「정보시스템연구」 제28권 제1호
한국정보시스템학회
2019년 3월, pp. 01~16
http://dx.doi.org/10.5859/KAIS.2019.28.1.1

온라인리뷰의 랭킹모델링을 위한 양과 질의 인과모형 분석*

이창용** · 김근형***

〈목 차〉

Ⅰ. 서 론

웹2.0 시대를 맞아 소셜네트워크서비스가 보편적으로 사용되고 있으며 이를 통하여 많은 온라인리뷰들이 생성되고 활용되고 있다. 온라인리뷰란 인터넷 또는 소셜네트워크 상에서 자신이 경험한 상품, 서비스에 관련된 의견이나 감성 등을 텍스트 형식으로 다른 소비자에게 전하는 소비자의 경험 정보로서, 사용후기, 구매후기, 댓글 등 여러 용어로 지칭되는 온라인 구전의 일종이다(박은주, 정유진, 2013). 온라인리뷰는 리뷰작성자들의 상품 및 서비스의 소비에 대한 경험담이나 의견 등을 포함하며 다른 잠재적 소비자들의 소비행위에 영향을 미친다. 온라인리뷰에 대한 선행 연구 결과들에 의하면, 리뷰의 정보 특성에 따라 소비자행동이 달라지고(Dwayne and Kevin, 2001), 정보 특성의 방향성(긍정/부정)이 태도 또는 구매 의도에 영향을 미쳤으며(Peterson and Meria, 2003), 평가적인 정보 특성(객관성/주관성)이 리뷰 수용자에게 영향을 미치는 것으로 나타났다(Jeon and Jung, 2006). 결국, 특정 상품 및 서비스나 속성에 대한 온라인리뷰의 긍정 또는 부정 정

* 이 논문은 2018학년도 제주대학교 교원성과지원사업에 의하여 연구되었음.
** HB네트웍스, ioptimus2012@naver.com(주저자)
*** 제주대학교 경영정보학과, khkim@jejunu.ac.kr(교신저자)

도에 대한 정확한 분석과 이에 기반한 랭킹은 소비자들에게는 매우 중요한 정보가 될 수 있다.

온라인리뷰를 작성할 때 작성자는 상품 및 서비스에 대한 경험에 대하여 주관적이고 감성적인 의견을 텍스트 형태로 표현할 뿐만 아니라 정량적인 평가점수도 함께 매긴다. 즉, 온라인리뷰에는 작성자가 경험한 상품 및 서비스에 대한 평가내용이 포함되어 있으며 그 평가내용에는 텍스트 형태의 정성적인 관점과 평가점수 형태의 정량적인 관점이 동시에 표출되어 있다. 일반적으로, 텍스트는 명사와 동사, 형용사가 포함되어 있는 자연어 형태이며 평가점수는 숫자형태이다.

온라인리뷰와 관련한 많은 연구들은 리뷰의 텍스트를 분석하기 위하여 자연어처리 기반의 분석기술의 개선이 주류를 이루어 왔다(윤홍준,2010; 장재영,2009;홍태호,2014). 작성자의 의견이 정성적으로 표현된 텍스트에서 상품 및 서비스의 속성에 대응하는 명사를 추출하고 그 속성의 긍정이나 부정정도를 나타내는 형용사 또는 동사를 보다 정확히 추출하기 위해서는 자연어처리 기술이 중요한 역할을 한다. 자연어처리 기술의 혁신적인 발전 없이는 텍스트 분석의 정확성을 높이는 것도 한계가 있을 수밖에 없다. 따라서, 온라인리뷰의 텍스트를 분석할 때 자연어처리 기술뿐만 아니라 다른 보완적인 방법으로 텍스트 분석의 정확성을 높일 수 있는 방안이 필요하다.

온라인리뷰에는 텍스트뿐만 아니라 평가점수와 리뷰길이 등과 같은 정량적인 속성도 존재한다. 온라인리뷰의 평가점수 및 리뷰길이 등이 텍스트에서 표현하고자 하는 정성적 의견과 유의미한 상관관계를 가진다면, 자연어처리 기술과는 다른 방법으로 텍스트 분석의 정확성을 높일 수 있는 보완적인 데이터로 활용될 수 있다. 자연어처리 기반의 텍스트 분석이 정성적 관점의 분석 방법이라면, 리뷰길이 및 평가점수를 고려하여 분석하는 것은 정량적 관점의 분석방법이라 할 수 있다. 현재까지의 연구현황을 보았을 때 온라인리뷰의 분석과정에서 정성적 관점과 정량적 관점을 동시에 고려하여 분석하는 연구는 전무한 실정이다.

본 연구에서는 온라인리뷰를 보다 정확하고 효과적으로 분석할 수 있는 방안을 제안한다. 온라인리뷰의 정성적 데이터인 텍스트뿐만 아니라 정량적 데이터인 평가점수 및 리뷰길이를 활용하여 온라인리뷰를 분석할 수 있는 랭킹 분석모델을 제안한다. 본 연구에서는 첫째, 온라인리뷰의 텍스트에 대한 자연어처리 기반의 분석결과와 리뷰길이 및 평가점수와의 회귀적 관계를 가설검증 형태로 고찰하며 둘째, 이를 기반으로 온라인리뷰를 보다 정확하게 평가할 수 있는 가중치 기반의 랭킹모델링에 대한 방안을 제시한다. 본 논문에서 제안하는 랭킹 분석모델은 온라인리뷰의 정성적 관점과 정량적 관점을 포괄하여 기존의 오피니언마이닝에 의한 분석모델을 더 일반화시킬 수 있다는 점에서 그 가치와 의의가 있다.

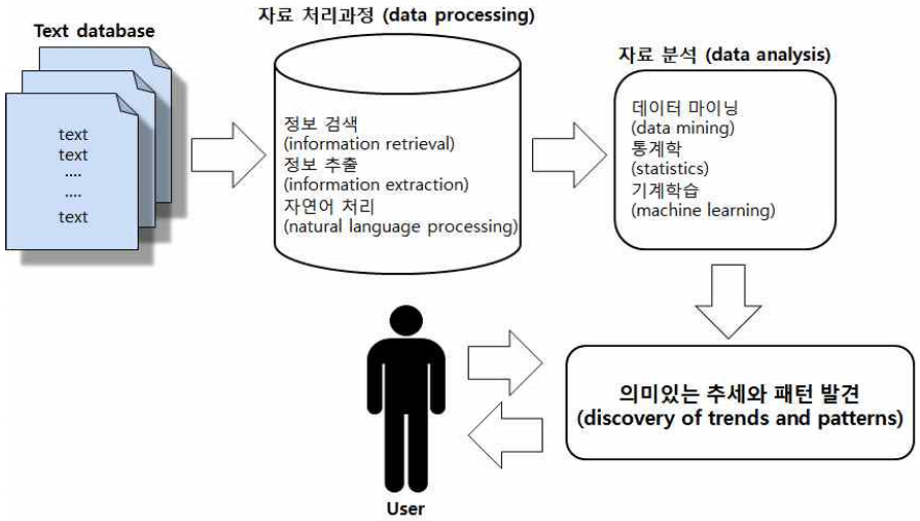
Ⅱ. 이론적 배경

2.1 텍스트마이닝과 오피니언마이닝

텍스트마이닝(Text Mining)은 다양한 텍스

트 정보원천으로부터 자동적으로 정보를 추출함으로써 이전에 알려지지 않았던 새로운 정보를 발견하는 정보처리기술이다(김근형, 2009). 텍스트마이닝은 데이터마이닝과 비슷한 개념이지만, 좁은 의미의 데이터마이닝은 테이블 구조의 정형화된 데이터구조에 최적화된 기술인 반면, 텍스트마이닝은 텍스트문서, e-메일, HTML화일과 같은 비구조화(unstructured) 또는 반구조화(semi-structured)된 데이터를 처리하는 것을 목적으로 한다. 온라인 고객리뷰(online customer review)는 일종의 텍스트로써 텍스트마이닝 기술에 의하여 분석 처리될 수 있다. 텍스트마이닝을 위하여 비구조화된 형태의 문서들이 수집되면 전처리(preprocessing)과정을 거쳐서 텍스트 분석이 수월한 형태로 변환되어야 한다. 전처리 과정에서는 불용어들이 제거되고 자연어처리기술에 기반한 형태소 분석기술 등을 활용하여 어근, 접두사/접미사, 품사(POS, part-of- speech) 등 다양한 언어적 속성의 구조를 파악하는 작업이 이루어진다. <그림 1>은 텍스트마이닝의 단계별 과정을 나타내고 있다(심영석, 2016).

오피니언마이닝(Opinion Mining)은 텍스트마이닝의 한 분야로서 텍스트를 분석하여 글쓴이의 의견이나 평가, 태도, 감성 등을 추출해내는 기법이다(장재영,2009). 텍스트 문서에 나타난 의견의 극성을 분석하는 감성분석이 가장 핵심적이다(김문지, 송은정, 김윤희, 2015). 어떤 의견의 주제가 무엇인지 보다는, 그 의견을 작성한 사람이 주제에 대하여 어떠한 감성을 가지고 있는지를 분석하는 기법이다. 예를 들어 그 주제에 대하여 긍정 또는 부정 감성 정도를 판단하여 분석한다(장재영, 2009). 즉, 오피니언마이닝은 웹사이트에 게시되어 있는 온라인 고객리뷰들을 분석 대상으로 하여 고객 의견에 대한 긍정(positive) 의견과 부정(negative)의견의 분포 등을 분석할 수 있다. 오피니언마이닝에서는 사용자의 감성과 의견을 수치화하여 객관적인 자료로 만들기 때문에 자연어처리기술이 필수적으로 활용된다.



<그림 1> 텍스트 마이닝의 단계별 과정

오피니언마이닝은 특징추출, 의견분류, 요약 및 표현 등의 3단계로 처리된다(양정연, 2009). 특징추출단계에서는 유용한 정보라고 판단되는 여러 특징들을 정의하고 추출해내는 단계이다. 이때 단순히 특징만을 추출하는 것이 아니라 해당 특징이 어떤 의미를 가지는가에 대한 의견을 나타내는 어휘정보도 함께 추출된다. 의견분류단계에서는 추출된 특징과 의견을 나타내는 어휘가 해당 정보소스에서 어떤 의미로 사용되었는가에 대한 판단 및 분류를 하는 단계이다. 요약 및 표현단계에서는 의견성향이 밝혀진 의견정보들을 요약하여 전체 정보의 내용을 효과적으로 사용자에게 전달하는 단계이다.

오피니언마이닝이 가장 성공적으로 적용되는 분야는 온라인 쇼핑몰에서 사용자의 상품평(product reviews)에 대한 분석이다. 실제 사용자가 작성한 상품평은 하나의 상품에 대해 사용자의 좋고 나쁨에 대한 감성을 표현한 결과이다. 따라서 개개인에 따라 긍정 또는 부정적인 의견으로 나눠지며 오피니언마이닝에서는 이러한 감성분석(sentiment analysis)을 통해 상품평을 자동으로 분류 요약하여 유용한 상품정보로 활용될 수 있다. 오피니언 마이닝 기술은 사용자들의 상품평을 효과적으로 추출하여 요약함으로써 잠재적 소비자들로 하여금 모든 상품평을 살펴보지 않더라도 상품에 대한 다양한 의견들을 쉽게 열람 및 검색할 수 있는 기반기술을 제공한다(장재영, 2009).

2.2 온라인리뷰와 랭킹모델

온라인리뷰는 비구조화된 텍스트형태로 표현되지만 평가점수와 날짜, 리뷰길이 등과 같은 구조적 데이터도 포함한다. <그림2>는 성산일출봉에 대한 온라인리뷰의 예를 나타내고 있다. 고객의 의견과 감성을 표현한 텍스트문장이외에도 평가점수(4개의 겹원)와 날짜 등이 나타나 있다. 온라인리뷰에서 평가점수는 해당 상품 및 서비스에 대한 호감 정도를 숫자로 표현한 것으로, <그림2>에서는 5점 척도 기준으로 4점을 부여한 예를 나타내고 있다.

<그림 1> 성산일출봉에 대한 온라인리뷰 예

텍스트는 명사와 형용사, 동사 등의 구성요소로 이루어진다. 명사는 일반적으로 상품 및 서비스의 속성이나 특징을 나타낸다. 의견이나 감성은 형용사나 동사를 통하여 표현된다(장재영, 2009). 텍스트를 통하여 표현되는 의견은 긍정의견, 부정의견, 조정의견으로 구분할 수 있다. 긍정의견은 긍정적인 형용사와 동사가 많이 포함된 텍스트이다. <그림1>의 고객리뷰에서 유채꽃이나 억새철은 '아름답다'라는 긍정 형용사와 함께 표현되어 있으므로 긍정의견에 속한다. 마찬가지로 부정의견은 부정적인 형용사와 동사가 많이 포함된 텍스트이다. 조정의견은 긍정 품사와 부정 품사를 동일선 상에서 검토하여 서로 상쇄시켜 도출된 의견을 의미한다. <그림1>에서 '정상' 이라는 속성은 '어렵다'라는 부정의미의 형용사와 '없다'라는 부정의미의 동사가 서로 상쇄되어 부정의미가 사라진다.

윤홍준(2010)은 긍정의견, 부정의견, 조정의견의 극성을 정량적으로 계산하기 위하여 감성극성 엔트로피(entropy) 계산식을 사용하였다. 이것은 상품평의 평가가 긍정과 부정 중 어느 한 방향으로 기울었는지 가늠하는 계산 방법으로서 엔트로피의 값에 따라 상품평의 극성이 선명한 정도를 파악할 수 있다.

$$H_d = (-\frac{p}{s}log\frac{p}{s}) + (-\frac{n}{s}log\frac{n}{s})$$

위 수식에서 p는 긍정적인 감성단어의 출현 횟수, n은 부정적인 감성단어의 출현 횟수, 그리고 s는 감성 단어 전체의 출현 회수를 나타낸다. 엔트로피의 값은 0에서 1 사이의 값으로 계산되며 0에 가까울수록 상품평의 극성이 뚜렷함을 의미하므로 사용자의 의견이 일관성 있게 표현된 상품평인 것으로 판단할 수 있다.

랭킹은 어떤 기준을 근거로 매겨진 순위를 의미한다. 랭킹모델은 그 순위를 도출하는 방법을 의미한다. 인터넷 상에서 검색엔진을 사용하여 자료를 검색할 때 검색 키워드와의 관련정도를 기준으로 검색결과에 대한 순위 즉, 랭킹(ranking)을 도출할 수 있다. 황원석(2013)의 연구에서는 검색키워드와의 관련성이 높은 순서로 논문을 효과적으로 검색할 수 있는 기법을 제안하였다. 오선주(2013)의 연구에서는 소셜네트워크에서 구성원 간 친밀도를 기준으로 관계의 순위를 매기는 랭킹 방법을 제안하였다.

온라인리뷰의 분석결과를 통하여 특정 상품이나 속성 등에 대한 호감도 또는 비호감도 순위 즉, 랭킹을 매길 수 있다. 호감도 랭킹은 평가점수 순으로 부여할 수도 있고 텍스트에 대한 정성적 분석을 통하여 긍정의견 정도에 따라 매길 수도 있다. 비호감도 랭킹은 호감도 랭킹을 역으로 산정하여 부여할 수 있다. 본 논문에서는 온라인리뷰 분석을 통한 호감도 랭킹 또는 비호감도 랭킹을 보다 정확히 부여하기 위한 새로운 방법을 제시하고자 한다.

2.3 선행 연구

김근형과 오성열(2009)의 연구에서는 텍스트마이닝 기술을 사회과학적 관점의 연구에 적용할 수 있도록 하기 위한 방법론을 제시하였다. 온라인 고객리뷰 데이터를 분석하여 고객특성을 도출하기 위하여 텍스트마이닝의 문서분류기술, 범주화기술, 정보추출 기술 등을 적용하기 위한 방법론을 제안하였으며 방법론 기반의 분석사례를 제시하였다. 장재영(2009)의 연구에서는 오피니언마이닝 기술을 이용하여 온라인 고객리뷰 형태의 상품평 의견을 자동으로 분류할 수 있는 감성분석 알고리즘을 제안하였다. 이 알고리즘은 온라인 쇼핑몰에 등록된 개별 상품평을 대상으로 긍정 및 부정의견을 판단하여 요약된 결과를 제공하는 기능을 하였다. 윤홍준 등(2010)의 연구에서는 대량의 상품평을 검색할 때 검색자의 목적에 최적화된 검색 우선순위를 부여하는 기법을 제안하였다. 상품평은 감성적이며 주관적인 의견을 포함하고 있기 때문에 상품평 검색은 일반 웹 검색과는 달라야 할 것이라는 아이디어를 기반으로 하고 있다. 웹 검색이 사용하는 사용자 검색어뿐만 아니라 상품평 내의 주관적인 의견의 포함 여부 및 감성 극성의 엔트로피 등을 고려하여 상품평의 검색 우선순위를 판단하는 기법을 제안하였다. 홍태호와 이태원(2014)는 온라인상점(amazon.com)에서의 영화평을 대상으로 문서

수준의 용어정보를 추출한 후 정보력이 높은 용어들을 대상으로 문서빈도, 정보획득량, 카이제곱 통계량을 도출하는 오피니언마이닝에 대한 연구를 하였다.

문성균 등(2014)의 연구에서는 온라인에서 브랜드나 기업이 어떻게 소비자들과 커뮤니케이션하고 있는지 살펴보기 위해 브랜드 트위터의 트윗 메시지를 유형화하고 그러한 메시지 유형 중에서 소비자들은 어떤 메시지 유형에 더욱 호의적인 반응을 나타냈는지 분석하였다. 조혜진 등(2015)의 연구에서는 온라인 주식 뉴스 텍스트에 한국어 역접관계를 고려한 오피니언 방의법 규칙 알고리즘을 제안하였다. 또한, 이 알고리즘을 적용하여 새로운 감성사전을 도출하고 코스피지수를 활용한 평판분석을 통하여 감성사전에 대한 정확도를 검증하였다. 양금과 이영찬(2015)의 연구에서는 중국관광객들을 대상으로 하여 자아일치성과 기능일치성이 구전효과와의 선형관계에서 자아구성이 어떤 조절효과를 나타내는지 분석하고 있다. 이삼열 등(2016)의 연구에서는 문화관련 온라인리뷰를 분석하여 다양한 문화예술 소비심리들이 어떻게 분포되어 있는지 고찰하고 있다. 사공원 등(2016)의 연구에서는 온라인리뷰에 담겨진 고객의 호텔서비스에 대한 감성을 도출하고 있으며 호텔간 서비스품질 비교와 시간에 따른 변화양상 등을 분석하고 있다.

이상에서 살펴본 바와 같이, 오피니언마이닝이나 온라인리뷰와 관련한 기존 연구들은 온라인리뷰의 텍스트를 보다 정확히 분석하기 위한 오피니언마이닝 알고리즘을 개선하는 연구이거나 온라인리뷰 자체를 분석하는 연구, 또는 온라인리뷰를 바라보는 사용자의 행태를 분석하는 연구들이다. 오피니언마이닝 알고리즘을 개선하는 연구는 자연어처리기술을 기반으로 하는데 비하여 본 논문은 온라인리뷰의 텍스트를 보다 정확히 분석하기 위하여 정량적인 데이터를 활용하자는 아이디어로서 기존 연구들과 차별성이 존재한다.

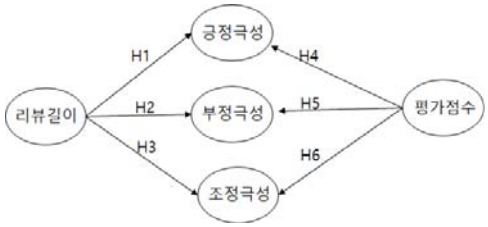
Ⅲ. 연구설계

3.1 연구 모형

본 논문에서는 온라인리뷰를 크게 정성적 데이터와 정량적 데이터의 요소를 갖는 것으로 간주하였다. 정성적 데이터는 작성자의 의견이 표현된 텍스트며 정량적 데이터는 평가점수와 리뷰길이이다. 텍스트는 오피니언마이닝 기술에 의하여 긍정의견 정도, 부정의견 정도, 조정의견 정도 등으로 가공할 수 있다. 리뷰길이는 텍스트를 구성하는 문자의 개수로부터 계산될 수 있다. 평가점수는 5점 척도 기반의 숫자로 표현된다.

본 논문에서는 온라인리뷰의 이러한 구성요소들 즉, 리뷰길이, 긍정극성 부정극성, 조정극성의, 평가점수 등을 연구변수로 설정하고 이들 사이의 선형회귀관계를 검증하기 위하여 H1～H6까지의 가설을 설정한다. 긍정극성 변수는 텍스트에서 긍정의견이 어느 정도 많은지를 나타내고, 부정극성 변수는 텍스트에서 부정의견이 얼마나 많은지를 나타낸다. 조정극성은 텍스트에서 긍정의견과 부정의견을 상쇄시킨 결과, 긍정의견이 얼마나 많은지 또는 부정의견이 얼마나 많은지를 의미한다. 리뷰길이는 텍스트의

길이를 의미하며 평가점수는 온라인리뷰를 5점 척도 기반으로 평가한 결과를 나타낸다. 이러한 변수들 사이의 선형회귀관계를 바탕으로 온라인리뷰의 긍정의견, 부정의견, 조정의견 정도 등을 보다 효과적으로 분석하고 호감도 랭킹 등을 정확히 매길 수 있는 랭킹모델링 방안을 제안한다.



<그림3> 연구모형

3.2 연구가설

상품 및 서비스에 대한 온라인리뷰에서 높은 호감도를 나타내는 긍정적인 극성을 표현하기 위해서는 긍정적인 감성단어를 가능한 많이 활용하면서 상품평을 작성하는 것이 일반적이다. 윤홍준 등(2010)의 연구에서 활용한 엔트로피 계산식도 긍정적인 감성단어의 출현횟수가 많을수록 긍정극성이 높게 판명되도록 설계되었다. 긍정적인 감성단어가 많아지기 위해서는 온라인리뷰의 길이가 길어져야 한다. 마찬가지로, 온라인리뷰의 텍스트 내에 부정적인 극성을 표현하기 위해서는 부정적인 감성단어를 가능한 많이 활용하면서 상품평을 작성할 것이다. 부정적인 감성단어가 많아지기 위해서는 온라인리뷰의 길이가 역시 길어질 것이다. 따라서 다음과 같이, 연구가설 H1과 H2를 수립하였다.

H1: 리뷰길이는 긍정극성에 유의한 영향을 미칠 것이다.

H2: 리뷰길이는 부정극성에 유의한 영향을 미칠 것이다.

조정극성은 긍정의견과 부정의견을 서로 상쇄시켜 도출된 극성을 의미한다. 긍정극성을 갖는 조정극성은 긍정적인 감성단어의 출현횟수는 많고 부정적인 감성단어의 출현횟수는 적을 것이지만 긍정단어가 많이 표현되기 위해서는 리뷰길이가 길어야 할 것이다. 마찬가지로, 부정극성을 갖는 조정극성은 부정적인 감성단어의 출현횟수는 많고 부정적인 감성단어의 출현횟수는 적을 것이지만 부정단어가 많이 표현되기 위해서는 리뷰길이가 길어야 할 것이다. 따라서 다음과 같이, 연구가설 H3을 수립하였다.

H3: 리뷰길이는 조정극성에 유의한 영향을 미칠 것이다.

평가점수를 높게 매기는 사용자는 해당 상품 및 서비스에 대한 호감도가 높기 때문이다. 해당 상품 및 서비스에 대한 호감도가 높으면 온라인리뷰의 텍스트를 통하여 긍정의견을 많이 표출할 것이며 긍정극성 또한 높아질 것이다. 일반적으로 텍스트 내에 긍정 단어 등을 이용하여 긍정의견을 많이 포함할 경우 해당 상품 및 서비스에 대한 호감도가 높을 것이며 높은 호감을 갖는 작성자는 높은 평가점수를 매길 것으로 예측할 수 있다. 마찬가지로, 온라인리뷰의 텍스트를 통하여 부정의견을 많이 표출하면 부정극성이 높아질 것이며 부정의견이 많으면 평가점수는 낮게 매겨질 것으로 예측할 수 있다. 따라서 다음과 같이, 연구가설 H5와 H6

를 수립하였다.

H4: 평가점수는 긍정극성에 유의한 영향을 미칠 것이다.
H5: 평가점수는 부정극성에 유의한 영향을 미칠 것이다.

조정극성은 긍정의견과 부정의견을 서로 상쇄하여 도출된 극성이지만 결과적으로는 긍정극성이나 부정극성으로 치우칠 것이다. 평가점수가 높을수록 긍정극성이 높을 것이라는 예측과 평가점수가 낮을수록 부정극성이 높을 것이라는 예측을 통하여 평가점수 또한 조정극성에 영향을 미칠 수 있다. 따라서 다음과 같이, 연구가설 H6을 수립하였다.

H6: 평가점수는 조정극성에 유의한 영향을 미칠 것이다.

3.3 연구변수의 측정

연구모형과 가설에서 언급된 연구변수들은 온라인리뷰의 구성요소들과 가공한 결과 등을 활용하여 측정하였다. 각 변수들에 대한 설명과 측정방법은 <표 1>에 명시되어 있다.

3.4 분석절차

연구변수들 사이의 회귀적 영향 관계를 파악하고 랭킹모델을 제시하기 위한 분석절차를 4단계로 설계하였다. 첫 번째 단계에서는 온라인리뷰가 게시된 웹사이트로부터 제주 명소 관련 온라인리뷰 데이터를 수집하고 각 리뷰의 텍스트에서 형용사를 추출한다. 두 번째 단계에서는 추출된 형용사를 포함해 리뷰 데이터를 정형화하고 이를 표준화하여 가공 처리한다. 세 번째 단계에서는 수립된 가설들을 검증하기 위해 오피니언 마이닝 기법으로 가공된 데이터를 가지고 상관분석 및 회귀분석을 실시한다. 네 번째 단계에서는 가설검증 결과를 바탕으로 랭킹모델을 제안한다.

Ⅳ. 실증분석

4.1 온라인리뷰 데이터의 수집

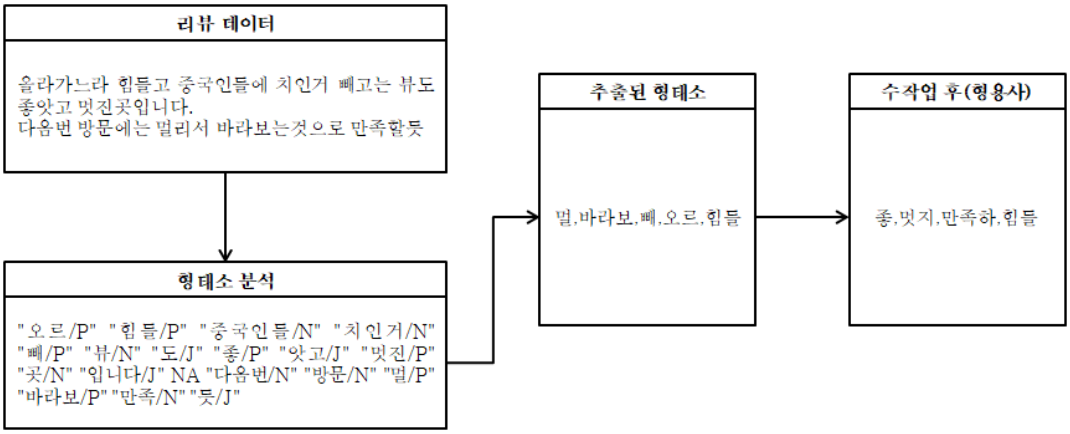
본 논문에서는 관광웹사이트인 트립어드바이저(TripAdvisor) 사이트에 게시된 관광리뷰 중 제주 명소 관련 리뷰 데이터를 분석 대상으로 선정하였다. 이 데이터는 195개의 제주 관광명소에 대한 2697개의 온라인리뷰를 포함한다. 각 온라인리뷰에는 관광지명, 리뷰제목, 리뷰내

<표 1> 변수 측정 방법

연구변수	변수 설명 및 측정 방법
긍정극성	리뷰의 긍정 형용사 개수를 5점 척도 기반으로 변환
부정극성	리뷰의 부정 형용사 개수를 5점 척도 기반으로 변환
조정극성	긍정의견에서 부정의견을 상쇄한 형용사 개수를 5점 척도 기반으로 변환
리뷰길이	리뷰의 텍스트에 사용된 문자의 개수를 5점 척도 기반으로 변환
정량점수	리뷰 작성자가 리뷰의 대상이 되는 상품 및 서비스에 대하여 부여한 5점척도 기반의 평점

<표 2> 온라인리뷰 데이터의 구조화

관광지명	리뷰제목	리뷰내용	일련번호	리뷰작성일	평점	계절
성산 일출봉	두번 가봤는데 너무	다 좋았네요.	1	2016/08	5	여름
성산 일출봉	역시 해뜰때 봐야죠	해뜰때 보기나 혹은 해가지고나서 주변 야경을	1	2016/05	5	봄
성산 일출봉	힘들지만 굿	뷰도 좋았고 멋진곳입니다.	1	2016/03	5	봄
성산 일출봉	쫑권런너무많음	정말 이것만 보면 예쁜데, 중국인 관광객들이 너	1	2016/02	5	겨울
성산 일출봉	새해맞이 성산일출봉	하였습니다. 갈때마다 느끼지만 웅장한 자연경	1	2016/12	5	겨울
성산 일출봉	유네스코 자연유산의	올라가는데 20분이면 충분하고 정말 힘들다 싶	1	2016/10	5	가을
성산 일출봉	한해의 끝과 시작		1	2016/12	4	겨울
성산 일출봉	제주도에서 반드시 들	성산 일출봉은 일출 보기로도 유명한 곳이다. 유	1	2016/12	4	겨울
성산 일출봉	가족방문기	유채꽃사진 찍는곳이 돈을받아 눈살이 찌푸려	1	2016/12	5	겨울
성산 일출봉	제주의 절경	제주에서도 이름난 곳이라 찾는 사람도 많네요.	1	2016/03	5	봄



<그림 4> 형용사 추출 과정

용, 리뷰작성일, 방문날짜, 평점 등의 항목을 포함한다. <표2>는 웹크롤링(Web Crawaling)을 통하여 수집한 온라인리뷰를 엑셀화일로 구조화시킨 내용을 나타내고 있다. 구조화된 온라인리뷰 데이터는 통계 소프트웨어인 R을 이용하여 분석하였다.

4.2 형용사 추출

온라인리뷰의 텍스트에서 표현된 감성적인 의견에 따라 긍정 및 부정극성을 파악할 수 있는데, 그 역할을 가장 잘 나타내는 형용사를 추출하고자 하였다. 온라인리뷰의 텍스트에서 형태소 분석을 위해 KoNLP 패키지의 SimplePos22() 함수를 사용하였다. SimplePos22() 함수는 KAIST에서 개발한 한글 형태소 분석 함수이다. 형용사에 해당하는 형태소를 추출하기 위해 stringr 패키지의 str_match() 함수를 사용하였다. 추출 결과에서 결측치(NA), 불용어를 제거하기 위해 !is.na() 함수, gsub() 함수를 사용하였다. 그런데 SimplePos22(), str_match() 등 함수의 성능 문제로 인하여 추출되지 못한 형용사가 존재하거나 다른 품사의 형태소도 함께 추출되는 문제가 있었다. 추출되지 못한 형용사는 수작업을 통하여 보완하였다. <그림4>는 이러한 과정을 나타내고 있다.

4.3 온라인리뷰의 정형화 및 표준화

비정형데이터인 온라인리뷰를 정형 데이터로 변환하기 위해 추출한 형용사들의 극성에 따라 긍정/부정으로 분류하여 각 리뷰별로 긍정의견(긍정 형용사 개수), 부정의견(부정 형용사 개수)을 계산하고, 또 긍정형용사와 부정형용사를 서로 상쇄하여 조정의견을 계산하였다. R에서 nchar() 함수를 사용하여 각 리뷰에 포함된 텍스트의 길이도 계산하였다. 정형화된 데이터들은 5점 척도 기반으로 변환하여 표준화하였다. 표준화는 각 데이터를 1~5 사이의 값으로 변환하기 위해 각 컬럼별로 범위기준을 다르게 설정하였다. 예를 들어 긍정의견의 경우 3을 기준으로 3 이하는 기본값으로 하고 4 이상은 3~5 사이의 값으로 변환하였다. 리뷰길이의 경우 250자를 기준으로 250자 이하는 1, 250자 이상은 2~5 사이의 값으로 변환하였다. <그림 5>는 표준화 과정을 나타내고 있으며, <표 3>은 표준화 결과를 나타내고 있다. 온라

리뷰내용	형용사
네요.	말로 형용할 수 없,어렵지 않,시원하,좋,힘들
해뜰때 보거나 혹은 해가진걸보고나서 주변 야경을 볼때	아깝지 않
좋았고 멋진곳입니다.	좋,멋지,만족하,힘들
정말 이것만 보려	예쁘
니다. 갈때마다 느끼지만 웅장한 자연경관에 항상 모	웅장하,나르,맛있,비싸
을라가는데 20분이면	멋지,힘들
	적당하,좋
성산 일출봉은 일출	유명하,멋지,매력,좋
유채꽃사진 찍는곳이 돈을받아 눈살이 찌푸려집니다	좋,눈살이 찌푸리,아름답
제주에서도 이름난 곳이라 찾는 사람도 많네요. 봄에	아름답,이름나

긍정형용사	부정형용사	조정(긍정-부정)	리뷰길이
4	1	3	206
1	0	1	55
3	1	2	64
1	0	1	80
3	1	2	156
1	1	0	76
2	0	2	82
4	0	4	161
2	1	1	86
2	0	2	67

긍정의견	부정의견	조정의견	리뷰길이
3	1	4	1
1	1	3	1
3	1	4	1
1	1	3	1
3	1	4	1
1	1	3	1
2	1	4	1
3	1	4	1
2	1	3	1
2	1	4	1

<그림 5> 온라인리뷰의 정형화 및 표준화 결과

<표 3> 가공된 데이터

관광지명	리뷰제목	리뷰작성일	계절	긍정의견	부정의견	조정의견	리뷰길이	정량점수(평점)
성산 일출봉	두번 가봤는데 너무 좋의	2016/08	여름	3	1	4	1	5
성산 일출봉	역시 해뜰때 봐야죠	2016/05	봄	1	1	3	1	5
성산 일출봉	힘들지만 굿	2016/03	봄	3	1	4	1	5
성산 일출봉	줄귀런너무많음	2016/02	겨울	1	1	3	1	5
성산 일출봉	새해맞이 성산일출봉	2016/12	겨울	3	1	4	1	5
성산 일출봉	유네스코 자연유산의 경	2016/10	가을	1	1	3	1	5
성산 일출봉	한해의 끝과 시작	2016/12	겨울	2	1	4	1	4
성산 일출봉	제주도에서 반드시 들러	2016/12	겨울	3	1	4	1	4
성산 일출봉	가족방문기	2016/12	겨울	2	1	3	1	5
성산 일출봉	제주의 절경	2016/03	봄	2	1	4	1	5

인리뷰의 원천데이터에는 관광지명, 리뷰제목, 리뷰내용, 리뷰작성일, 방문날짜, 평점 등의 데이터 항목이 있었지만, 이러한 데이터항목들 중에서 필요한 항목만을 선택하고 가공하여 <표 3>과 같이 긍정의견, 부정의견, 조정의견, 리뷰길이, 정량점수 등의 스키마(schema)로 구성된 데이터를 도출하였다.

4.4 연구 가설의 검증

연구 모형에 설정된 연구변수들 사이에서 서로 상관성이 있는지를 알아보기 위해 상관분석을 실시하였으며 그 결과는 다음 <표 4>와 같다.

평가점수는 리뷰길이를 제외한 모든 변수들과 유의미한 상관관계를 보이고 있다. 이는 평가점수가 긍정극성, 부정극성, 조정극성 등에 유의미한 영향을 미칠 가능성을 시사하고 있다. 리뷰길이는 긍정극성 및 부정극성과 유의미한 상관관계를 보이고 있는 것으로 보아, 긍정극성과 부정극성에는 유의미한 영향을 미칠 가능성이 있으나 조정극성에는 그렇지 않을 수 있음을 나타내고 있다.

연구 가설 H1과 H4를 검증하기 위하여 리뷰길이와 평가점수를 독립변수, 긍정극성을 종속변수로 두어 다중회귀분석을 실시하였다. <표 5>와 같이 리뷰길이와 평가점수는 긍정의견에 $p < 0.001$ 수준에서 정(+)의 영향을 미치고 있음을 알 수 있다. 가설 H1과 H4는 채택되었다.

연구 가설 H2와 H5를 검증하기 위하여 리뷰길이와 평가점수를 독립변수, 부정극성을 종속변수로 두어 다중회귀분석을 실시하였다. <표 6>과 같이 리뷰길이와 평가점수는 부정의견에 $p < 0.001$ 수준에서 정(+)의 영향을 미치고 있음을 알 수 있다. 가설 H2와 H5는 채택되었다.

<표 4> 연구변수들 간의 상관분석

변수	평균	표준편차	1	2	3	4	5
평가점수	4.26	.847	1				
긍정극성	1.83	.847	.243***	1			
부정극성	1.12	.407	-.276***	-.125***	1		
조정극성	3.38	.677	.370***	.765***	-.508***	1	
리뷰길이	1.06	.295	.015	.151***	.152***	.019	1

<표 5> 긍정극성과의 다중회귀분석결과(H1, H4)

독립변수	B	표준오차	Beta	t-value	Sig.t
리뷰길이	.423	.053	.147	7.982	.000***
평가점수	.241	.018	.241	14.040	.000***
R2 = .081, F=118.514					

<표 6> 부정극성과의 다중회귀분석결과(H2, H5)

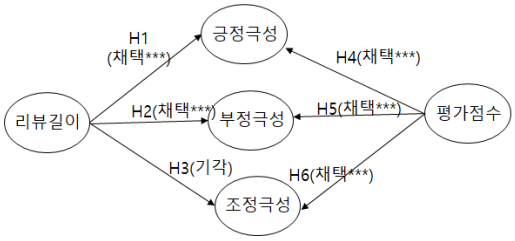
독립변수	B	표준오차	Beta	t-value	Sig.t
리뷰길이	-.134	.009	-.279	-15.244	.000***
평가점수	.241	.018	.241	14.040	.000***
R2 = .101, F=150.970					

<표 7> 조정극성과의 다중회귀분석결과(H3,H6

독립변수	B	표준오차	Beta	t-value	Sig.t
리뷰길이	.032	.041	.014	.768	.442
평가점수	.296	.014	.370	20.681	.000***
R2 = .037, F=214.447					

연구 가설 H3과 H6을 검증하기 위하여 리뷰길이와 평가점수를 독립변수, 조정극성을 종속변수로 두어 다중회귀분석을 실시하였다. <표 7>과 같이 평가점수는 조정극성에 유의한 영향을 미쳤으나, 리뷰길이는 조정극성에 유의한 영향을 미치지 않았다. 따라서, 가설 H6은 채택되었으나 H3은 기각되었다.

<그림 6>은 가설검증 결과를 나타내고 있다.



***:p<0.001

<그림 6> 가설검증 결과

4.5 랭킹모델링을 위한 제언

온라인리뷰에서 언급되는 상품, 서비스, 세부 속성 등을 대상으로 한 긍정극성 또는 부정극성 등에 대한 랭킹(Ranking)은 소비자들에게 단순하면서도 강렬한 정보로서의 역할을 할 수 있는데, 온라인리뷰의 텍스트에 포함된 감성적 또는 긍정/부정적 의견과 정량적인 평가점수 등을 종합 적용하여 도출되어야 정확하게 랭킹을 부여할 수 있다. 따라서 특정 상품 및 서비스에 대한 온라인리뷰의 긍정 정도 및 부정 정도 등에 대한 랭킹을 분석할 때 <그림7>과 같은 분석모델을 제안한다.

$$\text{최종긍정극성} = \alpha \cdot \text{긍정극성} + \beta \cdot \text{리뷰길이} + \gamma \cdot \text{평가점수}$$

$$\text{최종부정극성} = \alpha \cdot \text{부정극성} + \beta \cdot \text{리뷰길이} + \gamma \cdot \text{평가점수}$$

$$\text{최종조정극성} = \alpha \cdot \text{조정극성} + \gamma \cdot \text{평가점수}$$

<그림 7> 랭킹 분석모델

최종긍정극성은 앞서 채택된 가설 H1과 H4를 바탕으로 도출하였다. 긍정극성은 리뷰길이 및 평가점수와 유의한 영향 관계를 갖고 있다는 점을 근거로 하여 최종긍정극성을 계산할 때 리뷰길이와 평가점수를 고려하여 계산한다. 최종부정극성은 앞서 채택된 가설 H2와 H5를 근거로 도출되었다. 부정극성은 각각 리뷰길이 및 평가점수와 유의한 영향 관계를 갖고 있다는 점을 토대로 최종부정극성 점수 계산할 때 부정극성, 리뷰길이, 평가점수 등을 조합하여 계산한다. 최종조정극성은 앞서 채택된 가설 H6을 바탕으로 도출하였는데, 조정극성은 평가점수와 유의한 영향 관계를 갖고 있다는 점을 토대로 점수 계산 시, 조정극성과 평가점수를 조합하여 계산한다. 여기서, 각 수식마다 α, β, γ 등의 가중치 값을 0에서 1사이의 값으로 적당하게 설정하여야 한다.

기존의 오피니언마이닝 알고리즘은 온라인

리뷰를 분석할 때 최종적으로 조정극성만을 도출하는데 비하여, 제안하는 랭킹 분석모델에서는 조정극성과 평가점수를 활용하여 보다 정확한 최종조정극성을 도출할 수 있다. 뿐만 아니라 최종조정극성을 해석할 때 보조 자료로 활용할 수 있는 최종긍정극성과 최종부정극성도 도출할 수 있다.

Ⅴ. 결 론

본 논문에서는 온라인리뷰 데이터의 정성적 요소와 정량적 요소와의 인과적 관계를 고찰하였으며 이를 기반으로 온라인리뷰의 보다 정확한 분석을 위한 랭킹 분석모델을 제시하였다. 랭킹 분석모델은 온라인리뷰에서 언급되는 상품이나 서비스 등에 대한 품평이 얼마나 긍정적인지 또는 부정적인지를 기준으로 순위를 도출하기 위한 모델이다. 랭킹 분석모델에 포함된 변수들은 온라인리뷰를 구성하는 정성적 요소라 할 수 있는 긍정극성, 부정극성, 조정극성과 정량적 요소라 할 수 있는 리뷰길이, 평가점수 등이다.

본 논문에서는 관광웹사이트인 트립어드바이저에 게시된 다량의 온라인리뷰를 대상으로 정량적 데이터인 평가점수와 리뷰길이를 추출하였고, 정성적 데이터인 텍스트리뷰에서 오피니언마이닝 기술 등으로 형용사 등을 추출 및 가공하였으며, 리뷰길이 및 평가점수와 긍정극성, 부정극성, 조정극성간의 상관관계 및 회귀적 관계를 분석하였다.

분석 결과, 정량적 데이터요소인 리뷰길이는 정성적 데이터요소인 긍정극성과 부정극성과 유의한 영향관계가 존재하였으나 조정극성과는 유의한 영향관계가 없었다. 정량적 데이터요소인 평가점수는 긍정극성, 부정극성, 조정극성과 부정극성과는 유의한 영향관계가 존재하였다. 랭킹 분석모델에서 최종적인 긍정극성은 오피니언마이닝에 의하여 도출된 긍정극성뿐만 아니라 리뷰길이 및 평가점수도 고려하여 도출하도록 하였다. 랭킹 분석모델에서 최종적인 부정극성은 오피니언마이닝에 의하여 도출된 부정극성뿐만 아니라 리뷰길이 및 평가점수도 함께 고려하여 도출되도록 하였다. 랭킹 분석모델에서 최종적인 조정극성은 오피니언마이닝에 의하여 도출된 조정극성과 함께 및 평가점수도 고려하도록 하였다.

오피니언마이닝 알고리즘을 개선하는 기존의 연구는 자연어처리기술을 기반으로 온라인리뷰의 텍스트만을 정성적으로 분석하는데 비하여, 본 연구에서 제안한 랭킹 분석모델은 자연어처리 기술 기반의 텍스트 분석뿐만 온라인리뷰의 또 다른 구성요소인 리뷰길이와 평가점수 등과 같은 정량적인 데이터도 활용함으로써 랭킹분석의 정확도를 높일 수 있다. 본 논문에서 제안한 랭킹분석모델은 기존의 오피니언마이닝을 포괄하는 더 일반화된 모델이면서 분석정확도를 더욱 개선할 수 있다는 측면에서 의의가 있다.

온라인리뷰를 분석하는 과정에서 제주 명소 관련 리뷰 데이터에 한정하여 연구했다는 점은 본 논문의 한계이다. 추후, 다양한 유형의 온라인리뷰를 수집하여 추가적인 분석을 수행할 필요가 있다. 특히, 랭킹 분석모델에서 α, β, γ 값의 범위를 구체화시키지 못한 부분은 차후의 연구과제로 남겨둔다. 다양한 데이터를 기반으

로 시뮬레이션을 하면서 α, β, ɣ 값의 범위를 좁히는 연구가 더 필요하다. 또한, 온라인리뷰의 텍스트에서 형용사를 추출할 때 한글 자연어처리 기술의 한계로 인하여 수작업 과정이 포함되었음을 밝히며, 향후 향상된 자연어처리 소프트웨어를 사용하여 온라인리뷰를 보다 정확하게 분석할 수 있기를 기대한다.

참고문헌

김근형, 오성열, “온라인 고객리뷰 분석을 통한 시장세분화에 텍스트마이닝 기술을 적용하기 위한 방법론,” 한국콘텐츠학회논문지, 제9권, 제8호, 2009, pp.272-284.

김문지, 송은정, 김윤희, “온라인 리뷰 데이터의 오피니언마이닝을 통한 콘텐츠 만족도 분석 시스템 설계,” 한국인터넷정보학회 추계학술발표대회, 제17권, 제3호, 2016, pp.107-113.

문성균, 유희숙, 권건우, “트위터 메시지 유형이 메시지 수용자 반응에 미치는 영향에 관한 내용분석 연구,” 정보시스템연구, 제23권, 제4호, 2014, pp.1-24.

박은주, 정유진, “온라인 리뷰 탐색이 화장품 구매의도에 미치는 영향,” 한국생활과학회지, 제22권, 제2호, 2013, pp.343-355.

백현미, 안중호, 하상욱, “상품 가격에 따른 온라인 리뷰 유익성 결정 요인에 관한 연구,” 한국전자거래학회지, 제26권, 제3호, 2011. pp.93-112.

사공원, 하성호, 박경배, “온라인 후기에 내재된 고객의 감성분석과 LQI차원별 호텔서비스 품질평가,” 정보시스템연구, 제25권, 제3호, 2016, pp.217-245.

심영석, “텍스트 마이닝(Text Mining)을 이용한 관광지 이미지 결정요인에 관한 연구,” 세종대학교 석사학위논문, 2016.

양금, 이영찬, “중국관광객의 온라인구전에 대한 자아일치성과 기능일치성의 효과,” 정보시스템연구, 제24권, 제4호, 2015, pp.1-23.

양정연, 명재석, 이상구, “상품 리뷰요약에서의 문맥정보를 이용한 의견 분류 방법,” 정보과학회논문지:데이터베이스, 제36권, 제4호, 2009, pp.255-262.

오선주, “소셜네트워크에서 관계 랭킹 모델,” 한국전자거래학회지, 제18권, 제3호, 2013, pp.93-105.

윤홍준, 김한준, 장재영, "오피니언마이닝기술을 이용한 효율적 상품평 검색 기법," 정보과학회논문지, 제16권, 제2호, 2010, pp.222-226.

유상욱, “다차원 순위 모델 기반의 관광후기 분석”, 제주대학교 석사학위논문, 2017.

이승준, “온라인 고객 리뷰의 특성이 리뷰 유용성에 미치는 영향”, 홍익대학교 석사학위논문, 2012.

이삼열, 천영준, 곽규태, “한국인의 문화예술 소비심리 분석”, 문화정책총론, 제30권, 제1호, 2016, pp.204-226.

장재영, “온라인 쇼핑몰의 상품평 자동분류를 위한 감성분석 알고리즘,” 한국전자거래학회지, 제14권, 제4호, 2009, pp.19-33.

조혜진, 서진훈, 최진탁, “주식뉴스 콘텐츠를 활용한 오피니언마이닝 기반의 OAR 감성사전 알고리즘 기법,” 한국정보기술학회논문지, 제13권, 제3호, 2015, pp.111-119.

홍태호, 이태원, “온라인 상점 고객의 감성분류를 위한 용어기반 오피니언마이닝,” 2014년 한국경영정보학회 춘계학술대회, 2014, pp.619-624.

황원석, 채수민, 김상욱, 최호진, “논문 검색 엔진을 한 랭킹 방법,” 한국정보과학회논문지, 제40권, 제5호, 2013, pp.345-357.

Dwayne, D. G., Kevin, P. G., and Stephen, W. B. “Generating positive word of mouth communication through customer-employee relationships,” *International Journal of Service Industry Management*, Vol.12, No.1, 2001, pp.44-59.

Jeon, W. Y. and Jung, H. J. “Effects of online reviews on evaluation and purchase intention of a product in internet shopping: the role of gender differences,” *Journal of The Korean Psychological Association*, Vol.7, No.1, 2006, pp.113-129.

Peterson, R. A. and Meria, C. M. “Consumer information search behavior and internet,” *Psychology and Marketing*, Vol.20, No.2, 2003, pp.99-121.

이 창 용 (Lee, Changyong)

제주대 경영학사와 대학원에서 관광융합소프트웨어 전공으로 석사학위를 취득하였다. 현재 ㈜HB네트웍스에서 근무하고 있다.

김 근 형 (Kim, Keunhyung)

서강대 컴퓨터공학사와 대학원에서 석사 및 박사학위를 취득하였으며 현재 제주대학교 경영정보학과에 교수로 재직하고 있다.

<Abstract>

Causal model analysis between quantity and quality for deriving ranking model of Online reviews

Lee, Changyong · Kim, Keunhyung

Purpose

The purpose of this study is to analyze causal relationship between quantity and quality for deriving ranking model of Online reviews. Thus, we propose implications for deriving the ranking model for retrieving Online reviews more effectively.

Design/methodology/approach

We collected Online review from Tripadvisor web sites which might be a kind of world-famous tourism web sites. We transformed the natural text reviews to quantified data which consists of quantified positive opinions, quantified negative opinions, quantified modification opinions, reviews lengths and grade scores by using opinion mining technologies in R package. We executed corelation and regression analysis about the data.

Findings

According to the empirical analysis result, this study confirmed that the review length influenced positive opinion, negative opinion and modification opinion. We also confirmed that negative opinion and modification opinion influenced the grade score.

Keyword: Online reviews, Causal model, quantity, quality, ranking model

* 이 논문은 2018년 11월 5일 접수, 2018년 12월 11일 1차 심사, 2019년 2월 14일 2차 심사, 2019년 2월 19일 게재 확정되었습니다.