ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 March: 20(1): 133~148 http://dx.doi.org/10.13088/jiis.2014.20.1.133

오피니언 분류의 감성사전 활용효과에 대한 연구*

김승우
국민대학교 Business IT 전문대학원
(swk224@naver.com)

김남규
국민대학교 Business IT 전문대학원
(ngkim@kookmin.ac.kr)

최근 다양한 정보채널들의 등장으로 인해 빅데이터에 대한 관심이 높아지고 있다. 이와 같은 현상의 가장 큰 원인은, 스마트기기의 사용이 활성화 됨에 따라 사용자가 생성하는 텍스트, 사진, 동영상과 같은 비정형 데이터의 양이 크게 증가하고 있는 것에서 찾을 수 있다. 특히 비정형 데이터 중에서도 텍스트 데이터의 경우, 사용자들의 의견 및 다양한 정보를 명확하게 표현하고 있다는 특징이 있다. 따라서 이러한 텍스트에 대한 분석을 통해 새로운 가치를 창출하고자 하는 시도가 활발히 이루어지고 있다. 텍스트 분석을 위해 필요한 기술은 대표적으로 텍스트 마이닝과 오피니언 마이닝이 있다. 텍스트 마이닝과 오피니언 마이닝은 모두 텍스트 데이터를 입력 데이터로 사용할 뿐 아니라 파싱, 필터링 등 자연어 처리 기술을 사용한다는 측면에서 많은 공통점을 갖고 있다. 특히 문서의 분류 및 예측에 있어서 목적 변수가 긍정 또는 부정의 감성을 나타내는 경우에는, 전통적 텍스트 마이닝, 또는 감성사전 기반의 오피니언 마이닝의 두 가지 방법론에 의해 오피니언 분류를 수행할 수 있다. 따라서 텍스트 마이닝과 오피니언 마이닝의 특징을 구분하는 가장 명확한 기준은 입력 데이터의 형태, 분석의 목적, 분석의 결과물이 아닌 감성사전의 사용 여부라고 할 수 있다. 따라서 본 연구에서는 오피니언 분류라는 동일한 목적에 대해 텍스트 마이닝과 오피니언 마이닝을 각각 사용하여 예측 모델을 수립하는 과정을 비교하고, 결과로 도출된 모델의 예측 정확도를 비교하였다. 오피니언 분류 실험을 위해 영화 리뷰 2,000건에 대한 실험을 수행하였으며, 실험 결과 오피니언 마이닝을 통해 수립된 모델이 텍스트 마이닝 모델에 비해 전체 구간의 예측 정확도 평균이 높게 나타나고, 예측의 확실성이 강한 문서일수록 예측 정확성이 높게 나타나는 일관적인 성향을 나타내는 등 더욱 바람직한 특성을 보였다.

논문접수일 : 2013년 12월15일 게재확정일 : 2013년 12월 21일
투고유형 : 학술대회우수논문 교신저자 : 김남규

1. 서론

최근 스마트 기기의 보급 및 다양한 정보채널들의 등장으로 인해, 정보가 생성, 유통, 저장되는 양이 기하급수적으로 증가하고 있다. 이에 따라 데이터의 양 자체가 문제의 일부분이 되는 빅데이터(Big Data) 분석 기술(O'Reilly Radar Team, 2011)에 관심이 높아지고 있다. 빅데이터는 기존의 방법이나 도구로는 수집, 저장, 검색, 분석, 시각화가 어려운 정형 또는 비정형 데이터를 의미하며 (McKinsey, 2011), 방대하고, 다양한 데이터를 적시에 처리하고 분석해야 한다는 특징이 있다. Gartner 그룹이 향후 유망 기술을 예측하여 발표하는 보고서에는 빅데이터 관련 기술이 향후 2~5년 내에 IT분야에서 자리 잡을 주요 기술로 2년 연속 포함된 바 있다(Gartner, 2012). 국내

* 이 논문은 2013년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2013S1A5A2A01017304).

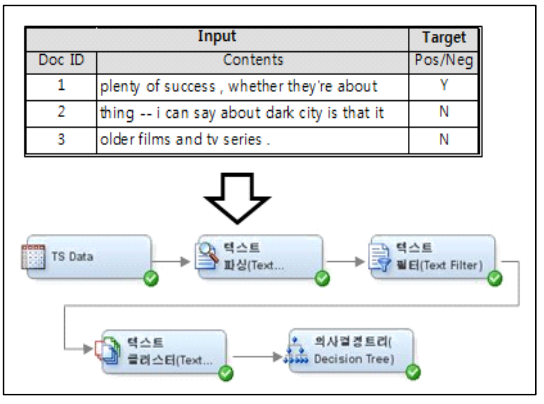
에서는 빅데이터 시장규모가 매년 70% 후반대의 고성장을 유지할 것으로 전망되고 있으며, 2015년에는 시장규모가 약 2만6300만 달러에 이를 것으로 예측되고 있다. 이처럼 빅데이터에 대한 관심은 정부, 학계, 업계 등 모든 분야에서 높아지고 있으며, 이에 따라 빅데이터 분석에 대한 수요, 활용방안, 그리고 분석 기술도 점차 다양해지며 발전하고 있다.

빅데이터에 대한 관심이 높아지고 있는 현상의 가장 큰 원인은, 스마트기기의 사용이 활성화됨에 따라 사용자가 생성하는 텍스트, 사진, 동영상과 같은 비정형 데이터의 양이 크게 증가하고 있는 것에서 찾을 수 있다. 특히 비정형 데이터 중에서도 텍스트 데이터의 경우, 사용자들의 의견 및 다양한 정보를 명확하게 표현하고 있다는 특징이 있다. 따라서 이러한 텍스트에 대한 분석을 통해 새로운 가치를 창출하고자 하는 시도가 활발히 이루어지고 있다. 예를 들어, 과거 기업은 중요한 의사결정을 내리기 위한 주요 근거 자료로 재무제표를 활용하였다. 재무제표는 기업성과 전반을 보여주는 유용한 자료이지만, 일정 시점 이전에 발생한 과거 실적을 토대로 작성되었다는 점에서 현재, 또는 가까운 미래의 의사결정을 위한 자료로 활용되기에는 부적합한 측면이 있다. 따라서 현 시점의 다양한 이슈나 의견을 실시간으로 파악하여 의사결정에 반영하기 위한 노력이 계속되었으며, 소셜미디어를 통해 실시간으로 대량 발생하는 텍스트에 대한 분석은 훌륭한 대안 중 하나로 인식되고 있다. 이처럼 텍스트 분석은 우리에게 기존의 문제 해결을 위한 다양한 가능성을 주고 있으며, 이에 따라 텍스트 분석 기술에 대한 관심은 점점 높아지고 있다.

텍스트 분석 기술을 활용한 분야 중 대표적인 것으로 오피니언 마이닝(Opinion Mining)을 들 수 있다. 오피니언 마이닝은 감성 분석(Sentiment Analysis)으로 불리기도 하며, 제품, 서비스, 조직, 개인, 이슈, 사건, 토픽, 그리고 이들의 여러 속성에 대한 사람들의 의견, 감성, 평가, 태도, 감정 등을 분석하는 일련의 과정을 의미한다(Liu, 2012). 대다수의 사람들은 효율적인 의사결정을 하기 위해 타인의 의견을 참고하곤 한다. 기업은 제품 및 서비스 개발의 방향을 결정하기 위해 소비자들의 욕구 및 불만에 대한 의견을 알고 싶어 하는데, 이를 위해 과거에는 설문지배포 및 전화상담 등 다양한 형태의 조사를 수행하였다. 하지만 최근에는 빅데이터 분석 업체를 통해 트위터 혹은 페이스북 등의 소셜미디어를 분석하는 등 소비자들의 의견을 수집하는 방법이 크게 변화했다. 소셜미디어를 통해 공개된 의견을 분석하는 방식은 과거의 전통적인 설문조사에 비해 보다 다양한 측면의 객관적인 의견을 거의 실시간으로 파악할 수 있다는 점에서, 수시로 변화하는 소비자들의 니즈에 선제적으로 대응할 수 있다는 장점을 갖는다. 뿐만 아니라 보다 적은 비용으로 많은 의견을 수집할 수 있기 때문에 그 활용가치가 매우 높다고 할 수 있다. 따라서 방대한 양의 텍스트 데이터에 담긴 사용자의 의견을 분석하기 위한 오피니언 마이닝에 대해 수요와 관심이 집중되는 것은 매우 자연스러운 현상이라고 할 수 있다.

오피니언 마이닝은 기본적으로 문서가 긍정, 부정, 또는 중립 중 어떤 견해를 갖고 있는지 판별하는 일련의 과정이라 볼 수 있으며, 분석은 각 문서 최소 단위인 어휘의 감성 극성(Sentiment Polarity)에 기반하여 이루어진다. 즉, 주요 어휘의 감성 극성이 미리 정의된 감성사전(Sentiment Lexicon)을 구축한 후, 새로 주어진 문서에 출현

한 어휘의 감성 극성에 따라 문서전체의 감성 극성을 분류하게 된다. 오피니언 마이닝은 텍스트 데이터에 대한 분석을 통해 의미있는 지식을 도출하는 기법인 텍스트 마이닝(Text Mining)과 많은 기술을 공유하고 있다. 따라서 오피니언 마이닝을 텍스트 마이닝의 하위 개념으로 인식하거나, 두 개념을 특별한 구분 없이 혼용하는 경우가 비일비재하다. 예를 들어 <Figure 1>의 경우 입력 문서에 대한 기본적인 자연어처리 및 텍스트 분석을 통해 목적 변수를 예측하는 간단한 흐름도를 보여주고 있으며, 목적 변수는 '긍정/부정'으로 주어져 있다. 이 경우 <Figure 1>의 과정은 전통적인 텍스트 마이닝의 과정에 따라 수행되었다는 점에서는 텍스트 마이닝으로 분류될 것이며, 분석의 목적이 긍정/부정을 예측하는 것이라는 점에서는 오피니언 마이닝으로 분류될 수 있을 것이다.



〈Figure 1〉 Opinion Classification Using Text Mining

이러한 모호성에도 불구하고, 텍스트 마이닝과 오피니언 마이닝은 감성사전의 사용 여부에 따라 방법론 측면에서 명확하게 구분된다. 즉 오피니언 마이닝은 미리 구축된 감성사전을 사용하는 반면, 텍스트 마이닝은 입력 데이터에 대한 학습을 통해 다른 데이터를 예측한다는 차이점을 갖는다. 이러한 관점에 따르면 문서의 분류 및 예측에 있어서 목적 변수가 긍정 또는 부정의 감성을 나타내는 경우에는, 전통적 텍스트 마이닝, 또는 감성사전 기반의 오피니언 마이닝의 두 가지 방법론에 의해 감성 예측을 수행할 수 있는 것이다. 따라서 본 연구에서는 텍스트 마이닝과 오피니언 마이닝의 목적 및 방법론 측면의 차이점을 고찰하고, 동일한 데이터에 대해 각각의 방법론에 의한 문서 분류 및 예측을 수행함으로써 각 방법론의 특징을 보다 명확히 파악하고자 한다.

2. 관련연구

텍스트 마이닝은 기존의 데이터 마이닝 프로세스에 자연어 처리 기술을 적용함으로써, 일반적인 정형 데이터뿐 아니라 비정형 텍스트 데이터까지 분석 대상을 확장한 것이다. 자연어 처리의 대상인 텍스트는 분석 목적에 따라 행렬, 계층, 벡터 등의 다양한 형태로 표현된다 (Stanvrianou et al., 2007). 기본적으로 텍스트는 벡터공간모델을 이용하여 표현되며, 각 문서에 사용된 용어(Term)의 빈도에 따라 해당 문서의 주요 단어(word) 및 주제, 특성이 요약된다. 여기서 용어의 빈도는 단순한 빈도가 아닌 TF-IDF (Term Frequency-Inverse Document Frequency) (Han and Kamber, 2011)가 많이 사용된다. TF-IDF는 여러 문서에서 자주 출현하는 일반적인 단어의 가중치를 낮게 부여하고, 특정 문서에서 출현하는 비일반적인 단어의 가중치를 높게 부

여하는 방식으로 계산된다. 이런 방식으로 각 문서는 용어 수만큼의 차원과 TF-IDF를 값으로 갖는 벡터로 표현되게 되는데, 문서 내 용어의 수가 너무 많기 때문에 SVD(Singular Value Decomposition) 등을 이용하여 차원을 축소하게 된다(Albright, 2006). 이러한 기술을 기반으로 하여 텍스트 입력을 수치 값으로 변환함으로써, 이후 과정에서 기존의 정형 데이터와 함께 텍스트 데이터에 대한 군집화(Cho and Kim, 2011; Hyun et al., 2013), 예측 등의 작업을 수행할 수 있다.

오피니언 마이닝(Dave et al., 2003)은 사용자가 다양한 매체를 통해 표출한 의견을 추출, 분류, 이해, 자산화하는 과정을 의미한다. 가장 기본적인 수준의 오피니언 마이닝은 문서 단위로 이루어지며, 이 차원의 연구(Pang et al., 2002; Turney, 2002)는 하나의 문서는 하나의 개체에 대한 하나의 감성을 표현한다는 가정 하에 수행된다. 보다 세부적 단위의 연구는 각 문장을 분석 대상으로 간주하며, 주관성 구분(Subjectivity Classification)(Wiebe et al., 1999)과 구와 절 단위의 분석을 주요 이슈로 다루고 있다. 가장 세분화된 대상에 대한 연구(Hu and Liu, 2004)는 각 개체 및 개체의 속성을 분석 단위로 한다. 이 수준의 분석은 개체의 개별 속성에 대한 감성까지 파악할 수 있다는 장점을 갖지만, 개체 인식, 개체의 부분 요소 및 속성 파악 등 많은 난제들을 포함하고 있어서 다른 두 수준의 분석에 비해 보다 난해한 것으로 인식되고 있다. 이 외에 비교급 문장(Jindal and Liu, 2006), 조건문(Narayanan et al., 2009), 풍자적 표현(Tsur et al., 2010), 문장 간 관계(Asher et al., 2008)에 대한 분석, 용어의 시맨틱을 반영한 분석(Yu et al., 2012) 등 자연어 처리에 대한 깊은 이해를 바탕으로 오피니언 마이닝의 정확도를 향상시키고자 하는 연구가 활발하게 수행되고 있지만, 분석의 가장 근간이 되는 감성 사전의 구축에 대한 연구는 그 중요성에 비해 상대적으로 덜 주목받고 있다.

문서의 감성 값은 해당 문서가 포함하고 있는 감성어(Sentiment word)의 종류 및 빈도수에 의해 결정되기 때문에, 감성어는 문서에서 감성을 추출하기 위한 가장 중요한 단서로 작용한다. 이러한 감성어 또는 감성어구(Sentiment Phrase)가 갖는 감성 값을 정의해 둔 것을 감성 사전이라 하며, 감성 사전은 명확한 감성을 갖는 시드(Seed) 어휘에 감성 값을 부여하고 이를 기본으로 다른 어휘의 감성 값을 추가 도출하는 방식으로 구축된다. 감성 사전 구축 방법론은 크게 사전기반 접근법(Hu and Liu, 2004; Kamps et al., 2004; Kim and Hovy, 2004)과 말뭉치기반 접근법(Hazivassiloglou and McKeown, 1997; Ding et al., 2009)으로 분류된다. 사전기반 접근법은 WordNet 등의 사전에서 나타난 시드 어휘와 다른 어휘들 간의 유사성 및 거리 관계를 통해 감성 사전을 구축하는 방식이다. 말뭉치기반 접근법은 실제로 수집된 문장들에 대한 구문 분석을 통해 감성 사전을 구축하는 방식이다. 또한 이러한 두 가지 전통적 접근법에 의해 구축된 감성 사전은 동일 어휘라도 사용되는 상황이나 목적에 따라 상이한 감성 값을 갖는 현상을 정확하게 반영하지 못한다는 한 계를 갖는다는 부작용을 극복하기 위해, 분석하고자 하는 주제 및 목적에 따라서 특화된 목적지향 감성사전을 구축하고자 하는 시도(Yu et al., 2013)도 최근 이루어진 바 있다.

3. 텍스트 마이닝 기반 오피니언 분류 모델과 오피니언 마이닝 기반 예측 모델의 비교

3.1 텍스트 마이닝과 오피니언 마이닝의 특성 비교

텍스트 마이닝과 오피니언 마이닝은 모두 텍스트 데이터를 입력 데이터로 사용할 뿐 아니라, 파싱, 필터링 등 자연어 처리 기술을 사용한다는 측면에서 많은 공통점을 갖고 있다. 일반적으로는 텍스트 마이닝이 보다 넓은 의미의 텍스트 분석 기법으로 인식되고 있으며, 이 중 감성사전을 이용하여 문서의 감성을 분류하는 특수한 분석 기법을 오피니언 마이닝으로 이해하고 있다. 하지만 두 기법은 목적, 산출물, 주요 모듈 측면에서 <Table 1>과 같은 차이를 보이므로, 서로 구분하여 이해할 필요가 있다.

〈Table 1〉 Comparison of Opinion mining and Text mining

	Text Mining	Opinion Mining
Technique	Parsing, Filtering and Natural Language Processing Techniques.	
Purpose	Through the analysis of text data who get the knowledge that is meaningful.	The classification sentiment of text document. (Positive or Negative)
Input	Text document.	
Output	Topic extraction, Document clustering, Document classification Etc.	Sentiment identifying the document.
Main Module	TF-IDF, SVD	Create the Sentiment Laxicon and refer.

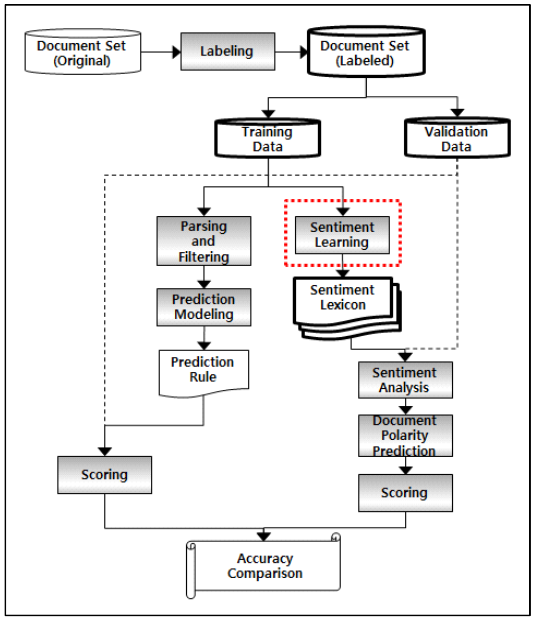
<Table 1>에서 나타난 바와 같이 텍스트 마이닝은 텍스트 데이터 분석을 통해 의미있는 지식을 창출하는 과정 및 기술을 의미하며, 이를 통해 도출되는 결과물은 토픽, 문서 클러스터, 문서 분류 등이 있다. 한편 오피니언 마이닝은 문서의 감성을 식별하는 한정된 목적으로 사용되며, 본질적으로는 문서 분류의 특수한 경우로 이해될 수 있다. 텍스트 마이닝의 근간을 이루는 주요 모듈은 TF-IDF에 근거한 문서간 유사도 계산, 그리고 SVD를 활용한 차원 축소 과정을 들 수 있다. 한편 오피니언 마이닝은 각 용어별로 감성값을 부여하는 감성사전 구축, 그리고 구축된 감성사전을 활용하여 특정 문서 전체의 감성값을 계산하는 문서 극성 계산이 주요 모듈을 이루고 있다.

하지만 <Table 1>에 제시된 분류 기준은 두 기법을 배타적으로 구분한다기 보다는, 관점에 따라 동일 분석을 서로 다른 기법으로 인식하게 되는 혼란을 야기한다. 예를 들면 영화에 대한 감상평과, 해당 감상평이 긍정인지 부정인지를 나타내는 목적 변수가 존재한다고 하자. 이에 대해 전통적인 텍스트 마이닝 기법에 근거하여 각 문서별 SVD값 또는 토픽을 도출하고, 이를 활용하여 목적 변수를 예측하는 의사결정나무 또는 인공신경망을 도출했다고 가정하자. 이렇게 도출된 모형은 새로운 입력 문서(감상평)에 대해 목적 변수(긍정/부정)를 예측하는 데 사용될 수 있다. 이 경우 이러한 분석은 전통적 텍스트 마이닝 기법을 사용했지만 예측하고자 하는 목적 변수가 긍정/부정의 감성이라는 측면에서, <Table 1>의 분류에서 텍스트 마이닝의 성격과 오피니언 마이닝의 성격을 동시에 갖는다. 이러한 예와 같이 분석의 목적 또는 분석의 결과물 관점에서 두 기법을 구분하는 것은 서로 중첩(overlap)되는 영역을 가지므로 두 기법 구분의 명확한 기준이 되기 어렵다.

따라서 본 연구에서는 텍스트 마이닝과 오피니언 마이닝의 구분을 위해 주요 모듈인 감성사전 활용 여부에 주목하고자 한다. 즉 분석의 결과물 및 분석 목적에 관계없이 감성사전의 활용 여부에 따라 두 가지의 방법론을 구분하고, 각 방법론이 예측 정확도 측면에서 어떠한 차이를 보이는지 실험을 통해 규명하고자 한다. 보다 구체적으로는 긍정/부정 분류를 목적으로 하는 동일한 데이터 셋에 대해, 감성사전을 활용하지 않는 전통적인 텍스트 마이닝 기반 예측 모델과 감성사전 기반의 오피니언 마이닝 모델을 각각 구축하고 이들의 성능을 비교해 보고자 한다.

3.2 오피니언 분류의 감성사전 활용효과 측정을 위한 연구 모형

본 연구에서 진행하고자 하는 전체 연구 모형은 <Figure 2>에 제시되어 있다. 원통형은 데이터를, 직사각형은 프로세스를 나타낸다. 그 외의 도형은 중간 또는 최종 산출물을 의미한다.



〈Figure 2〉 Experimental Model

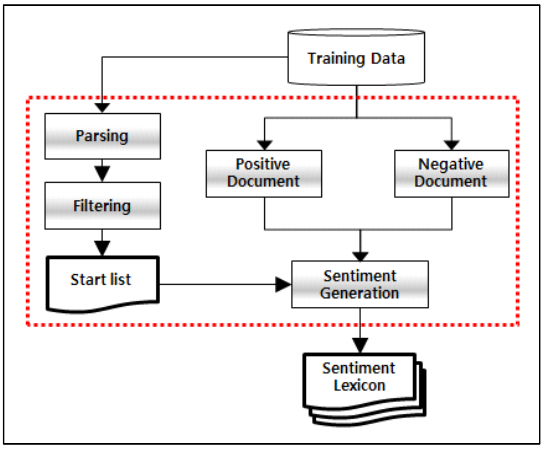
<Figure 2>에서 Labeling은 원 문서에 목적변수를 부여하는 과정을 의미한다. 목적변수가 부여된 문서들은 각각 학습(Training) 데이터와 검증(Validation) 데이터로 나누어진다. 학습 데이터 하단의 좌측 프로세스는 텍스트 마이닝을, 우측 프로세스는 오피니언 마이닝을 의미한다. 먼저 텍스트 마이닝 프로세스를 살펴보면, 파싱, 필터링을 통해 문서를 단어들로 분해하고, 해당 단어의 품사(Role)를 결정한다. 그리고 이에 대한 토픽 분석 및 SVD값 도출을 통해 의사결정나무 또는 인공신경망 등의 예측 모델을 생성한다. 예측 모델에 따라 규칙을 도출하고, 이 규칙을 검증 데이터에 적용한 결과의 정확도를 스코어링 프로세스에서 도출한다. 한편, 오피니언 마이닝 프로세스는 점선 사각형으로 구분된 감성학습(Sentiment Learning) 과정을 통해 감성사전(Sentiment Lexicon)을 구축한다. 감성학습 과정은 이후 절에서 보다 자세히 설명한다. 구축된 감성사전에 따라 검증 데이터의 각 문서의 문서별 감성 극성(Document Polarity)을 도출하며, 도출된 감성 극성을 실제 값과 비교함으로써 정확도를 계산한다. 문서의 감성 극성 도출과정 역시 이후 절에서 자세히 소개한다. 마지막으로 정확성 비교(Accuracy Comparison) 단계에서는 텍스트 마이닝과 오피니언 마이닝의 예측력을 비교한다.

3.3 감성사전 구축 및 문서 감성 극성 도출

본 연구에서 도입한 위한 감성사전 구축 모형은 <Figure 3>과 같다. 감성사전은 이미 구축된 범용 감성사전을 도입하는 방식과, 연구 목적에

따른 특수 감성사전을 직접 구축하는 방식으로 크게 구분될 수 있다. 하지만 범용 감성사전을 사용하는 방식은, 사용한 감성사전의 품질에 따라 오피니언 마이닝 기반 예측 정확성이 크게 좌우될 수 있으므로, 텍스트 마이닝과 오피니언 마이닝의 예측 정확성을 비교하기 위한 본 연구에는 적합하지 않은 것으로 판단된다. 따라서 본 연구에서는 주어진 학습 데이터에 근거하여 본 주제에 특화된 감성사전을 구축하고, 이를 검증 데이터의 감성 분류에 사용함으로써 두 방법론의 차이를 보다 공정하게 평가하고자 한다.

<Figure 3>의 좌측 프로세스는 학습 데이터에 대한 기본적인 자연어 처리를 통해 관심 용어(Start list)를 생성하는 과정을 나타낸다. 또한 감성사전 생성(Sentiment Generation) 프로세스는 이러한 관심 용어가 긍정 문서와 부정 문서에서 각각 출현한 회수를 기반으로 각 용어의 감성 지수(Sentiment Score)를 계산하는 과정을 의미한다. 감성 지수는 <Figure 4>의 식에 의해 계산되며, 용어와 감성 지수의 쌍의 집합이 감성사전(Sentiment Dictionary)에 저장된다.



〈Figure 3〉 Sentiment Lexicon Construction

t = Term
$Count_p(t)$ = Number of positive documents, including terms t
$Count_n(t)$ = Number of negative documents, including terms t

$$\text{Sentiment Score}(t) = \frac{[Count_p(t) - Count_n(t)]}{\max[Count_p(t), Count_n(t)]}$$

〈Figure 4〉 Sentiment Score of Each Term

<Figure 4>의 식에서 각 용어의 감성 지수는 최소 -1.0에서 최대 +1.0까지의 값을 갖게 된다. 이 때 +0.3 ~ +1.0 사이의 값을 갖는 용어는 긍정 용어로, -1.0 ~ -0.3 사이의 값을 갖는 용어는 부정 용어로 인식한다. 한편 -0.3 ~ +0.3 사이의 값을 갖는 용어는 중성적 성향으로 인해 결과를 왜곡시킬 우려가 있으므로 사전에 포함시키지 않았다.

감성사전을 이용하여 각 문서의 감성 극성(Sentiment Polarity)을 도출하는 식이 <Figure 5>에 나타나있다. 즉 문서 전체의 감성지수는 문서에 포함된 용어의 감성지수의 합으로 계산되며, 이 값이 0보다 큰 경우 해당 문서의 감성 극성은 긍정으로, 그렇지 않은 경우 부정으로 판별된다. 물론 보다 정교한 감성 극성 방안이 많은 연구에서 소개되었으나, 본 연구의 목적은 동일 데이터에 대한 텍스트 마이닝과 오피니언 마이닝의 분석 과정 및 결과를 비교하는 것이므로 가장 간략한 모델을 사용한다.

T = All term sets contained in the document d
Sentiment Score(t) = Sentiment value of term t

$$\text{Sentiment } Polarity(d) = \begin{cases} Positive & if \sum_{i=1}^{T} Sentiment\ Score(ti) > 0 \\ \text{Negative} & \text{Otherwise} \end{cases}$$

〈Figure 5〉 Sentiment Polarity of Each Document

4. 실험

4.1 데이터 소개

두 가지 방법론의 분석 과정 및 결과를 객관적으로 비교하기 위해, 동일한 데이터에 대한 각 방법론의 예측 정확성을 비교하는 실험이 반드시 필요하다. 이러한 용도의 데이터 중 널리 알려진 것으로 Internet Movie Database(IMDb)에서 제공하는 영화 리뷰 데이터를 들 수 있다 (available at: http://www.cs.cornell.edu/people/pabo/movie-review-data/). IMDb는 아카이브를 통해 42,000개 이상의 영화 리뷰 글을 평점과 함께 제공하고 있으며, 많은 연구가 이 데이터를 직접 이용하거나 가공된 형태의 데이터를 이용하여 제안 방법론의 예측 정확성을 평가하고 있다. 가공된 데이터로는 Pang and Lee(2004)가 제공하는 Movie Review Data가 자주 활용되고 있다. 해당 사이트는 IMDb로부터 수집한 데이터를 원본 형태뿐 아니라 감성 극성 분석, 감성 값 분석, 그리고 주관성 분석을 위한 형태로 가공한 데이터를 제공하고 있다. 이 데이터는 20개 이내의 리뷰를 작성한 312명의 리뷰 총 2,000건으로 구성되어 있으며, 별점(Star Ranking)을 이용하여 긍정적 리뷰인지 부정적 리뷰인지를 나타내고 있다. 이렇게 분류된 문서 수는 긍정 문서 1,000개, 부정 문서 1,000개이다. 긍정과 부정으로 분류된 리뷰는 긍정과 부정 각각 1,000개이다<Figure 6>.

DocNum	Contents	Target
1	had plenty of success, whether they're	P
2	along from a suspect studio, with every	P
3	deserves to .	P
~	~	~
1998	want to see what mick jagger looks like	N
1999	citizens, all get together and stop jim	N
2000	thing -- I can say about dark city is that	N

〈Figure 6〉 Experiment Data

4.2 텍스트 마이닝과 오피니언 마이닝의 예측력 비교 실험

실험을 위한 전체 과정은 <Figure 2>의 전체 연구 모형, <Figure 3>의 감성사전 구축 모형에 따른다. 실험을 위한 학습 데이터는 긍정 문서 500건, 부정 문서 500건으로 구성되었으며, 검증 데이터 역시 같은 수로 구성되었다. 텍스트 마이닝을 통한 오피니언 분류의 예측력은 SAS Enterprise Miner 7.1을 통해 측정하였으며, 파싱, 필터링, 클러스터링, 의사결정나무모델링의 순서로 분석이 진행되었다. 마지막으로 스코어링을 통해 이렇게 생성된 의사결정나무모형을 검증 데이터에 적용하여 예측력을 측정했다.

학습 데이터로부터 오피니언 마이닝의 분석에 필요한 감성사전을 구축하였으며, 그 결과의 일부가 <Figure 7>에 제시되어 있다. 감성 사전은 용어, 품사, 긍정 문서에서의 출현 회수, 부정 문서에서의 출현 회수, 감성 지수로 구성되어 있다.

또한 학습 데이터를 통해 구축된 감성사전을

Term	Role	Count (Positive)	Count (Negative)	Score
unfortunate	Noun	10	0	1
subtle	Adj	88	11	0.875
lord	Noun	48	6	0.875
interests	Noun	8	1	0.875
fake	Adj	8	15	-0.46667
trash	Noun	4	15	-0.73333
stupid	Adj	10	94	-0.89362

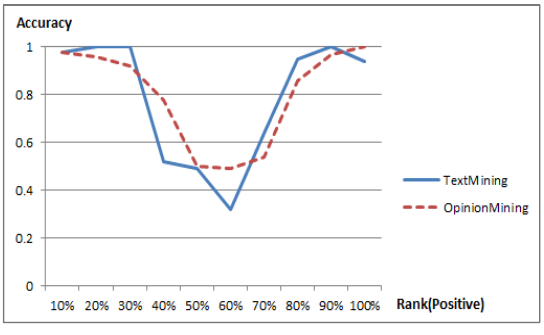
〈Figure 7〉 Constructed Sentiment Lexicon (Part)

검증 데이터의 문서에 적용하여 <Figure 5>의 식에 따라 각 문서의 감성 극성을 도출하였으며, 실제 감성 극성과 예측 감성 극성의 일치 여부에 따른 정확도를 계산하였다.

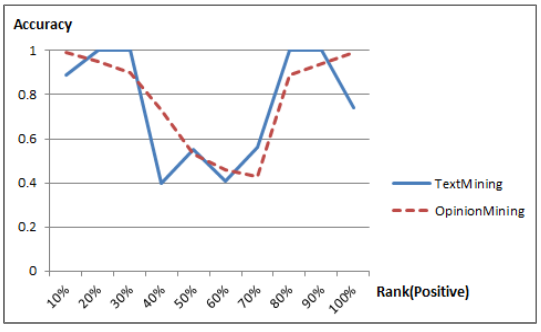
4.3 결과 분석

초기 실험은 전체 품사의 용어를 대상으로 수행하였으며, 예측 순위별 정확도 비교 결과가 <Figure 8>에 요약되어 있다. 그래프에서 x축은 각 방법론에 따른 긍정 예측 순위(비누적)를 보여주며, y축은 해당 순위 내 문서의 예측 정확도를 보여준다. 예를 들어 20% 구간은 감성 지수의 값이 상위 10% ~ 20%에 속하는 문서들의 집합을 나타내며, y축은 이 문서들에 대한 예측 정확도를 나타낸다. 즉 좌측과 우측은 각각 긍정/부정 지수가 높은 문서를 나타내며, 중앙은 중립 성향이 강한 문서를 나타낸다. 전반적으로 긍정/부정 성향이 강한 문서일수록 중립 성향의 문서에 비해 예측 정확도가 높게 나타남을 볼 수 있다.

한편 주요 품사(부사, 형용사, 명사)만을 사용하여 긍정 예측 순위별 정확도를 분석한 결과는 <Figure 9>에 나타나있다. 이 경우 역시 긍정/부정 성향의 문서가 중립 성향의 문서에 비해 높은 예측 정확도를 보임을 알 수 있다. 또한 <Figure 8>과 <Figure 9>에서 텍스트 마이닝의 예측 결과가 다소 불규칙한 형태를 보이는 것에 비해, 오피니언 마이닝의 예측 결과는 매끄러운 곡선 형태로 나타남을 알 수있다. <Figure 9>의 예측 정확도를 수치로 나타낸 표가 <Table 2>에 나타나 있다. 분석 결과 오니피언 마이닝이 텍스트 마이



〈Figure 8〉 Classification Accuracy with Positive Rank Varied



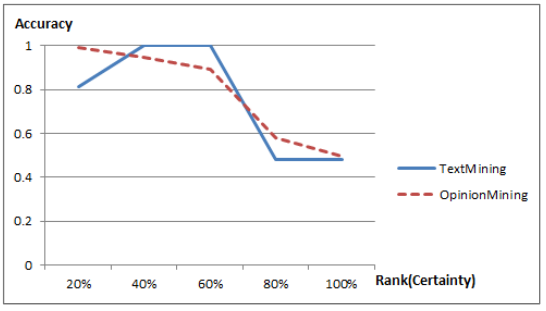
〈Figure 9〉 Classification Accuracy with Positive Rank Varied (for Main Parts of Speech)

〈Table 2〉 Classification Accuracy with Positive Rank Varied (for Main Parts of Speech)

Adj,Adv,Noun	Text Mining	Opinion Mining
10%	0.89	0.99
20%	1	0.95
30%	1	0.9
40%	0.4	0.73
50%	0.55	0.53
60%	0.41	0.46
70%	0.56	0.43
80%	1	0.89
90%	1	0.94
100%	0.74	0.99
Average	0.755	0.781

닝에 비해 보다 바람직한 특성을 나타내는 것으로 파악된다. 이러한 판단 근거는 문서 전체에 대한 예측력 평균이 오피니언 마이닝이 더 높게 나타난다는 점, 그리고 긍정/부정 성향이 강하게 나타나는 문서에서의 예측력이 오피니언 마이닝이 더 높게 나타난다는 점에서 찾을 수 있다.

<Figure 9>는 긍정 예측 순위에 따른 정확도를 보이고 있지만, 이를 통해 확실성(Certainty)을 비교하기에는 어려움이 있다. 따라서 긍정 예측 순위가 아닌 확실성 순위에 따라 <Figure 9>를 재구성한 그래프가 <Figure 10>에 나타나있다.



〈Figure 10〉 Classification Accuracy with Certainty Varied (for Main Parts of Speech)

<Figure 10>에서 x축은 예측의 확실성에 따른 순위, 즉 문서의 감성 극성 예측값의 절대값의 크기에 따른 순위를 나타낸다. 따라서 확실성이 높은 좌측 영역에 속하는 문서가 높은 예측력을 나타낼수록 바람직한 특성을 갖는 예측 모델이라고 할 수 있다. <Figure 10>에서 오피니언 마이닝의 경우 뚜렷한 우하향 형태를 보이는 반면 텍스트 마이닝의 경우 다소 불규칙한 형태를 보이고 있다. 따라서 확실성 측면에서도 오피니언 마이닝이 텍스트 마이닝에 비해 보다 바람직한 결과를 나타냄을 알 수 있다. <Table 3>은 <Figure 10>에 대한 수치를 보이고 있다.

〈Table 3〉 Classification Accuracy with Certainty Varied (for Main Parts of Speech)

Adj,Adv,Noun	Text Mining	Opinion Mining
20%	0.815	0.99
40%	1	0.945
60%	1	0.895
80%	0.48	0.58
100%	0.48	0.495
Average	0.755	0.781

본 실험의 의의는 오피니언 분류라는 동일한 목적 하에서 영화 리뷰 데이터라는 동일한 데이터에 대해 오피니언 마이닝과 텍스트 마이닝의 두 가지 분석을 수행함으로써, 두 방법론의 분석 과정과 결과를 비교해 보는 데 있다. 실험 결과를 요약하면 다음과 같다. 우선 전체 구간의 예측 정확도 평균은 오피니언 마이닝이 다소 높게 나타났다. 또한 긍정/부정 극성이 강할수록, 즉 확실성이 강할수록 예측 정확성이 높게 나타나는 성향은 오피니언 마이닝에서 더욱 뚜렷하게 나타났다. 결론적으로 본 실험에서는 예측 정확성 측면에서 오피니언 마이닝이 텍스트 마이닝에 비해 보다 바람직한 특징을 나타냈다.

5. 결론

텍스트 마이닝과 오피니언 마이닝은 모두 텍스트 데이터를 입력 데이터로 사용할 뿐 아니라 파싱, 필터링 등 자연어 처리 기술을 사용한다는 측면에서 많은 공통점을 갖고 있다. 특히 문서의 분류 및 예측에 있어서 목적 변수가 긍정 또는 부정의 감성을 나타내는 경우에는, 전통적 텍스트 마이닝, 또는 감성사전 기반의 오피니언 마이

닝의 두 가지 방법론에 의해 오피니언 분류를 수행할 수 있다. 따라서 텍스트 마이닝과 오피니언 마이닝의 특징을 구분하는 가장 명확한 기준은 입력 데이터의 형태, 분석의 목적, 분석의 결과물이 아닌 감성사전의 사용 여부라고 할 수 있다. 따라서 본 연구에서는 오피니언 분류라는 동일한 목적에 대해 텍스트 마이닝과 오피니언 마이닝을 각각 사용하여 예측 모델을 수립하는 과정을 비교하고, 결과로 도출된 모델의 예측 정확도를 비교하였다.

실험 결과 오피니언 마이닝을 통해 수립된 모델이 텍스트 마이닝 모델에 비해 전체 구간의 예측 정확도 평균이 높게 나타나고, 예측의 확실성이 강한 문서일수록 예측 정확성이 높게 나타나는 일관적인 성향을 나타내는 등 더욱 바람직한 특성을 보였다. 또한 오피니언 마이닝은 한 번 구축된 사전을 유사 도메인에서 다시 활용할 수 있으므로 규칙 학습에 소요되는 시간을 줄일 수 있다는 장점과, 특정 문서가 특정 감성으로 분리된 이유를 감성 사전을 통해 설명할 수 있다는 설명력 측면의 장점을 갖는다.

하지만 본 연구에서 수행한 실험의 결과를 확대 해석하기에는 무리가 따른다. 우선 실험 데이터가 영화 리뷰 데이터로 한정되어 있다. 또한 텍스트 마이닝 과정에서 파싱, 필터링 등의 단계에서 다양한 파라메터의 설정이 가능하다. 물론 오피니언 마이닝 과정에서도 감성사전 구축, 그리고 문서의 감성 극성 판별 과정에서 수많은 판단이 존재한다. 예를 들면 감성사전을 구축할 때, 중립 성향의 단어로 분류되어 감성사전에서 제외되는 감성 지수의 범위를 어떻게 설정할 것인지, 감성사전이 Unigram만 포함할 것인지, Bigram, Trigram까지 포함할 것인지 등 다양한 판단이 존재하게 되며, 이러한 판단은 예측 성능에 영향을 끼칠 수 있다. 그럼에도 불구하고 본 연구는 오피니언 분류라는 동일한 목적을 위해 텍스트 마이닝과 오피니언 마이닝의 두 가지 분석을 수행함으로써, 두 방법론의 분석 과정과 결과의 차이를 비교한 연구라는 점에서 의의가 인정될 수 있다. 향후 실험 데이터의 확대, 실험의 정밀성 및 다양성 향상 등을 통해 두 방법론에 대한 보다 엄밀한 평가가 이루어져야 할 것이다.

참고문헌

Albright, R., *Taming Text with the SVD*, SAS Institute Inc., 2006.

Asher, N., F. Benamara, and Y. Y. Mathieu, “Distilling Opinion in Discourse: A Preliminary Study,” *Proceedings of the International Conference on Computational Linguistics*, (2008), 7-10.

Cho, I. and N. Kim, "Recommending Core and Connecting Keywords of Research Area Using Social Network and Data Mining Techniques," *Journal of Intelligence and Information Systems*, Vol. 17, No. 1(2011), 127-138.

Dave, K., S. Lawrence, and D. M. Pennock, “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,” *Proceedings of International Conference on World Wide Web*, (2003), 519-528.

Ding, X., B. Liu, and L. Zhang, “Entity Discovery and Assignment for Opinion Mining Applications,” *Proceedings of ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, (2009), 1125-1134.

Gartner Inc., *2012 Hype Cycle for Emerging Technologies*, Gartner Inc., 2012.

Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd edition., Morgan Kaufmann Publishers, 2011.

Hazivassiloglou, V. and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives," *Proceedings of Annual Meeting of the Association for Computational Linguistics*, (1997), 174-181.

Hu, M. and B. Liu, "Mining and Summarizing Customer Reviews," *Proceddings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2004).

Hyun, Y., H. Han, H. Choi, J. Park, K. Lee, K. Kwahk, and N. Kim, "Methodology Using Text Analysis for Packaging R&D Information Services on Pending National Issues," *Journal of Information Technology Applications & Management*, Vol. 20, No. 3(2013), 231-257.

Jindal, N. and B. Liu, "Mining Compareative Sentences and Relations," *Proceeding of National Conference on Artificial Intelligence*, Vol. 2(2006), 1331-1336.

Kamps, J., M. Marx, R. J. Mokken, and M. D. Rijke, "Using WordNet to Measure Semantic Orientation of Adjectives," *Proceedings of International Conference on Language Resources and Evaluation*, Vol. 4(2004), 1115-1118.

Kim, S. and E. Hovy, "Determining the sentiment of Opinions," *Proceedings of International Conference on Computational Linguistics*, No. 1367(2004).

Liu, B., *Sentiment Analysis and Opinion Mining*, Morgan and Claypool Publishers, 2012.

McKinsey Global Institute, *Big Data: The next Frontier for Innovation, Competition, and Productivity*, McKinsey and Company, 2011.

Narayanan, R., B. Liu, and A. Choudhary, "Sentiment Analysis of Conditional Sentences," *Proceeding of Conference on Empirical Methods in Natural Language Processing*, Vol. 1(2009), 180-189.

O'Reilly Radar Team, *Big Data Now: Current Perspectives from O'Reilly Radar*, O'Reilly, 2011.

Pang, B., L. Lee, and S. Vaithyanathan, "Thumbs Up?:Sentiment Classification using Machine Learning Techniques," *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Vol. 10(2002), 79-86.

Stanvrianou, A., P. Andritsos, and N. Nicoloyannis, "Overview and Semantic Issues of Text Mining", *ACM SIGMOD Record*, Vol. 36, No. 3(2007), 23-34.

Tsur, O., D. Davidov, and A. Rappoport, "A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews," *Proceedings of the International AAAI Conference on Weblogs and Social Media*, (2010), 162-169.

Turney, P. D., "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proceedings of Annual Meeting of the Association for computational Linguistics*, (2002), 417-424.

Wiebe, J., R. F. Bruce, and T. P. O'Hara, "Development and Use of a Gold-Standard Data Set for Subjectivity Classifications," *Proceedings of the Association for*

Computational Linguistics, (1999), 246-253.

Yu, E., J. Kim, C. Lee, and N. Kim, "Using Ontologies for Semantic Text Mining," *The Journal of Information Systems*, Vol. 21, No. 3(2012), 137-161.

Yu, E., Y. Kim, N. Kim, and S. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligence and Information Systems*, Vol. 19, No. 1(2013), 95-110.

Abstract

A Study on the Effect of Using Sentiment Lexicon in Opinion Classification

Seungwoo Kim* · Namgyu Kim**

Recently, with the advent of various information channels, the number of has continued to grow. The main cause of this phenomenon can be found in the significant increase of unstructured data, as the use of smart devices enables users to create data in the form of text, audio, images, and video. In various types of unstructured data, the user's opinion and a variety of information is clearly expressed in text data such as news, reports, papers, and various articles. Thus, active attempts have been made to create new value by analyzing these texts.

The representative techniques used in text analysis are text mining and opinion mining. These share certain important characteristics; for example, they not only use text documents as input data, but also use many natural language processing techniques such as filtering and parsing. Therefore, opinion mining is usually recognized as a sub-concept of text mining, or, in many cases, the two terms are used interchangeably in the literature. Suppose that the purpose of a certain classification analysis is to predict a positive or negative opinion contained in some documents. If we focus on the classification process, the analysis can be regarded as a traditional text mining case. However, if we observe that the target of the analysis is a positive or negative opinion, the analysis can be regarded as a typical example of opinion mining. In other words, two methods (i.e., text mining and opinion mining) are available for opinion classification. Thus, in order to distinguish between the two, a precise definition of each method is needed. In this paper, we found that it is very difficult to distinguish between the two methods clearly with respect to the purpose of analysis and the type of results.

We conclude that the most definitive criterion to distinguish text mining from opinion mining is whether an analysis utilizes any kind of sentiment lexicon. We first established two prediction models, one based on opinion mining and the other on text mining. Next, we compared the main processes used by

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-5209, E-mail: ngkim@kookmin.ac.kr

the two prediction models. Finally, we compared their prediction accuracy. We then analyzed 2,000 movie reviews. The results revealed that the prediction model based on opinion mining showed higher average prediction accuracy compared to the text mining model. Moreover, in the lift chart generated by the opinion mining based model, the prediction accuracy for the documents with strong certainty was higher than that for the documents with weak certainty. Most of all, opinion mining has a meaningful advantage in that it can reduce learning time dramatically, because a sentiment lexicon generated once can be reused in a similar application domain. Additionally, the classification results can be clearly explained by using a sentiment lexicon.

This study has two limitations. First, the results of the experiments cannot be generalized, mainly because the experiment is limited to a small number of movie reviews. Additionally, various parameters in the parsing and filtering steps of the text mining may have affected the accuracy of the prediction models. However, this research contributes a performance and comparison of text mining analysis and opinion mining analysis for opinion classification. In future research, a more precise evaluation of the two methods should be made through intensive experiments.

Key Words : Sentiment Lexicon, BigData Analysis, Opinion Mining, Text Mining

저 자 소 개

김 승 우
그리스도대학교 경영정보학부에서 학사 학위를 취득하였으며, 현재 국민대학교 비즈니스IT전문대학원 비즈니스IT전공 석사 과정에 재학 중이다. 주요 관심분야는 텍스트 마이닝, 오피니언 마이닝, 데이터 마이닝, 데이터베이스 등이다.

김 남 규
현재 국민대학교 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학회 이사, JITAM 편집위원을 역임하였으며, 한국경영정보학회, 한국지능정보시스템학회, 한국정보시스템학회 종신회원 및 한국생산성본부 자문위원으로 활동 중이다. 주요 관심분야는 시맨틱 데이터 관리, 데이터베이스 설계 및 데이터 마이닝 등이다.