aut, cre](2015). OpenNLP: supports the most common NLP tasks Package. R Package Version 1.5.3. https://cran.r-project.org/web/packages/openNLP/index.html
3) Manan Shah(2015). SentR: Provides functional sentiment analysis Package. R Package Version 1.0. https://github.com/mananshah99/sentR

3.2.1 단어의 극성 판단

감성사전을 구축하기 위해 본 연구에서는 수집된 리뷰에서 단어 간의 극성이 명확하게 구분되는 '형용사' 만을 활용하였다. 형용사 추출에 앞서 'quot', 'br' 등 웹 크롤링 과정에서 발생한 불필요한 단어와 의미를 알 수 없는 단어를 제거하였으며 숫자, 특수기호 등의 불용어를 제거하였다. 최종적으로 추출된 형용사들의 SO-PMI 점수를 계산해 단어의 극성을 판단한다.

PMI는 두 단어 간의 유사성을 나타내는 점수로 PMI값이 높으면 두 단어 사이의 유사성이 높다는 것을 의미한다. 이렇게 계산한 PMI값은 SO-PMI 계산에 활용한다. SO-PMI 적용에 앞서 분석 데이터의 긍정기준단어 집합과 부정기준단어 집합을 우선적으로 설정한 이후 분석대상 단어와 긍정기준단어와의 PMI 합에서 부정기준단어와의 PMI 합을 뺀 점수인 SO-PMI 점수를 계산한다. SO-PMI 점수가 높으면 긍정기준단어 집합과 유사성이 높으며 부정기준단어 집합과 유사성이 낮으므로 긍정단어로 해석할 수 있다. 마찬가지로 SO-PMI점수가 낮으면 부정단어로 해석할 수 있다.

SO-PMI 값은 기준단어집합에 영향을 받으므로 기준단어집합을 적절하게 설정하는 것이 중요하다. 본 연구에서 기준단어집합을 만들기 위해 훈련데이터의 리뷰별점(Rating) 상위 30%를 긍정단어집합, 하위 30%를 부정단어집합으로 설정하였다. 두 집합에서 생성된 단어의 빈도수를 비교하여 해당 단어의 빈도수가 더 높은 집합에 귀속시킨다. 이렇게 생성된 긍정과 부정 상위 15개의 단어를 기준단어집단으로 만들었다. 기준단어집단의 개수선정은 선행연구에 근거하였다(Song et al, 2010). 본 연구에서 선정한 'action'의 기준단어집합은 다음과 같다(<표 2> 참고).

〈Table 2〉 Standard words on action genre

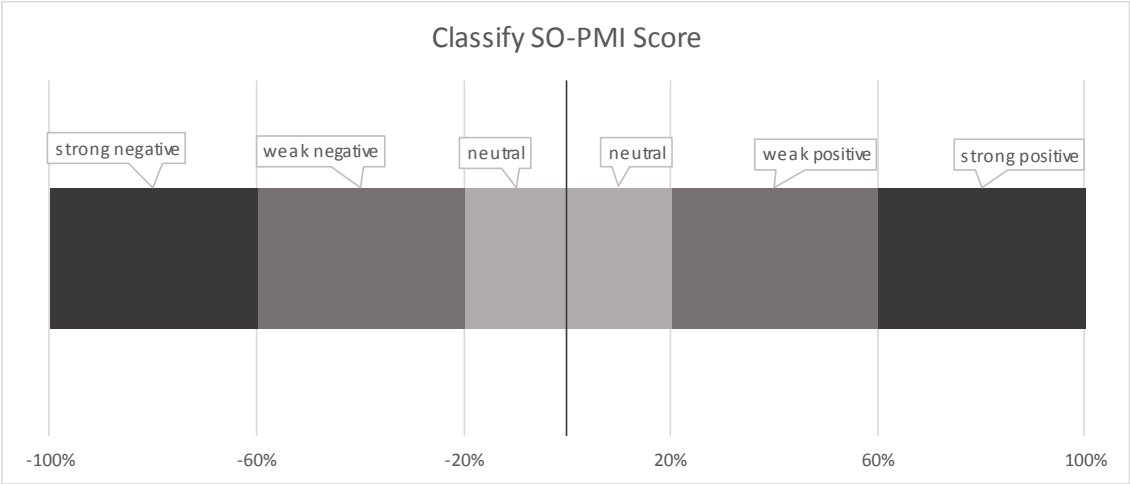
Positive words		Negative words	
Words	Frequency	Words	Frequency
best	6499	bad	2356
dark	5648	even	2097
great	5632	much	1232
good	5543	worst	817
many	2681	right	534
new	2278	stupid	503
special	2239	big	474
real	1943	terrible	428
amazing	1889	pretty	412
perfect	1576	indian	403
little	1571	awful	381
comic	1423	worse	347
original	1339	half	344
brilliant	1157	horrible	316
enough	1144	evil	264

3.2.2 감성사전 구축

SO-PMI를 통해 단어별 극성점수를 판단한 후 이를 토대로 감성사전을 구축한다. 감성사전에서 각 단어는 'strong positive', 'weak positive', 'strong negative', 'weak negative,' 'neutral' 5개로 구분한다. SO-PMI가 0 이상이면서 그 값이 0 이상 값 중 상위 40%에 해당하면 'strong positive', 상위 80%에 해당하면 'weak positive', 나머지는 'neutral'로 정의했으며 0 이하에서도 마찬가지의 규칙을 적용했다. SO-PMI를 기준으로 단어의 극성을 분류한 규칙은 다음 식 (3)과 같다(<그림2> 참조). 또한 본 연구에서 'action'의 단어들의 극성을 분류한 예는 다음과 같다(<표 3>참조).

$$Classify(SO-PMI)$$

$$\begin{cases} \geq 0 \begin{cases} \geq 60\%, Strongpositive \\ 20\% \leq SO-PMI \leq 60\%, weakpositive \\ \leq 20\%, neutral \end{cases} \\ \leq 0 \begin{cases} \leq -60\%, Strongnegative \\ -60\% \leq SO-PMI \leq -20\%, weaknegative \\ \geq -20\%, neutral \end{cases} \end{cases} \quad (3)$$



〈Figure 2〉 Classify SO-PMI score

〈Table 3〉 Classified words on action genre

Positive		Neutral		Negative	
Words	Scale	Words	Scale	Words	Scale
afraid	strong	able	weak	awful	strong
awesome	weak	base	weak	bad	strong
cinematic	weak	bigger	weak	complete	weak
clown	strong	common	weak	damn	weak
crazy	strong	cool	weak	dead	strong
epic	weak	entire	weak	direct	weak
fresh	weak	fine	weak	decent	weak
future	weak	full	weak	dumb	strong
impressed	strong	good	weak	fake	strong
incredible	strong	last	weak	horrible	strong
loud	strong	live	weak	low	strong
new	weak	next	weak	obvious	weak
psychotic	strong	open	weak	ridiculous	strong
scary	strong	sad	weak	silly	weak
scifi	weak	sure	weak	slow	weak
special	weak	surprised	weak	strange	weak
modern	weak	typical	weak	stupid	strong
strong	weak	normal	weak	terrible	strong
superb	strong	various	weak	weak	weak
technical	weak	violent	weak	worst	strong

3.3.3 감성사전을 활용한 별점(Rating) 예측방안

본 연구에서 활용한 SentR 패키지는 6,518개의 단어로 구성된 감성사전을 제공하며, 라이브러리 내 저장된 감성사전을 활용해 나이브베이지안 분류를 진행한다. 본 연구에서는 기존의 SentR 패키지에서 제공한 범용 감성사전에 훈련용 데이터를 바탕으로 만든 장르별 형용사 감성사전을 추가하여 장르별 맞춤형 감성사전을 구축한다. 이후 새로 구축한 맞춤형 감성사전을 'SentR' 라이브러리에 적용하여 계산된 긍정/부정 비율과 기존의 범용 감성사전으로 계산된 긍정/부정 비율을 테스트용 데이터의 리뷰별 별점 예측에 적용하여 비교한다. 예측은 'sentR' 패키지에서 제공하는 나이브베이지안 분류(Naive Bayesian classification)를 사용하며 나이브베이지안 분류식(Naive Bayesian classification)은 다음 식 (4)와 같다.

$$C_{NB}(d) = Argmax_{c_j \in C} P(c_j) \prod_{i=1}^{n} P(w_i|c_j) \quad (4)$$

여기서 C_{NB}는 문서의 클래스를 결정하는 함수이고 $w_i(i=1\sim n)$는 문서 d에 포함된 단어들을 표시하고 c_j는 문서의 클래스(긍정, 부정)를 의미한다. $P(w_i|c_j)$는 w_i가 c_j에 속할 확률이다. 본 연구에서는 나이브베이지안 분류식을 활용해 해당 문서가 긍정에 속할 확률과 부정에 속할 확률을 각각 계산한다. 이를 바탕으로 본 연구에서는 'sentR' 패키지를 통해 계산된 각 리뷰별 나이브베이지안 분류의 긍정/부정 스코어(score)를 단어의 극성 판별에 사용하며, 기준 별점(Rating)과 비교하기 위해 10점 척도로 변환하여 별점(Rating) 예측에 활용한다. 긍정/부정 스코어를 기존 별점과 같은 10점 척도로 변환하기 위해 훈련데이터에서 각 영화장르별 리뷰의 1점부터 10점까지의 별점비율을 추출한다. 훈련데이터를 통해 만들어진 별점비율을 나이브베이지안을 통해 계산된 긍정/부정 스코어에 1점부터 10점까지 순서대로 할당해 긍정/부정 스코어를 10점 척도로 변환한다.

4. 연구 결과

4.1 장르별 별점(Rating) 예측 평가

본 연구에서는 범용감성사전과 맞춤형감성사전의 성능을 비교하기 위해 절대평균오차(Mean absolute error)를 사용했다. 절대평균오차는 각 리뷰별 실제 별점과 별점 예측치간 절대오차의 평균을 계산한 값으로 그 값이 작을수록 사전의 성능이 높은 것으로 해석할 수 있다.

또한 절대평균오차를 기준으로 향상도(Improvement)는 다음 (5)와 같이 정의될 수 있다. 즉, 향상도는 범용감성사전 사용 시에 비해서, 맞춤형 감성사전 사용한 경우 몇 퍼센트 향상되었는지를 보여준다.

$$Improvement = 1 - \frac{MAE_{Modified\,Dictionary}}{MAE_{Original\,Dictionary}} \quad (5)$$

본 논문에서 제안한 SO-PMI를 활용한 감성사전 구축을 통해 별점을 예측한 결과는 다음과 같다(<표 4> 참조). 본 연구에서 '액션', '코메디',

〈Table 4〉 Rating prediction accuracy of movie genre

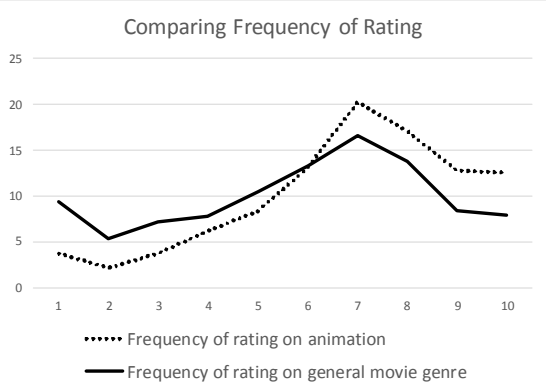
Movie genre	Type of dictionary	MAE	Paired sample T-test (p-value)	Improve-ment
Action	Original dictionary	2.315	0.000***	2.76%
	Modified dictionary	2.251		
Animation	Original dictionary	2.144	0.165	-1.17%
	Modified dictionary	2.169		
Comedy	Original dictionary	2.440	0.000***	2.21%
	Modified dictionary	2.386		
Drama	Original dictionary	2.393	0.000***	1.80%
	Modified dictionary	2.350		
Horror	Original dictionary	2.516	0.030*	2.03%
	Modified dictionary	2.465		
Sci-fi	Original dictionary	2.572	0.000***	4.16%
	Modified dictionary	2.465		
Total	Original dictionary	2.411	0.000***	2.82%
	Modified dictionary	2.343		

significance level *** : p<0.001, ** : p<0.01, * : p<0.05.

‘드라마’, ‘공포’, ‘과학공상’ 5개 장르와 영화 전체데이터에 있어서는 새로 구축한 맞춤형 사전의 성능이 기존사전에 비해 좋은 것으로 분석되었다. 반면 ‘애니메이션’ 장르의 경우 두 사전의 성능차이가 유의수준을 벗어나지 못해 두 사전 간의 성능에 유의미한 차이가 없는 것으로 확인되었다. 가장 성능 향상이 큰 경우는 ‘과학공상’으로 4.16%의 증가를 보였다. 또한 분석대상인 장르별 리뷰의 개수에도 차이가 있었다. ‘액션’은 13,894개로 가장 많았으며 ‘드라마’ 9318개, ‘코메디’ 9,126개, ‘과학공상’ 4,426개, ‘공포’ 2,449개 순서를 보였다. 성능 향상이 확인되지 않았던 ‘애니메이션’의 경우 2,870개로 상대적으로 다른 장르에 비해 영화리뷰의 개수가 적은 것을 알 수 있다.

4.2 애니메이션 장르 성능 차이 분석

기존 리뷰개수가 다른 장르에 비해 상대적으로 적었던 ‘애니메이션’의 경우 유의미한 성능 향상을 이끌어내지 못했다. 이것은 ‘애니메이션’ 데이터의 특징에서 비롯한 것으로 생각된다. ‘애니메이션’ 장르의 별점의 경우 다른 일반적인 영화장르에 비해 긍정비율이 압도적으로 높았다(<그림 3> 참조). 이러한 경향은 SO-PMI의 기준단어를 설정하는데 영향을 주었다(<표 5> 참조). 일반적인 영화의 경우 긍정기준단어로 설정한 단어의 빈도수가 부정기준단어의 단어 빈도수에 비해 평균적으로 약 4.5배 이상 높은 경향을 보인다. 그러나 ‘애니메이션’의 경우 긍정기준단어의 단어빈도수가 부정기준단어의 단어빈도수에 비해 평균 약 46배 이상 높은 경향을 보였다. 이는 일반적인 영화에 비해 10배 이상 높은 수치로 맞춤형 감성사전 구축을 위한 단어의 극성 설정



〈Figure 3〉 Comparing Frequency of rating between animation and general movie

에 문제점을 가져왔다. ‘애니메이션’ 훈련데이터를 통해 구축한 맞춤형 감성사전은 일반적인 영화의 감성사전과 다른 경향을 보였다(<표 6> 참조). <표 6>은 ‘애니메이션’ 장르와 일반 영화에서 공통으로 사용된 단어 중 극성이 크게 차이나는 것을 수집한 자료이다. 일반적으로 긍정이라고 생각되는 ‘best’와 ‘beautiful’ 이 중립으로 설정되어있으며 부정으로 생각되는 ‘dead’와 ‘dull’은 긍정으로 설정되어 있다. 이러한 사전구축의 오류가 별점예측에 영향을 주어 성능의 향상이 이루어지지 않은 것으로 생각된다.

애니메이션 장르의 성능 향상이 이루어지지 않은 또 다른 이유는 본 연구가 단어들 간의 상관관계나 문장 구조를 고려하지 않고, 단지 개별 단어들의 출현빈도만을 고려하는 접근 방법에 기초하고 있기 때문인 것으로 생각된다. 본 연구에서는 문장 분석 시 형태소분석을 통해 품사를 분류하여 형용사만을 사용하여 텍스트분석을 진행하였기 때문에 ‘not’, ‘never’ 등의 부정어는 분석에 포함되지 않았다. 예를 들어 ‘okay’의 경우 일반 영화장르에서 부정어와 결합되어 사용되는

〈Table 5〉 Comparing frequency of standard words between animation and general movie

Animation				General movie			
Positive		Negative		Positive		Negative	
Words	Frequency	Words	Frequency	Words	Frequency	Words	Frequency
best	1477	better	66	best	17876	bad	6355
great	1454	right	24	great	16251	first	2985
good	1318	horrible	22	good	14980	worst	2733
much	1144	worse	19	many	9148	better	2294
first	1129	else	18	dark	7104	stupid	1285
little	1122	mean	18	new	5713	awful	1209
many	949	evil	17	little	5432	right	1185
japanese	693	awful	14	special	5179	terrible	1180
beautiful	658	quality	14	amazing	4596	worse	1057
old	572	american	13	real	4501	horrible	904
new	566	pointless	13	perfect	4420	poor	851
human	547	sound	12	original	4137	half	847
amazing	470	clumsy	11	top	3670	less	598
real	446	ridiculous	11	greatest	3530	ridiculous	507
wonderful	360	crazy	9	excellent	3164	indian	430

비율이 단독으로 사용되는 경우보다 압도적으로 높아 ‘strong negative’로 분류되었다. 그러나 애니메이션 장르에서는 일반 영화장르에 비해 ‘okay’단독으로 사용된 비율이 높아 ‘neutral’로 분류되었다. ‘interested’의 경우도 마찬가지로 일반 영화장르에 비해 애니메이션 장르에서 단독으로 사용된 비율이 높아 ‘strong positive’로 분류되었다. 이렇게 별도의 부정어에 대한 고려를 하지 않은 것이 영화장르 전체에 영향을 주었으며 특히 애니메이션 장르의 성능 향상을 이끌어내지 못한데 주요한 영향을 준 것으로 생각된다. 따라서 부정어와 동시에 출현한 경우에 대한 적절한 처리를 통해서 추후 성능 향상이 가능할 것으로 보인다.

반면 기존 감성사전에 수록된 단어의 극성과 차이가 가장 클 것으로 예상했던 ‘액션’과 ‘공포’, ‘과학공상’ 장르에서는 새로 구축한 맞춤형 사전의 성능이 더 뛰어남을 알 수 있었다. 이를 바탕으로 SO-PMI를 활용한 맞춤형 감성사전 구축방안은 향후 기존 감성사전의 단어극성과 다를 것으로 예상되었던 경제기사, 가전제품 리뷰, 상품 불만접수 등 다양한 분야에도 폭넓게 적용되어 감성분석의 정확성 향상에 기여할 수 있을 것으로 생각된다.

〈Table 6〉 Comparing polarity of words between animation and general movie

Words	Polarity	
	Animation	General movie
afraid	strong positive	neutral
amazing	neutral	strong positive
basic	weak positive	weak negative
beautiful	neutral	strong positive
best	neutral	strong positive
brilliant	neutral	strong positive
clear	strong positive	neutral
common	strong positive	neutral
cool	weak positive	weak negative
dark	neutral	strong positive
dead	weak positive	strong negative
dull	strong positive	strong negative
excellent	neutral	strong positive
final	weak negative	weak positive
general	weak positive	weak negative
great	neutral	strong positive
greatest	weak negative	strong positive
hilarious	neutral	strong negative
horrible	weak positive	strong negative
interested	strong positive	neutral
interesting	weak negative	weak positive
japanese	neutral	strong positive
low	weak positive	strong negative
major	strong positive	neutral
memorable	neutral	strong positive
middle	strong positive	weak negative
negative	weak positive	strong negative
nice	strong negative	neutral
okay	neutral	strong negative
older	weak positive	strong positive
open	strong positive	weak negative
original	weak negative	weak positive
perfect	neutral	strong positive
plain	strong positive	strong negative
poor	neutral	strong negative
predictable	neutral	strong negative
realistic	weak negative	strong positive
scary	strong positive	weak negative
weak	weak positive	weak negative
white	strong positive	weak negative
wrong	neutral	strong negative

5. 결론

본 연구에서는 SO-PMI를 사용한 단어의 극성 판별을 통해 감성사전 구축에 대한 연구를 수행하였다. 연구결과 새로 구축한 맞춤형 감성사전 사용 시, 기존의 범용감성사전에 비해 평균 2.82%의 예측정확도의 향상을 보였다. '과학공상' 장르의 경우 4.16%의 향상되어 가장 큰 차이를 보였으며, 리뷰 개수가 가장 많았던 '액션'의 경우 예측정확도가 2.76% 향상되었다. 하지만, 리뷰 개수가 가장 적었던 '애니메이션'은 통계적으로 유의하지는 않았지만, 1.17%의 성능 감소를 보였다.

본 연구는 SO-PMI를 활용하여 영화 장르별 단어의 극성을 새로 정의하였으며, 이를 감성사전 구축에 사용해 감성분석의 실질적 성능 향상을 확인한 것이 주요 학문적 기여라고 할 수 있다. 그러나 각 분야별 감성분석에 필요한 맞춤형 감성사전 구축을 위해 본 연구가 갖는 연구 한계점은 다음과 같으며 이후 추가적인 연구가 필요하다.

첫째, 본 연구에서 훈련데이터를 활용한 맞춤형 감성사전 구축 시 수집한 영화 장르별 리뷰 중 단어의 극성이 뚜렷하다고 예상되는 형용사만을 한정하여 사용하였다. 그러나 동사, 부사 등 다른 여러 품사의 단어도 감성분석에 활용이 가능할 것으로 예상되는 바 여러 가지 품사를 포함한 감성사전 개선에 대한 연구가 진행되어야 할 것으로 생각한다.

둘째, 본 연구에서 '애니메이션' 장르의 예측정확도가 개선되지 않은 이유를 분석하여 감성사전 구축 방안을 개선해야 한다. 본 논문에서는 훈련데이터 수집 시 데이터의 긍정/부정 비율을 고려하지 않았으며 부정어 처리를 진행하지 않았기 때문에 예측정확도에 문제가 발생한 것으로 결론지었다. 추후 연구에서 이러한 문제점을 개선할 수 있는 방법을 제시한다면 맞춤형 감성사전의 감성분석 예측 정확도가 향상될 것으로 생각된다.

셋째, 본 연구에서는 데이터의 특성이 명확하게 구분되는 영화장르 데이터를 활용했으며 이를 위해 영화정보를 폭넓게 제공하는 IMDb 웹사이트의 데이터를 수집 하였다. 그러나 IMDb의 영화리뷰는 평균 200단어가 넘는 긴 글로 구성되어있는 만큼, 장문에 대한 SO-PMI분석이 단문 위주로 구성된 다른 분야에 적용가능할지 연구가 필요하다. 또한 영화 이외의 다른 분야에서도 맞춤형 감성사전이 일관성 있게 성능을 보일지 추가적인 연구가 진행되어야 한다.

참고문헌(References)

Adhitama P., S. H. Kim and I. S. Na, "Twitter Trending Topic Classification using Naive Bayes Classifier," *Proceedings of the Korean Information Science Society Conference,* Vol.40(2013), 879~881.

An J. K. and H. W. Kim, "Building a Korean Sentiment Lexicon Using Collective Intelligence," *Journal of Intelligent Information Systems*, Vol.21, No.2(2015), 49~67.

Chang J. Y., "A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in On-Line Shopping Mall," *The Journal of Society for e-Business Studies*, Vol.14, No.4(2009), 19~33.

Cho T. M., H. N. Cho, J. D. Lee and J. H. Lee,

"TV Drama Rating Prediction based on Sentiment Analysis of Viewers` Comments," *Proceedings of the Korean Institute of Intelligent Systems Conference*, Vol.24, No.1 (2014), 83~84.

Jin W., H. H. Ho and R. K. Srihari, "OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction," *KDD Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*(2009), 1195~1204.

Jo, E. K., "The Current State of Affairs of the Sentiment Analysis and Case Study Based on Corpus," *The Journal of Linguistic Science,* Vol.61(2012), 259~282.

Jo H. J., J. H. Seo and J. T. Choi, "OAR Algorithm Technology Based on Opinion Mining Utilizing Stock News Contents," *Journal of Korean Institute of Information Technology*, Vol.13, No.2(2015), 111~119.

Kim J. H., Y. J. Oh and S. H. Chae, "The Construction of a Domain-Specific Sentiment Dictionary Using Graph-based Semi-supervised Learning Method," *Korean Journal of the Science of Emotion and Sensibility*, Vol.18, No.4(2015), 97~104.

Kim K. P. and Y. S. Kwon, "Performance Comparison of Naive Bayesian Learning and Centroid-Based Classification for e-Mail Classification," *IE Interfaces* Vol.18, No.1 (2005), 10~21.

Kim S. W. and N. G. Kim, "A Study on the Effect of Using Sentiment Lexicon in Opinion Classification," *Journal of Intelligent Information Systems*, Vol.20, No.1(2014), 133~148.

Lee K. B., J. B. Baik and S. W. Lee, "Estimating a Pleasure-Displeasure Index of Word based on Word Similarity in SNS," *Journal of KIISE : Computing Practices and Letters*, Vol.20, No.3(2014), 159~164.

Oh S. H. and S. J. Kang, "Movie Retrieval System by Analyzing Sentimental Keyword from User's Movie Reviews," *Journal of the Korea Academia-Industrial cooperation Society*, Vol.14, No.3(2013), 1422~1427.

Scaffidi C., K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red Opal: Product-Feature Scoring from Reviews," *Proceedings of the 8th ACM conference on Electronic commerce*(2007), 182~191.

Seo J. H., H. J. Jo and J. T. Choi, "Design for Opinion Dictionary of Emotion Applying Rules for Antonym of the Korean Grammar," *Journal of Korean Institute of Information Technology,* Vol.13, No.2(2015), 109~117.

Song J. S., and S. W. Lee, "Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews," *Journal of KIISE: Software and Applications*, Vol.38, No.3 (2013), 157~168.

Song S. I., D. J. Lee and S. G. Lee, "Identifying Sentiment Polarity of Korean Vocabulary Using PMI," *Proceedings of the Korean Information Science Society Conference*, Vol.37, No.1(2010), 260~265.

Turney P. D. and M.L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus," *National Research Council, Institute for Information Technology, Technical Report*(2002), ERB-1094.

Turney P. D., and M. L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Transactions on Information Systems (TOIS),* Vol.21, No.4(2003), 315~346.

Yeon J. H., D. J. Lee, J. H. Shim and S. G. Lee, "Product Review Data and Sentiment Analytical Processing Modeling," *The Journal of Society for e-Business Studies,* Vol.16, No.4(2011), 125~137.

Yu E. J., Y. S. Kim, N. Y. Kim and S. R. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligent Information Systems*, Vol.19, No.1(2013), 95~10.

Abstract

Sentiment analysis on movie review through building modified sentiment dictionary by movie genre

Sang Hoon Lee* · Jing Cui** · Jong Woo Kim***

Due to the growth of internet data and the rapid development of internet technology, “big data” analysis is actively conducted to analyze enormous data for various purposes. Especially in recent years, a number of studies have been performed on the applications of text mining techniques in order to overcome the limitations of existing structured data analysis. Various studies on sentiment analysis, the part of text mining techniques, are actively studied to score opinions based on the distribution of polarity of words in documents. Usually, the sentiment analysis uses sentiment dictionary contains positivity and negativity of vocabularies. As a part of such studies, this study tries to construct sentiment dictionary which is customized to specific data domain. Using a common sentiment dictionary for sentiment analysis without considering data domain characteristic cannot reflect contextual expression only used in the specific data domain. So, we can expect using a modified sentiment dictionary customized to data domain can lead the improvement of sentiment analysis efficiency. Therefore, this study aims to suggest a way to construct customized dictionary to reflect characteristics of data domain. Especially, in this study, movie review data are divided by genre and construct genre-customized dictionaries. The performance of customized dictionary in sentiment analysis is compared with a common sentiment dictionary. In this study, IMDb data are chosen as the subject of analysis, and movie reviews are categorized by genre. Six genres in IMDb, ‘action’, ‘animation’, ‘comedy’, ‘drama’, ‘horror’, and ‘sci-fi’ are selected. Five highest ranking movies and five lowest ranking movies per genre are selected as training data set and two years’ movie data from 2012 September 2012 to June 2014 are collected as test data set. Using SO-PMI (Semantic Orientation from Point-wise Mutual Information) technique, we build customized sentiment dictionary per genre and compare prediction accuracy on review rating. As a result of the analysis, the prediction using customized

* Dept. of Mathematics, College of Natural Sciences, Hanyang University
** Dept. of Business Administration, Graduate School, Hanyang University
*** Corresponding Author: Jongwoo Kim
School of Business, Hanyang University
222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea
Tel: +82-2-2220-1067, Fax: +82-2-2220-1169, E-mail: kjw@hanyang.ac.kr

dictionaries improves prediction accuracy. The performance improvement is 2.82% in overall and is statistical significant. Especially, the customized dictionary on ‘sci-fi’ leads the highest accuracy improvement among six genres. Even though this study shows the usefulness of customized dictionaries in sentiment analysis, further studies are required to generalize the results. In this study, we only consider adjectives as additional terms in customized sentiment dictionary. Other part of text such as verb and adverb can be considered to improve sentiment analysis performance. Also, we need to apply customized sentiment dictionary to other domain such as product reviews.

Key Words : Sentiment Analysis, Sentiment Dictionary, PMI, SO-PMI

Received : March 11, 2016 Revised : April 7, 2016 Accepted : April 11, 2016
Publication Type : Regular Paper Corresponding Author : Jong Woo Kim

저 자 소 개

이 상 훈
현재 한양대학교 자연과학대학 수학과에 재학 중이다. 주요 연구 관심분야는 데이터마이닝 기법과 응용, 오피니언 마이닝 등이다.

최　정
현재 한양대학교 일반대학원 경영학과 석사과정을 수료하였다. 중국 하얼빈공업대학교 경영학과에서 학부를 마쳤으며, 주요 연구 관심분야는 데이터마이닝 기법과 응용, 오피니언 마이닝, 빅데이터 등이다.

김 종 우
현재 한양대학교 경영대학 경영학부 교수로 재직 중이다. 서울대학교 수학과에서 학사를 마쳤으며, 한국과학기술원에서 경영과학으로 석사학위를, 산업경영학으로 박사학위를 취득하였다. 주요 연구 관심분야는 데이터마이닝 기법과 응용, 오피니언 마이닝, 상품추천기술, 지능형 정보시스템, 집단지성, 사회 네트워크 분석, 클라우드 컴퓨팅 서비스 등이다.