ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 September: 22(3): 23~43 http://dx.doi.org/10.13088/jiis.2016.22.3.023

사용자 리뷰의 평가기준 별 이슈 식별 방법론: 호텔 리뷰 사이트를 중심으로

변성호
국민대학교 비즈니스IT전문대학원
(*formyjihad@kookmin.ac.kr*)

이동훈
국민대학교 비즈니스IT전문대학원
(*donghoonlee@kookmin.ac.kr*)

김남규
국민대학교 비즈니스IT전문대학원
(*ngkim@kookmin.ac.kr*)

최근 IT기술의 발전에 따라 많은 사람들이 자신들의 여가활동에 대한 경험을 공유하고 있으며, 역으로 다른 사람들의 여가활동에 대한 경험을 참고하여 더 나은 여가활동을 누릴 수 있는 기회를 얻게 되었다. 이러한 현상은 영화, 숙박, 음식, 여행 등 여가활동 전반에 걸쳐 나타나고 있으며, 그 중심에는 여가활동에 대한 정보를 요약하여 제공하는 수많은 사이트가 있다. 대부분의 여가활동 정보 사이트는 각 상품에 대한 평균 평점뿐만 아니라 상세 리뷰를 제공함으로써, 해당 상품을 구매하고자 하는 잠재고객의 의사결정을 지원하고 있다. 하지만 기존 대부분의 사이트는 한 단계의 평가기준에 따라 평점과 리뷰를 제공하기 때문에, 각 평가기준을 구성하는 세부 요소에 대한 특징과 평가기준 별 주요 이슈를 파악하기 위해서는 상당히 많은 수의 리뷰를 직접 읽어야 한다는 불편이 따른다. 즉 사용자는 자신이 중요한 것으로 생각하는 평가기준에 대한 조건을 파악하기 위해, 많은 수의 리뷰를 하나하나 읽어보는 과정에서 많은 시간과 노력을 소비하게 된다. 예를 들어 호텔의 접근성, 객실, 서비스, 음식 등 한 단계의 평가기준만을 사용하여 평점과 리뷰를 제공하는 사이트의 경우, 접근성 중 특히 지하철역과의 거리, 객실 중 특히 욕실의 상태를 살펴보고자 하는 사용자에게 필요한 정보를 충분히 제공하지 못하게 된다. 따라서 본 연구에서는 기존 여가활동 정보 사이트의 한계, 즉 평가기준별로 입력된 리뷰를 신뢰하기 어렵다는 점과 평가기준을 구성하고 있는 세부 내용을 파악하기 어렵다는 점을 극복하기 위한 방안을 제시하고자 한다.

본 연구에서 제안하는 방법론은 사용자가 별도의 구분 없이 입력한 리뷰를 그 내용에 따라 평가기준별로 자동 분류하고, 각 평가 기준 별 주요 이슈를 요약하여 제공한다. 제안 방법론은 최근 텍스트 분석에 활발하게 사용되고 있는 토픽 모델링(Topic Modeling)에 기반을 두고 있으며, 각 리뷰를 하나의 문서 단위로 사용하는 것이 아니라 리뷰를 문장 단위로 끊어 개별 리뷰 유닛(Review Unit)으로 분해한 뒤, 평가기준별로 리뷰 유닛을 재구성하여 분석한다는 측면에서 기존의 토픽 모델링 기반 연구와 큰 차이가 있다고 할 수 있다. 본 논문에서는 제안 방법론을 실제 호텔 정보 사이트에서 수집한 423건의 리뷰 문서에 적용하여 6가지 평가기준에 대해 총 4,860건의 리뷰 유닛을 재구성하고, 이에 대한 분석 결과를 소개함으로써 제안 방법론의 유용성을 간접적으로 보인다.

주제어 : 빅 데이터, 리뷰 분석, 텍스트 마이닝, 토픽 모델링

논문접수일 : 2016년 8월 17일 논문수정일 : 2016년 9월 19일 게재확정일 : 2016년 9월 24일
원고유형 : 일반논문(급행) 교신저자 : 김남규

1. 서론

최근 IT 기술의 발전을 통해 다양한 모바일 기기가 보급됨에 따라, 여러 여가활동에 대한 정보를 쉽게 공유하고 획득할 수 있는 환경이 조성되었다. 구체적으로 소셜 네트워크 서비스(Social Network Service)나 블로그, 개인 홈페이지 등에서 누구나 쉽게 여가활동에 대한 경험을 글로 작

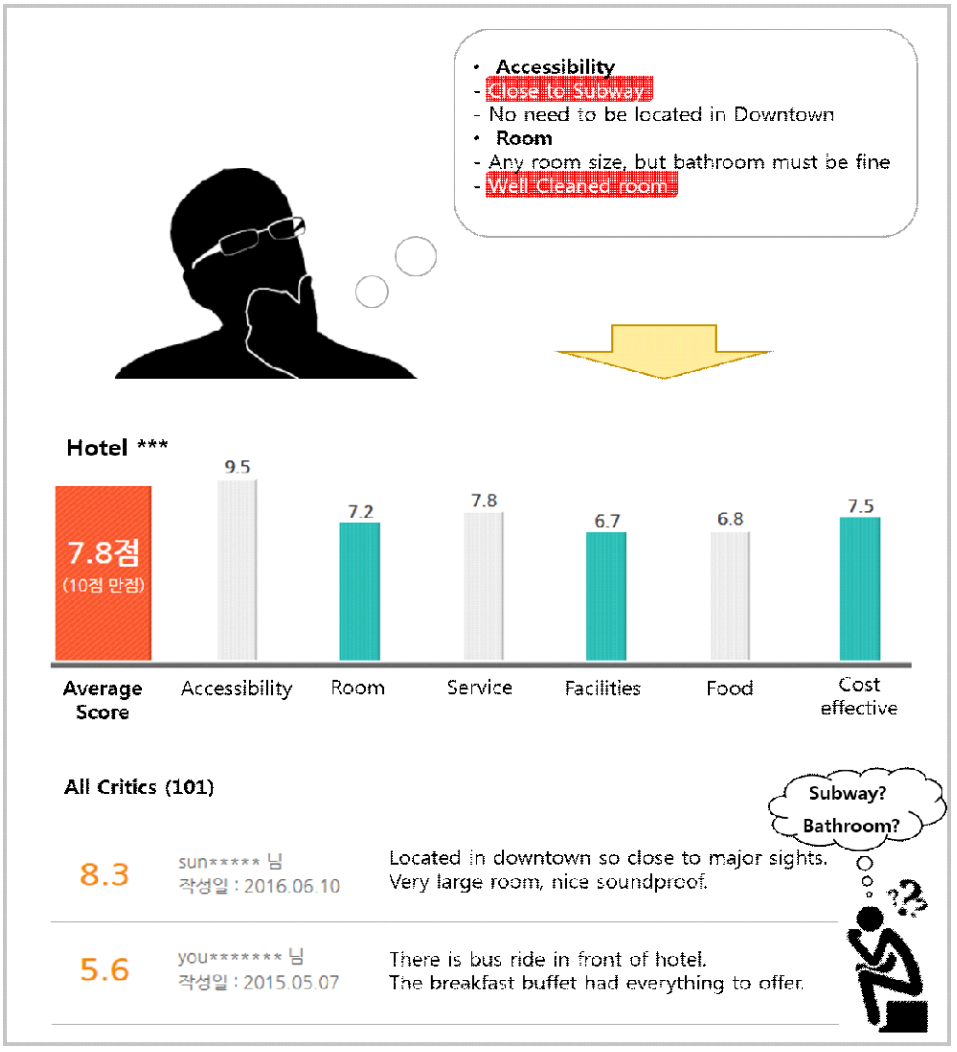
성하여 공유하고 있으며, 전문적으로 특정 여가활동의 정보만을 종합하여 제공하는 사이트도 다수 출현하였다. 심지어 잡지나 신문 등의 매체에 자신의 평점과 평론을 게시하던 영화 평론가, 소설 평론가 등의 전문가들도 이러한 여가활동 정보 사이트에서 전문적으로 상품에 대해 평점을 부여하고 리뷰를 작성하기 시작했으며, 일반 사용자들은 이러한 전문가의 리뷰뿐 아니라 같은 여가활동을 즐기는 다른 일반 사용자의 리뷰 또한 함께 참고하여 더 나은 여가활동을 누리기 위한 의사결정에 도움을 받고 있다.

이러한 현상은 특정 분야에 국한되지 않고, 여가활동 전반에 걸쳐 활발하게 이루어지고 있다. 예를 들어 Tripadvisor, Webtour, 그리고 VisitACity 등은 여행에 대한 종합 정보를 제공하고 있으며, 여행 관련 정보 중에서도 숙박에 대한 정보만을 특화 시켜 제공하는 Booking.com, Hotels.com, 그리고 Hotelpass.com 등도 사용자층이 점차 확대되고 있다. 또한 여행 이외에도 영화 정보가 활발하게 공유되는 Rotten Tomato, Cine21 등과 해외 드라마, 음악 정보 전문 사이트인 Metacritics 등도 여가활동 정보를 다루는 대표적인 사이트로 자리매김하고 있다.

이처럼 다양한 분야의 여가활동 정보 사이트들은 각각의 목적 및 관점에 따라 다양한 형태의 정보를 제공하고 있으나, 일반적으로 각 상품에 대한 전체 평점과 함께 해당 상품을 구매한 기존 고객들의 상세 리뷰를 제공하고 있다. 이들 사이트 대부분은 별도의 평가기준 없이 각 상품에 대한 통합 리뷰를 제공하고 있기 때문에(Rotten Tomato, Metacritics), 사용자가 특정 기준에 대한 리뷰를 선택적으로 조회하는 것이 거의 불가능하다. 물론 평가기준을 설정하고 이에 따라 평점 및 리뷰를 작성하게끔 유도하는 사이트들도 일부 존재하지만(Tripadvisor, Hotelpass.com), 리뷰를 작성하는 사용자들이 이러한 기준을 무시하고 동일 또는 유사한 내용을 여러 평가기준에 대해 입력하는 경우가 많다. 이로 인해, 예를 들면, 특정 호텔에 대한 접근성 평가기준의 리뷰를 조회했는데 접근성 뿐 아니라 서비스 및 음식에 대한 리뷰도 혼재되어 나타나게 되며, 이는 사용자 만족도 및 사이트 신뢰도 저하의 원인이 된다.

기존 여가활동 정보 사이트에서 제공되는 정보의 또 다른 한계는 평가기준의 구체성 측면에서 찾을 수 있다. 예를 들어 <Figure 1>의 사용자는 접근성 평가기준에서는 지하철역이 가까워야 하며 도심일 필요는 없다는 요구사항을 갖고 있으며, 객실 평가기준에서는 방 크기는 상관없으며 욕실은 넓어야 하고 청결 상태가 가장 중요하다는 요구사항을 갖고 있다. 그림의 예에서 접근성 평가기준의 평균 평점이 매우 높은 것으로 나타났지만, 그 근거, 즉 호텔이 지하철역과 가까운지, 호텔을 지나는 버스가 많은지, 도심에 위치하는지에 대한 정보를 요약하여 제시하지는 않는다. 따라서 접근성 중에서도 특히 지하철이 가까워야 하며 객실 중에서도 특히 욕실이 넓어야 한다는 요구사항을 만족시키는 호텔을 찾기 위해서는 상당수의 기존 리뷰를 직접 읽어봐야 하며, 이는 호텔 검색 및 선정에 필요한 시간과 노력이 대폭 증가하는 원인이 된다.

<Figure 1>은 일반적인 호텔 정보 사이트에서 제공되는 정보의 한계를 나타내고 있다. 즉 여가활동 정보 사이트를 사용하는 사람들은 접근성, 객실 등의 한 수준의 평가기준에 대한 요구사항을 나타내기 보다는, 접근성 중에서도 지하철, 버스, 지리적 위치, 그리고 객실 중에서도 욕실, 방, 청결도 등 두 수준 또는 이상의 평가기준에 대한 요구사항을 갖는 것이 일반적이다. 하지만



〈Figure 1〉 The Limitation of Traditional Hotel Information Site

현재 운영되고 있는 대부분의 여가활동 정보 사이트는 별도의 평가기준을 제시하지 않거나, 한 수준의 평가기준만을 사용하고 있으므로, 각 평가기준을 구성하는 세부 요소에 대한 특징과 평가기준 별 주요 이슈를 파악하기 위해서는 상당히 많은 수의 리뷰를 직접 읽어야 한다는 불편이 따른다. 물론 평가기준을 세분화하여 사용자가 다양한 수준의 기준에 대한 리뷰를 입력하게끔 유도하는 방식도 고려해볼 수 있다. 하지만 이러한 접근은 지나치게 많은 입력 구분으로 인해 사용자의 불편이 초래된다는 측면, 그리고 평가기준별로 모든 세부기준을 파악하기 어렵다는 측면에서 현실적인 대안이라고 보기 어렵다.

따라서 본 연구에서는 기존 여가활동 정보 사

이트의 한계, 즉 평가기준별로 입력된 리뷰를 신뢰하기 어렵다는 점과 평가기준을 구성하고 있는 세부 내용을 파악하기 어렵다는 점을 극복하기 위한 방안을 제시하고자 한다. 본 연구에서 제안하는 방법론은 사용자가 별도의 구분 없이 입력한 리뷰를 그 내용에 따라 평가기준별로 자동 분류하고, 각 평가기준별 주요 이슈를 발굴하고 요약하여 제공한다. 제안 방법론은 최근 텍스트 분석에 활발하게 사용되고 있는 토픽 모델링(Topic Modeling)에 기반을 두고 있으며, 각 리뷰를 하나의 문서 단위로 사용하는 것이 아니라 리뷰를 개별 리뷰 유닛(Unit)으로 분해한 뒤 평가기준별로 유닛을 재구성하여 분석한다는 측면에서 기존의 토픽 모델링 기반 연구와 큰 차이가 있다고 할 수 있다. 또한 제안 방법론을 호텔 정보 사이트에 적용하여 분석한 결과를 소개함으로써, 제안 방법론의 유용성을 간접적으로 보이고자 한다.

본 논문의 이후 구성은 다음과 같다. 2장에서는 본 연구와 직간접적으로 관련된 선행 연구의 성과를 요약하고, 3장에서는 본 연구의 제안 방법론을 간단한 예를 통해 소개한다. 4장에서는 제안 방법론을 실제 호텔 정보 사이트에 적용한 실험 결과를 소개하고, 마지막 장인 5장에서는 본 연구의 기여 및 한계, 후속 연구의 방향을 제시한다.

2. 관련 연구

웹사이트를 통한 정보공유가 활발해짐에 따라 사용자 리뷰의 생산과 소비가 활발해지고 있다. 이에 따라 특정 제품에 대한 리뷰들이 해당 기업과 관심 고객에게 필요한 정보를 제공하게 되었고(Archak et al., 2011), 다양한 형태로 제공되는 리뷰는 제품을 선택하는 중요한 요소로 자리 잡게 되었다. 사용자 리뷰는 여행, 영화, 그리고 쇼핑몰 등 특정 도메인에 종속되지 않고 다양한 곳에서 생산되고 있으며(Liu and Kim, 2015), 웹상의 거의 모든 소비 공간에서 필수 불가결한 요소로 인식되고 있다. 그러나 리뷰들은 사용자에 의해 자연어(Natural Language) 형태로 작성되어 비정형 또는 반정형 데이터로 제공(Buneman, 1997)되기 때문에, 이를 정량화하여 분석하기 위한 연구들이 지속적으로 수행되고 있다. 대표적인 예로 감성분석(Sentiment Analysis)과 감성사전(Lee et al., 2016) 및 토픽 모델링을(Chae et al., 2015) 활용한 연구, 색인어 추출(Term Extraction)과 TF-IDF 분석을 적용한 연구(Jeon and Ahn, 2015) 종속성 네트워크 모델을 이용하여 사용자 리뷰에서 추출된 특징들 간의 연관관계를 분석한 연구(Kim, 2010), 문맥 정보(상품분류, 상품특징, 표현어휘, 리뷰점수)를 활용하여 사용자 리뷰를 분류한 연구(Yang et al., 2009) 등을 들 수 있다. 또한 Yeon et al(2011)는 사용자 리뷰를 분석하기 위해 감성분석과 OLAP(On-Line Analytical Processing)을 활용한 온라인 감성 분석처리(On-Line Sentiment Analytical Processing) 방안을 제시 하였다. 최근에는 오피니언 마이닝(Opinion Mining)과 온톨로지(Ontology)를 활용하여 사용자 리뷰를 분석한 연구(Mun et al., 2016)도 수행된 바 있다.

이와 같이 리뷰 분석을 위한 다양한 연구가 이루어짐과 동시에 비정형 데이터로부터 정량화된 정보의 가독성을 높이고자 하는 시각화 연구도 수행되었다. Lee et al(2009)는 사용자가 정의한 사전을 통해 리뷰를 분류한 후 화면 중심에 게시물을 배치하고, 게시물의 연관성에 따라 분석 결

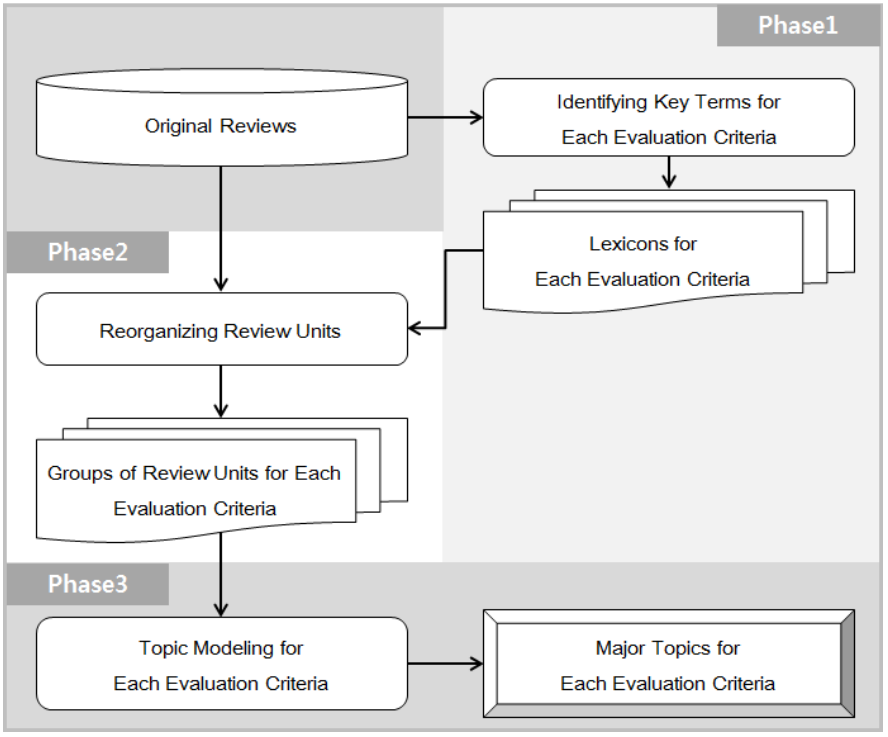
과를 태양계 형태와 유사하게 보여주는 방안을 제시하였다. 또한 이 연구에서는 클러스터링 기법을 사용하여 댓글을 분석하고, 키워드의 클러스터로 분류된 댓글들을 키워드 중심의 방사형으로 시각화하는 방안도 제시하였다. 이와 유사한 형태로 SkepticalLeft와 아고라에서 수집한 댓글을 파형 그래프와 비방향성 그래프로 시각화하는 방법(Lee et al., 2012)도 제시된 바 있다. 한편 호텔 예약 웹서비스인 TripAdvisor와 Booking.com의 리뷰를 분석하여 지도상에 오피니언 마이닝 결과를 시각화하는 방법(Bjørkelund et al., 2012)도 제시되었으나, 이 연구는 호텔의 각 평가기준 별 긍정 또는 부정 여부를 나타내지는 못한다는 한계를 가지고 있다. 최근에는 사용자의 리뷰뿐만 아니라 상품의 분류와 속성을 반영하는 리뷰 감정분석 결과를 워드 클라우드와 그래프로 시각화하는 시도(Choi et al., 2016)가 이루어졌으며, 온톨로지와 오피니언 마이닝을 활용한 분석 결과를 군집으로 시각화하는 새로운 방안도 제시되었다(Mun et al., 2016).

특히 본 연구의 목적과 가장 밀접한 관련이 있는 리뷰 재구성에 대한 연구도 꾸준히 수행된 바 있다. 가장 대표적인 방식은 리뷰를 정량화하여 요약하는 방식이라고 할 수 있다. 상품의 특징을 나타내는 단어의 빈도수(Gamon et al., 2005) 또는 이와 평점이 결합된 형태로 분석 결과를 제시하고(Scaffidi et al., 2007; Yang et al., 2008), 상품의 특징 별로 긍정/부정 여부를 도출하는 연구가 수행되었다(Liu et al., 2005). 최근에는 목적 별 리뷰정보를 도출하는 방안(Kim and Kim, 2016)이 제시되었다. 이 연구에서는 데이터 사전을 통해 특정 목적과 관련된 단어 정보를 추출한 후, 데이터 마이닝 기법인 군집화와 연관 규칙 학습을 통해 상품에 대한 고객들의 평가를 정량적으로 나타냈다. 또한 상품의 속성 별 긍정/부정 여부를 파악하기 위해 속성명 사전을 구축하고, 이에 따라 5개의 속성(분위기, 장소, 서비스, 음식, 가격)으로 리뷰를 분류하여 분석한 연구(Yeon et al., 2013)도 수행된 바 있다. 이 연구는 리뷰를 속성에 따라 분류하였다는 측면에서 본 연구와 유사성이 있으나, 분류된 리뷰의 주요 이슈에 대한 분석이 아닌 긍정/부정 여부를 분석하였다는 측면에서 제안 방법론과는 차이가 있다.

3. 제안 방법론

3.1 연구 범위

본 절에서는 제안하는 방법론의 범위 및 주요 과정에 대해 간략하게 소개한다. 본 장에서 제안하는 방법론은 편의상 호텔 정보 사이트를 예로 들어 기술되지만, 사용자 리뷰가 있는 모든 여가활동 정보 사이트에 동일하게 적용될 수 있다. 본 연구의 전체 모형은 다음 <Figure 2>와 같다. Phase 1은 리뷰 데이터로부터 리뷰에 사용되는 주요 용어를 발굴하여 평가기준 별 용어사전으로 만드는 과정으로, 3.2.1절에서 자세하게 다룬다. Phase 2는 리뷰 데이터로부터 문장 단위의 유닛을 추출한 후 이를 Phase 1의 결과인 평가기준 별 용어사전을 이용해 각 평가기준 별로 분류하는 과정으로, 3.2.2절에서 소개한다. 마지막으로 Phase 3에서는 각 리뷰 유닛 그룹 별 토픽 모델링을 통해 평가기준 별 이슈를 도출하는 과정으로, 이 내용은 3.2.3절에서 다룬다. 이해를 돕기 위해 본 장에서는 간략한 가상 예를 통해 방법론을 소개하며, 제안 방법론을 실제 데이터에 적용한 결과는 4장에서 소개한다.



〈Figure 2〉 Research Model

3.2 평가기준 별 이슈 식별

3.2.1 평가기준 별 용어사전 구축

여가활동 정보 사이트는 그 분야 및 특성에 따라 상이한 평가기준을 가질 수 있다. 예를 들어 영화 정보 사이트는 작품성, 스토리, 배우의 연기력 등의 기준을 가질 수 있고, 음악 정보 사이트는 멜로디, 가수의 가창력, 감정 전달력 등의 기준을 가질 수 있다. 기존의 사이트들은 이러한 평가기준이 명확히 구분되어 있지 않거나, 구분된 경우라도 사용자의 의도적/비의도적 오입력으로 인해 평가기준과 리뷰의 내용이 부합하지 않는 경우가 많다. 제안 방법론은 기존 사이트의 이러한 한계를 극복하기 위해 사용자의 리뷰를 각 평가기준에 따라 자동으로 분류하는 과정을 수행하며, 이 과정에는 평가기준 별 용어사전이 필수적으로 사용된다.

평가기준의 구체적 특성(Feature)을 나타내는 용어사전은 기존에 구축된 도메인별 용어사전 또는 온톨로지를 활용하여 구축할 수 있지만, 본 연구에서는 용어 빈도수 분석 결과를 활용하여 작은 규모의 용어사전을 직접 구축하였다. 구체적으로 우선 전체 리뷰에 출현한 용어 중 명사의 빈도수를 측정한 후, 이들 중 특정 임계값(Threshold) 이상 출현하는 용어만을 도출하였다. 그리고 이들 용어들이 각 평가기준의 어떤 항목의 묘사에 사용될 수 있는지를 2차원 행렬 형태로 정리하였다. 예를 들어 호텔 정보 사이

〈Table 1〉 Example of Lexicon for each Evaluation Criteria of Hotel Information Site

Term	Freq.	Evaluation Criteria					
		Accessibility	*Facility*	*Service*	*Food*	*Room*	*Price*
Location	180	O					
Subway Station	145	O					
Family	120						
Guest Room	101					O	
Kindness	92			O			
Bus	88	O					
Employee	85			O			
Cleaning	75			O		O	
Bathroom	60					O	
Soundproof	46					O	
. . .							

트의 평가기준으로는 접근성(Accessibility), 시설(Facility), 서비스(Service), 음식(Food), 객실(Room), 그리고 가격(Price) 등이 주로 사용되고 있으므로, 각 용어들이 이들 6가지 평가기준에 대응되는 여부를 <Table 1>과 같이 정리할 수 있다.

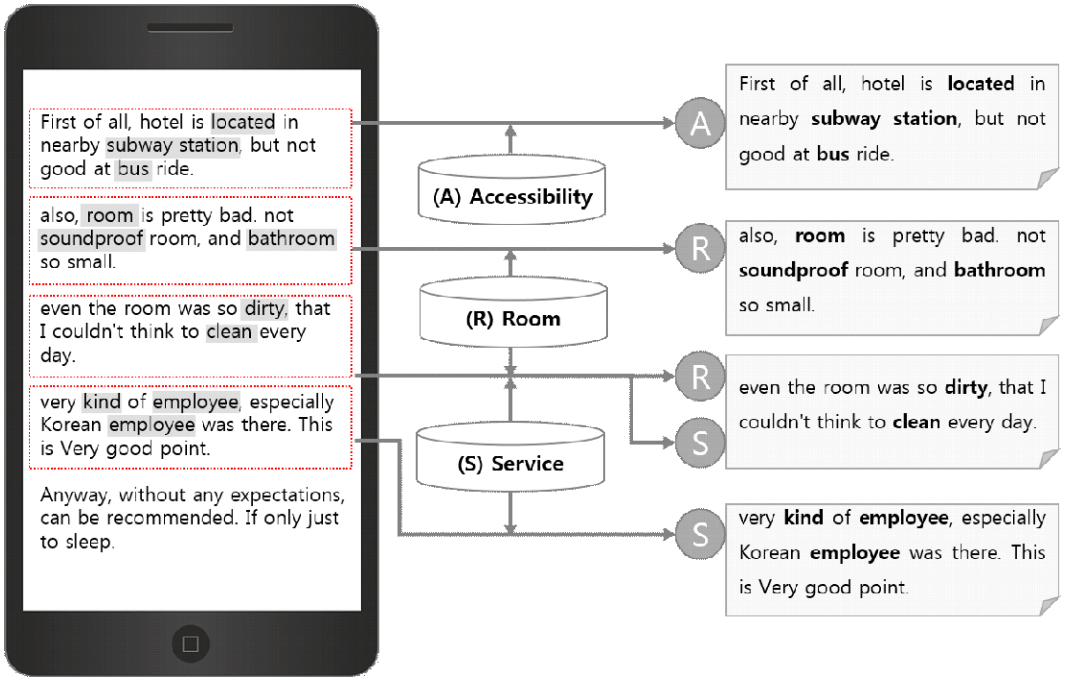
<Table 1>은 호텔 정보 사이트의 접근성, 서비스, 객실의 평가기준에 대한 용어사전에 각각 (위치, 지하철, 버스), (친절, 직원, 청소), 그리고 (객실, 청소, 욕실, 방음)의 용어가 수록된 가상 예를 보이고 있다. <Table 1>에서 용어 “청소”와 같이 여러 평가기준의 용어사전에 동시에 수록되는 용어가 있을 수 있으며, 반대로 “가족”과 같이 빈도수가 높을지라도 특정 평가기준의 특성을 나타내는 것으로 볼 수 없는 경우는 평가기준 별 용어사전에서 제외됨을 알 수 있다. 평가기준 별 용어집 구축 과정에서 발생할 수 있는 오차를 최소화하기 위해 연구원 두 명이 동일한 분류 작업을 수행하였다. 구체적으로 두 연구원의 분류가 일치하는 경우 해당 용어를 곧바로 용어사전에 추가하였으며, 서로 의견이 다른 경우 또 다른 연구원의 의견에 따라 용어를 분류하였다.

3.2.2 리뷰 유닛 재구조화

토픽 모델링을 활용한 대부분의 연구는 하나의 문서를 하나의 연구 단위로 사용한다. 이는 하나의 문서가 하나의 주제(Topic) 또는 하나의 감성(Sentiment 또는 Opinion)을 나타내고 있다는 가정에 기반을 둔다. 하지만 호텔 리뷰의 경우 비교적 짧은 글 내에도 여러 주제 및 감성이 포함되는 경우가 많다. 따라서 각 리뷰를 하나의 연구 단위로 사용하여 토픽 모델링을 수행하게 되면, 여러 평가기준이 뒤섞인 상태의 토픽이 도출되어 결과의 활용도가 매우 낮아지게 된다. 이러한 한계를 극복하기 위해 본 연구에서는 사용자의 리뷰를 평가기준에 따라 분리한 후 평가기준 별 토픽 모델링을 수행하며, 이는 제안 방법론의 가장 독창적인 부분 중 하나이다.

평가기준에 따른 리뷰 분리를 위해 앞에서 평가기준 별 용어사전을 구축하였으며, 본 부절에서는 이를 적용하여 각 리뷰를 리뷰 유닛으로 재구조화하는 과정을 소개한다. 우선 마침표(.) 또는 줄바꿈을 구분자(Delimiter)로 적용하여 각 리뷰를 문장 별로 분리하고 이를 리뷰 유닛이라고 정의한다. 다음으로 <Table 1>을 활용하여 각 리뷰 유닛이 어떤 평가기준에 대한 내용을 언급하고 있는지 파악한다. 예를 들어 <Figure 3>은 특정 호텔에 대한 하나의 리뷰로부터 4개의 리뷰 유닛을 도출하고, 이를 세 가지 평가기준으로 재구조화하는 예를 나타낸다.

<Figure 3>의 좌측은 가상의 리뷰 원문을 나타내며, 가운데 부분은 <Table 1>로부터 도출된 접근성(위치, 지하철, 버스), 객실(객실, 청소, 욕실, 방음), 그리고 서비스(친절, 직원, 청소)의 세 가지 용어사전을 나타낸다. 용어사전에 수록된 어휘는 좌측의 리뷰 원문에서 음영으로 구분되어 있다. 본 예에서 리뷰 원문은 전체 다섯 개의 문장으로 구성되어 있으며, 마지막 문장은 용어사전에 수록된 어떤 용어도 포함하고 있지 않으므로 이 단계에서 폐기된다. 나머지 네 문장의 경우 평가기준 별 용어사전에 수록된 용어의 포함 여부에 따라 해당 그룹의 리뷰 유닛으로 분류된다. 예를 들어 첫 문장은 접근성 그룹, 두 번째 문장은 객실 그룹, 네 번째 문장은 서비스 그룹의 리뷰 유닛으로 분류되며, 세 번째 문장의 경우 객실 그룹과 서비스 그룹에 동시에 속하게 된다. 이러한 과정을 통해 각 평가기준 별 리뷰 유닛의 그룹이 생성되며, 이후 각 그룹 별 토픽 모델링을 통해 평가기준 별 주요 이슈를 도출할 수 있다.



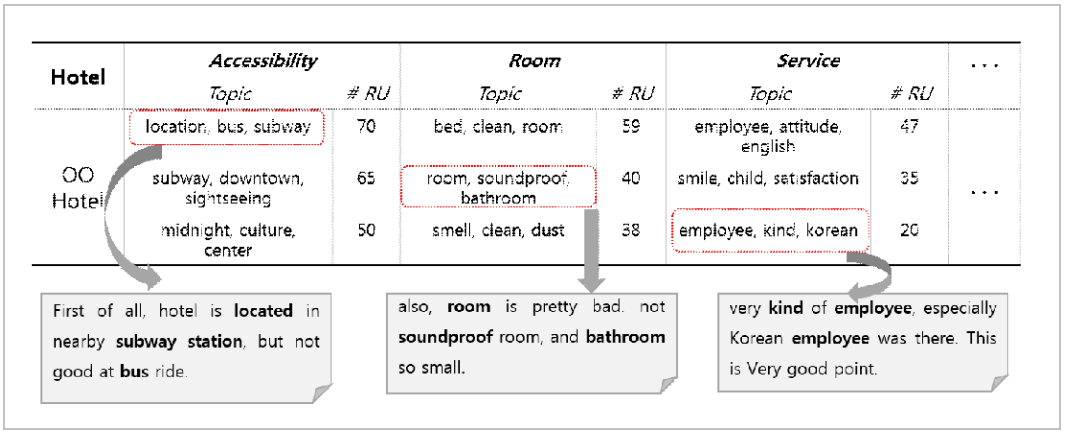
〈Figure 3〉 Review Unit Restructuration

3.2.3 평가기준 별 주요 이슈 도출

본 부절에서는 앞에서 도출한 각 호텔의 평가기준 별 리뷰 유닛에 대한 토픽 모델링 분석을 통해 평가기준 별 주요 이슈를 도출하는 과정을 소개한다. 토픽 모델링은 각 문서에 포함된 용어의 빈도수에 근거하여 유사 문서를 그룹화한 뒤, 각 그룹을 대표하는 주요 용어를 추출하여 해당 그룹의 토픽 키워드 집합을 제시하는 기법이다. 토픽 모델링은 많은 연구 및 서적에서 이미 소개되었을 뿐 아니라 상용 분석 도구를 통해 쉽게 수행 가능하므로, 본 연구에서는 이에 대한 자세한 과정 대신 주요 원리만을 요약하여 소개한다.

우선 전체 용어에 대한 차원 축소가 이루어지게 되며, 이 때 차원의 수가 곧 토픽의 수를 나타내게 된다. 각 용어는 각 차원, 즉 토픽에 대한 대응도를 갖게 되며, 이를 용어 가중치(Term Topic Weight)라고 한다. 이 때 용어 가중치가 주어진 용어 임계값(Term Cutoff) 이상인 경우 이 용어는 해당 토픽을 기술하는 용어로 분류되며, 임계값으로는 주로 각 토픽의 모든 용어 가중치의 "평균 + 1σ(Sigma, 표준편차)"가 사용된다. 또한 각 문서에 대해서도 문서 가중치(Document Topic Weight)가 계산되는데, 이 값은 문서에 포함된 각 용어의 TF-IDF 값과 용어 가중치의 곱의 표준합(Normalized Sum)으로 계산된다. 또한 문서 가중치의 값이 문서 임계값(Document Cutoff) 이상인 경우 이 문서는 해당 토픽에 속하는 문서로 분류되며, 임계값으로는 주로 각 토픽의 모든 문서 가중치의 "평균 + 1σ"가 사용된다. 이러한 방식을 통해 문서 집합으로부터 주요 토픽 및 각 토픽을 기술하는 주요 용어를 도출하고, 각 토픽에 해당되는 문서를 식별할 수 있다. 이러한 일련의 과정을 통해 각 호텔의 평가기준 별 이슈를 도출한 가상 예가 <Figure 4>에 제시되어 있다.

<Figure 4>는 <Figure 3>의 평가기준 별 리뷰 유닛에 대한 토픽 모델링을 통해 OO호텔의 접근성, 객실, 그리고 서비스에 대한 고객 리뷰의 주요 이슈를 요약한 예를 보이고 있다. 그림에서 "Topic"은 각 이슈의 주요 키워드를 나타내며, "#RU"는 각 이슈에 해당되는 리뷰 유닛의 수를 나타낸다. 한편 각 이슈, 즉 토픽에 대해 가장 높은 문서 가중치를 갖는 상위 몇 개의 문장 리뷰



〈Figure 4〉 Example of Main Issues for each Evaluation Criteria

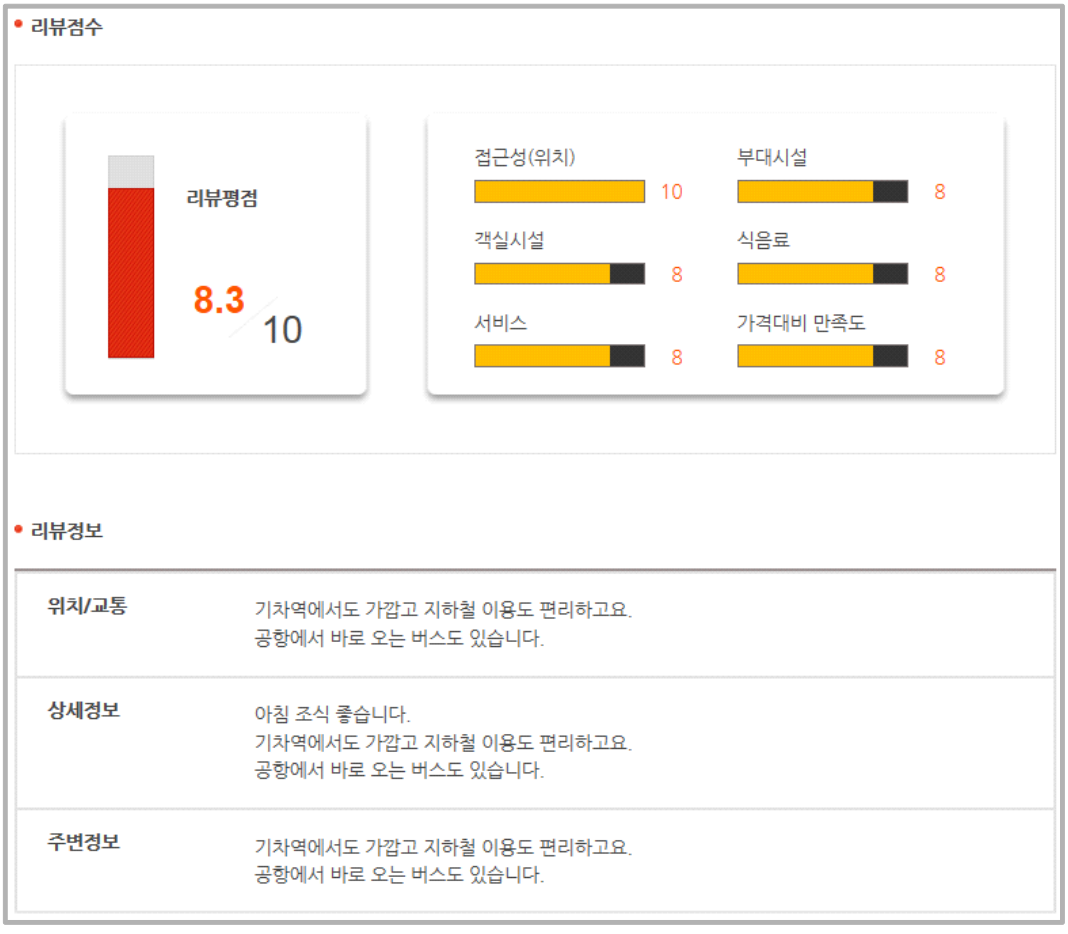
원문을 요약 표와 함께 제시함으로써, 해당 이슈의 내용 이해를 도울 수 있다. 예를 들어 <Figure 4>는 세 가지 이슈 각각에 대해 문서 가중치가 가장 높은 리뷰 유닛 한 가지씩을 제시한 경우를 보이고 있다.

본 장에서는 간략한 가상 예를 통해 제안 방법론을 소개하였다. 다음 장인 4장에서는 실제 데이터에 대해 제안 방법론을 적용한 실험의 과정 및 결과를 제시한다.

4. 실험

4.1 실험 데이터 소개

본 장에서는 실제 운영되고 있는 호텔 정보 사이트로부터 리뷰 데이터를 수집하고, 제안 방법론을 적용하여 각 호텔의 주요 평가기준 별 주요 이슈를 식별하는 실험을 수행한다. 실험 대상으로는 누적고객 470만 명 이상을 갖는 글로벌 호텔 정보 사이트인 'H' 사이트를 선정하였으며,



〈Figure 5〉 Example of Hotel Reviews in Site 'H'

데이터 수집에는 직접 제작한 크롤러를 사용하였다. 'H' 사이트는 접근성, 부대시설, 객실시설, 식음료, 서비스, 그리고 가격대비 만족도의 6개 평가기준을 사용하고 있으며, 전체 평점과 함께 각 기준별 평점도 제공하고 있다. 한편 텍스트 리뷰는 위치/교통, 상세정보, 그리고 주변정보의 세 가지 항목을 구분하여 입력받고 있으나, <Figure 5>와 같이 세 가지 항목을 엄밀하게 구분하지 않고 유사한 내용을 반복 입력한 리뷰가 많이 존재하였다.

따라서 본 실험에서는 3장에서 제안한 방법론을 사용하여 전체 리뷰를 리뷰 유닛으로 분해한 후, 각 문장을 평가기준에 따라 재구성하여 분석하였다. 구체적으로 2005년 8월부터 2016년 5월까지 20건 이상의 리뷰가 등록된 호텔 중 평균 평점이 8점대인 호텔 2개와 7점대인 호텔 3개를 선택하여 분석에 사용하였다. 최초 실험 설계 시에는 각 평점대 별로 리뷰의 수가 가장 많은 호텔을 선택하여 분석을 실시하고자 하였으나, 7점대 미만의 평점을 갖는 호텔의 경우 방문객이 적어 리뷰의 수가 20건을 넘기는 경우가 드문 것으로 나타나, 7점대와 8점대 평점을 갖는 호텔만을 분석에 사용하였다. 이러한 5개 호텔에 대해 수집한 리뷰의 총 수는 423건이며, 문장 분리 후 재구성된 리뷰 유닛의 총 수는 4,860건으로 나타났다. 실험 대상 호텔 5개의 평점, 리뷰 수, 그리고 리뷰 유닛의 수가 <Table 2>에 요약되어 있다.

4.2 평가기준 별 용어사전 구축

본 절에서는 호텔 정보 사이트의 6개 평가기준 별 용어사전 구축 과정 및 결과를 소개한다. 우선 4,860개 리뷰 유닛에 대해 형태소 분석, 파싱(Parsing) 및 필터링(Filtering)을 수행하였으며, 이 과정에서 SAS Enterprise Miner 14.1의 Text Parsing 및 Text Filtering 모듈을 사용하였다. 또한 불필요한 어휘를 제거하기 위해 이메일, URL, 기타 무의미한 단어 등 총 68,822개의 어휘를 수록한 불용어 사전(Stop List)을 적용하여 분석 결과를 정제하였다. 이렇게 정제된 각 용어를 평가기준 별 용어사전에 수록하는 과정은 두 연구원에 의해 동시에 수행되었다. 두 연구원의 분류가 일치하는 경우 해당 용어를 곧바로 용어사전에 추가하였으며, 서로 의견이 다른 경우 또 다른 연구원의 의견에 따라 용어를 분류하였다. 각 용어사전에 수록된 어휘의 수가 충분할 경우는 빈도수 3 미만의 용어는 사전에서 제거하였으며, 어휘의 수가 충분하지 않은 사전의 경우는

〈Table 2〉 Rating, Number of Reviews, Number of Review Units of the Five Hotels

Hotel	*Ratings*	*# of Reviews*	*# of Review Units*
Hotel A	8.33	128	1,962
Hotel B	8.36	89	919
Hotel C	7.83	92	849
Hotel D	7.85	92	775
Hotel E	7.86	22	355
	Average Ratings	**Total # of Reviews**	**Total # of Review Units**
	8.05	423	4,860

Accessibility (Total 24 Terms)

Terms	Freq.
역	463
버스	210
도보	159
근처	75
지하철	43
· · ·	

Facility (Total 26 Terms)

Terms	Freq.
시설	41
인터넷	10
식당	8
뷔페	4
주차장	3
· · ·	

Service (Total 21 Terms)

Terms	Freq.
서비스	35
친절	21
직원	11
팁	5
일본어	3
· · ·	

Food (Total 18 Terms)

Terms	Freq.
조식	16
식사	12
음료	10
커피	10
생수	5
· · ·	

Room (Total 36 Terms)

Terms	Freq.
객실	62
넓다	17
화장실	5
침대	5
냉장고	3
· · ·	

Price (Total 20 Terms)

Terms	Freq.
가격대	56
만족	30
저렴	20
비싸다	12
싸다	4
· · ·	

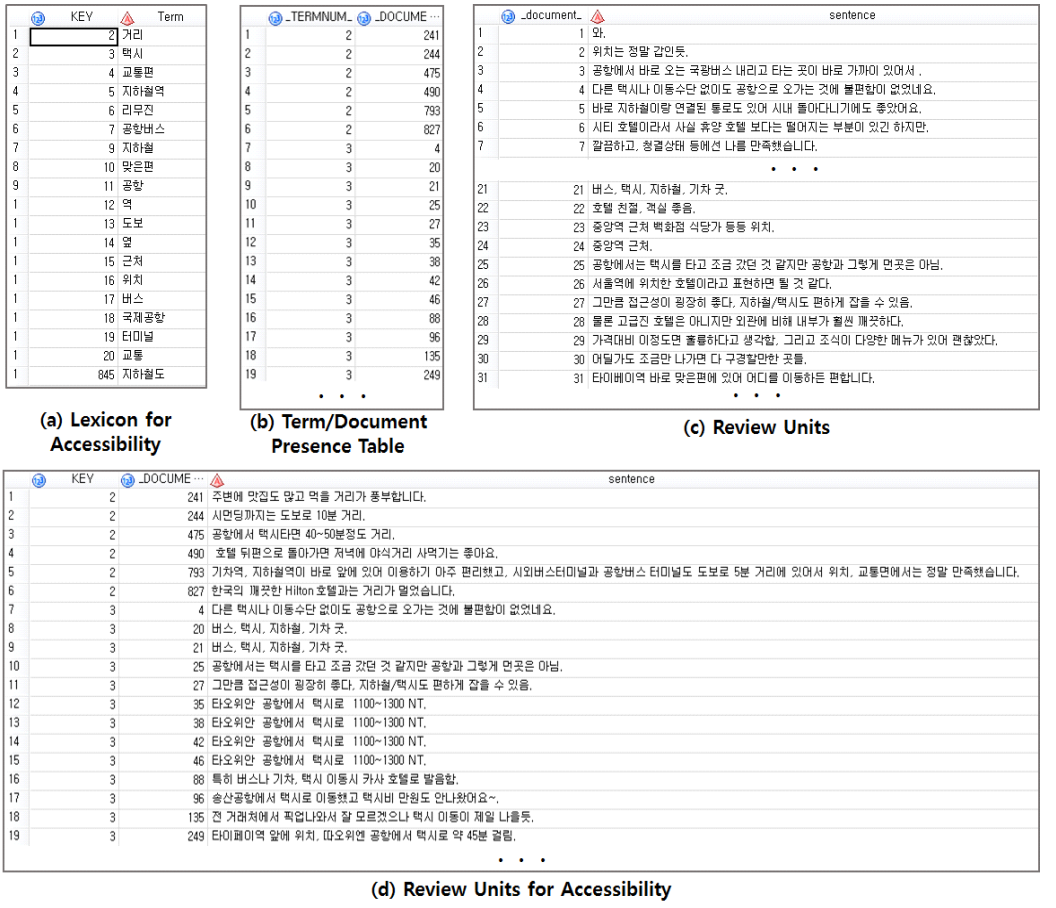
〈Figure 6〉 Lexicon of Each Evaluation Criteria (Part)

빈도수와 관계없이 전체 용어를 사전에 수록하였다. 이러한 과정을 거쳐 총 6개의 평가기준에 대해 전체 145개의 용어가 식별되었으며, 각 용어사전의 내용 일부가 <Figure 6>에 나타나있다.

4.3 리뷰 유닛 재구조화

본 절에서는 6개의 평가기준 별 용어사전을 사용하여 423개의 리뷰 원문을 4,860개의 리뷰 유닛으로 재구조화한 결과를 소개한다. 이 과정 역시 SAS Enterprise Miner 14.1의 Text Parsing을 주로 사용하였으며, 결과의 후처리에는 SAS Enterprise Guide 7.1을 사용하였다. 6개 평가기준 별 리뷰 유닛 도출 과정은 서로 유사하므로, 본 절에서는 접근성 기준에 대한 분석 과정만을 <Figure 7>에 소개한다.

<Figure 7(a)>는 접근성 용어사전의 수록 어휘를 나타내며, <Figure 7(b)>는 각 어휘가 출현한 문서를 나타낸다. 또한 리뷰 원문은 <Figure 3>



	KEY	Term
1	2	거리
2	3	택시
3	4	교통편
4	5	지하철역
5	6	리무진
6	7	공항버스
7	9	지하철
8	10	맞은편
9	11	공항
1	12	역
1	13	도보
1	14	옆
1	15	근처
1	16	위치
1	17	버스
1	18	국제공항
1	19	터미널
1	20	교통
1	845	지하철도

(a) Lexicon for Accessibility

	TERMNUM	_DOCUME…
1	2	241
2	2	244
3	2	475
4	2	490
5	2	793
6	2	827
7	3	4
8	3	20
9	3	21
10	3	25
11	3	27
12	3	35
13	3	38
14	3	42
15	3	46
16	3	88
17	3	96
18	3	135
19	3	249

(b) Term/Document Presence Table

	document	sentence
1	1	와.
2	2	위치는 정말 갑인듯.
3	3	공항에서 바로 오는 국광버스 내리고 타는 곳이 바로 가까이 있어서 .
4	4	다른 택시나 이동수단 없이도 공항으로 오가는 것에 불편함이 없었네요.
5	5	바로 지하철이랑 연결된 통로도 있어 시내 돌아다니기에도 좋았어요.
6	6	시티 호텔이라서 사실 휴양 호텔 보다는 떨어지는 부분이 있긴 하지만.
7	7	깔끔하고, 청결상태 등에선 나름 만족했습니다.
		• • •
21	21	버스, 택시, 지하철, 기차 굿.
22	22	호텔 친절, 객실 좋음.
23	23	중앙역 근처 백화점 식당가 등등 위치.
24	24	중앙역 근처.
25	25	공항에서는 택시를 타고 조금 갔던 것 같지만 공항과 그렇게 먼곳은 아님.
26	26	서울역에 위치한 호텔이라고 표현하면 될 것 같다.
27	27	그만큼 접근성이 굉장히 좋다, 지하철/택시도 편하게 잡을 수 있음.
28	28	물론 고급진 호텔은 아니지만 외관에 비해 내부가 훨씬 깨끗하다.
29	29	가격대비 이정도면 훌륭하다고 생각함, 그리고 조식이 다양한 메뉴가 있어 괜찮았다.
30	30	어딜가도 조금만 나가면 다 구경할만한 곳들.
31	31	타이베이역 바로 맞은편에 있어 어디를 이동하든 편합니다.
		• • •

(c) Review Units

	KEY	_DOCUME…	sentence
1	2	241	주변에 맛집도 많고 먹을 거리가 풍부합니다.
2	2	244	시먼딩까지는 도보로 10분 거리.
3	2	475	공항에서 택시타면 40~50분정도 거리.
4	2	490	호텔 뒤편으로 돌아가면 저녁에 야식거리 사먹기는 좋아요.
5	2	793	기차역, 지하철역이 바로 앞에 있어 이용하기 아주 편리했고, 시외버스터미널과 공항버스 터미널도 도보로 5분 거리에 있어서 위치, 교통면에서는 정말 만족했습니다.
6	2	827	한국의 깨끗한 Hilton 호텔과는 거리가 멀었습니다.
7	3	4	다른 택시나 이동수단 없이도 공항으로 오가는 것에 불편함이 없었네요.
8	3	20	버스, 택시, 지하철, 기차 굿.
9	3	21	버스, 택시, 지하철, 기차 굿.
10	3	25	공항에서는 택시를 타고 조금 갔던 것 같지만 공항과 그렇게 먼곳은 아님.
11	3	27	그만큼 접근성이 굉장히 좋다, 지하철/택시도 편하게 잡을 수 있음.
12	3	35	타오위안 공항에서 택시로 1100~1300 NT.
13	3	38	타오위안 공항에서 택시로 1100~1300 NT.
14	3	42	타오위안 공항에서 택시로 1100~1300 NT.
15	3	46	타오위안 공항에서 택시로 1100~1300 NT.
16	3	88	특히 버스나 기차, 택시 이동시 카사 호텔로 말음함.
17	3	96	송산공항에서 택시로 이동했고 택시비 만원도 안나왔어요~.
18	3	135	전 거래처에서 픽업나와서 잘 모르겠으나 택시 이동이 제일 나을듯.
19	3	249	타이페이역 앞에 위치, 따오위엔 공항에서 택시로 약 45분 걸림.
			• • •

(d) Review Units for Accessibility

〈Figure 7〉 Results of Review Unit Restructuration (Part)

에 소개된 방법에 의해 우선 문장 단위로 분리되며 <Figure 7(c)>와 같은 형태로 나타나며, 이들 문장 중 접근성 용어사전의 수록 어휘를 포함한 문장, 즉 <Figure 7(b)>에 나타난 문장만이 접근성 리뷰 유닛으로 구분된다. 이렇게 도출된 접근성 리뷰 유닛의 결과 일부가 <Figure 7(b)>에 나타나있다. <Figure 7>의 과정을 각 호텔 5개 및 각 평가기준 6개에 대해 총 30번 반복 수행함으로써, 5개 호텔의 6개 평가기준에 대한 리뷰 유닛을 구성하였다.

4.4 평가기준 별 주요 이슈 도출

본 절에서는 본 실험의 마지막 단계로, 앞에서 추출한 각 호텔의 평가기준 별 리뷰 유닛에 대한 토픽 모델링을 수행한 결과를 요약한다. 토픽 모델링에는 SAS Enterprise Miner 14.1의 Text Topic 모듈을 사용하여 수행하였으며, 4.2절에서 소개한 불용어 사전을 사용하여 토픽 키워드를 정제하였다. 이러한 과정을 통해 도출된 5개 호텔의 6개 평가기준에 대한 주요 이슈 및 각 이슈

	Accessibility		Facilities		Room	
	Topic	*# Doc.*	*Topic*	*# Doc.*	*Topic*	*# Doc.*
A Hotel	위치,+좋다,기차역,최고,건너편	38	시설,가격대,+깨끗하다,+만족스럽다,+오래되다	10	욕실,+없다,+넓다,객실,리노베이션	19
	공항,택시,버스,리무진,근처	36	부페,아침,+괜찮다,친절,리모델링	6	냄새,화장실,+좋다,리노베이션,+오래다	17
	교통,+좋다,+편리하다,객실,+편하다	33	+많다,주변,식당,편의점,지하	5	객실,+깨끗하다,상태,청소,+좋다	10
B Hotel	버스,정류장,중앙,+가깝다,리무진	70	인터넷,무료,+좋다,노트북,사용	19	객실,+좁다,+깔끔하다,+크다,+없다	38
	공항,버스,시간,노선,+되다	65	식당,지하,+작다,커피숍,식사	19	+깨끗하다,욕실,냄새,객실,화장실	35
	위치,+좋다,+아니다,+만족스럽다,가격	50	데스크,직원,프런트,친절,무료	18	화장실,+좋다,비데,인터넷,+크다	25
C Hotel	백화점,맞은편,+편리하다,버스,편의점	31	+시설,쇼핑,백화점,접근성,+편리하다	17	+좁다,객실,+깔끔하다,트윈,+괜찮다	17
	위치,최고,+없다,+크다,+편리하다	30	건물,지하,+깔끔하다,+좋다,백화점	15	+좋다,전망,위치,+괜찮다,나름	17
	출구,남쪽,오른쪽,삿뽀로,건물	22	식당,+좋다,음식,아침,백화점	12	냄새,화장실,담배,+없다,욕실	15
D Hotel	위치,최고,+없다,+좋다,도보	40	부페,아침,+맛있다,+괜찮다,+훌륭하다	8	+깨끗하다,객실,+괜찮다,친절,지하철	15
	교통,최고,+편리하다,메인,장점	36	건물,+크다,+없다,+깨끗하다,불편	8	냄새,+오래다,공팡이,처음,담배	11
	공항,버스,종점,공항버스,터미널	33	직원,+친절하다,+낡다,친절,+많다	7	창문,하루,에어컨,+없다,욕실	8
E Hotel	도보,근처,골목,중심부,+가능하다	9	인터넷,무선,무료,사용,+가능하다	3	+깨끗하다,화장실,싱글,냄새,+그렇다	6
	+지하철,근처,직진,생각,+편하다	9	건물,+없다,부대시설,+좋다,주차장	4	+없다,욕실,넹장고,+불편하다,건물	8
	위치,+좋다,기차역,+가깝다,근처	7	아침,커피,오후,시설,+좋다	2	+좁다,더블,+아니다,침대,객실	7

	Food		Service		Price	
	Topic	*# Doc.*	*Topic*	*# Doc.*	*Topic*	*# Doc.*
A Hotel	음식,지하철,+나쁘다,+유명하다,주변	14	서비스,+좋다,위치,가격대,영어	10	위치,서비스,+좋다,가격,+크다	5
	+많다,주변,백화점,편의점,근처	14	+친절하다,직원,서비스,+없다,교통	9	객실,리모델링,+깨끗하다,+아니다,고급	4
	식사,아침,+많다,+맛있다,+나쁘다	11	청소,객실,상태,+괜찮다,+눕다	8	가격대,+괜찮다,+훌륭하다,시설,+만족하다	3
B Hotel	식사,아침,+훌륭하다,가격,길거리	16	친절,+깨끗하다,직원,데스크,+아니다	27	가격,저렴,식사,+만족스럽다,+깨끗하다	20
	지하,식당,친절,+작다,+맛있다	11	영어,일본어,직원,아주머니,남자	23	가격대,+좋다,+훌륭하다,+괜찮다,+깨끗하다	18
	부페,+없다,+간단하다,일식당,+괜찮다	11	+친절하다,직원,서비스,+깨끗하다,+가능하다	21	만족,+깨끗하다,시설,친절,가격대	14
C Hotel	커피,무료,자판기,인터넷,전용	11	체크인,카운터,처음,객실,컴퓨터	7	가격대,+좋다,만족도,내부,+높다	7
	+많다,사람,분위기,+조용하다,식사	9	객실,청소,+없다,+불편하다,카운터	6	만족,시설,객실,위치,사용	7
	쇼핑,백화점,주변,식사,지하	8	+친절하다,직원,한국어,영어,소통	5	가격,+저렴하다,+만족하다,+좋다,+깔끔하다	5
D Hotel	식사,아침,수준,+만족스럽다,+많다	11	서비스,+좋다,위치,친절,+깨끗하다	13	만족,위치,교통,친절,결론	6
	+좋다,직원,+깔끔하다,+괜찮다,식사	9	+친절하다,직원,+깔끔하다,시설,+좋다	12	가격대,객실,+괜찮다,시설,+좋다	6
	+깨끗하다,객실,음식,+괜찮다,+깔끔하다	6	친절,직원,무지,사탈,+낡다	12	가격,+저렴하다,교통,+없다,객실	5
E Hotel	+없다,생수,음료,객실,서비스	4	+친절하다,직원,+좋다,+깔끔하다,시간	5	+저렴하다,+만족스럽다,가격,장점,조용	2
	김밥,삼각,과일,샐러드,우유	4	사용,친절,인터넷,+가능하다,데스크	4	안해,야지,만족,목격,+크다	1
	식사,아침,+간단하다,+많다,+아니다	6	서비스,음료,+좋다,아침,무료	4	+나쁘다,+낮다,목기,사탈,만족도	2

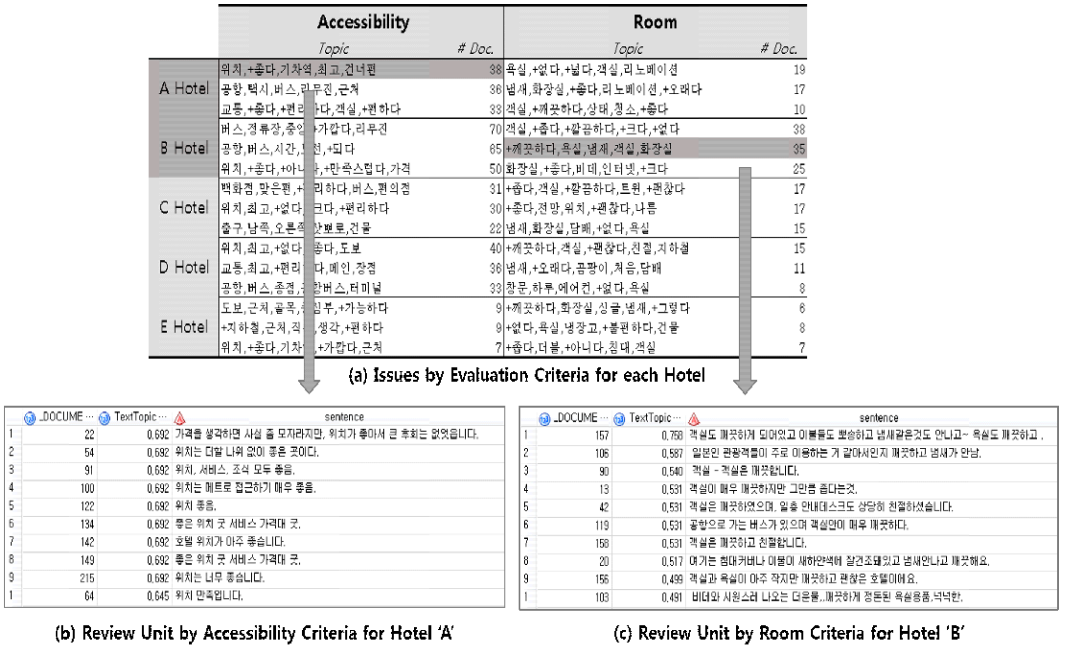
〈Figure 8〉 Main Issues of Each Evaluation Criteria for Five Hotels

에 대응되는 리뷰 유닛의 수를 요약한 결과가 <Figure 8>에 나타나있다.

<Figure 8>은 제안 방법론의 최종 결과물 중 하나로, 각 호텔별 평가기준에 대해 상위 이슈 3개씩을 키워드로 제공하고 있다. 상위 이슈는 각 이슈에 대응되는 문서의 수가 많은 순으로 선정하였다. 이와 같은 형태의 리뷰 요약 테이블을 통해 각 호텔의 각 평가기준에 대해 어떤 이슈가 있는지, 그리고 각 이슈를 다룬 리뷰 유닛의 수는 얼마인지 쉽게 파악할 수 있을 것으로 기대한다.

하지만 많은 이슈의 경우 <Figure 8>의 표에 제시된 정보만으로도 어떤 내용인지 구체적으로 파악이 가능한 반면, 이슈 키워드만으로는 구체적인 내용의 파악이 어려운 이슈도 존재한다. 예를 들어 A 호텔의 접근성 평가기준에 대한 이슈 중 “위치, 좋다, 기차역, 최고, 건너편” 이슈의 경우 ‘A 호텔이 기차역 부근에 존재하여 접근성이 매우 좋다’는 내용을 담고 있을 것으로 충분히 추측이 가능한 반면, B 호텔의 객실 평가기준에 대한 이슈 중 “깨끗하다, 욕실, 냄새, 객실, 화장실” 이슈의 경우 ‘깨끗하다’와 ‘냄새’라는 상충되는 용어가 존재하기 때문에 그 내용을 명확히 짐작하기에 어려움이 있다. 따라서 제안 방법론은 두 번째 최종 결과물로 <Figure 9>와 같이 각 이슈에 대응되는 리뷰 유닛 중 상위 10개를 함께 제공한다. 상위 리뷰 유닛은 해당 리뷰 유닛의 문서 가중치가 높은 순으로 선정하였다.

<Figure 9>는 제안 방법론의 일종의 활용 시나리오를 보이고 있다. 예를 들면 각 호텔별로



〈Figure 9〉 Correspondence Matrix Between Issues and Documents

<Figure 9(a)>의 요약 테이블을 제공한 뒤, 관심 있는 이슈를 선택함으로써 <Figure 9(b)> 또는 <Figure 9(c)>가 표시되는 형태로 서비스를 구현할 수 있다. 이러한 구현을 통해 앞에서 언급했던 B호텔의 객실에 대한 이슈, 즉 "깨끗하다, 욕실, 냄새, 객실, 화장실" 이슈의 경우 '객실과 욕실이 모두 깨끗하며 냄새가 나지 않았다'는 내용임을 파악할 수 있다.

본 장에서는 실제 호텔 정보 사이트의 리뷰에 대한 실험을 통해, 다양한 평가기준에 대한 내용이 혼재되어 있는 리뷰로부터 제안 방법론을 적용하여 각 평가기준 별 리뷰를 발췌할 수 있을 뿐 아니라 이들 리뷰를 평가기준 별로 요약하여 제공함으로써 각 평가기준에 대한 상세 이슈를 일목요연하게 파악할 수 있음을 보였다.

5. 결론

여가활동 정보 사이트에 대한 수요와 활용도가 높아짐에 따라, 정형/비정형 분석을 통해 사용자가 제공한 정보를 다양한 형태로 가공하여 제공하기 위한 시도가 활발하게 이루어지고 있다. 하지만 기존 대부분의 여가활동 정보 사이트들은 상품 및 서비스의 평가기준이 명확히 구분되어 있지 않거나, 구분된 경우라도 사용자의 의도적/비의도적 오입력으로 인해 평가기준과 리뷰의 내용이 부합하지 않는 경우가 많다. 또한 평가기준을 세부적으로 구분하여 제공하기 어렵기 때문에, 실제로 사용자가 궁금해 하는 부분에 대해 충분한 정보를 제공하기 어렵다는 한계를 갖는다. 따라서 본 연구에서는 다양한 평가기준

에 대한 내용이 혼재되어 있는 리뷰로부터 각 평가기준 별 리뷰 유닛을 재구성하고, 이들 리뷰 유닛의 주요 이슈를 평가기준 별로 요약하여 제공함으로써 각 상품 및 서비스의 평가기준 별 상세 이슈를 요약하여 제공하는 방법론을 제안하였다. 또한 누적고객 470만 명 이상을 갖는 글로벌 호텔 정보 사이트인 'H' 사이트에 소개된 호텔 5곳을 선정하여 리뷰 423개를 수집하고, 이를 4,860개의 리뷰 유닛으로 재구조화하여 각 호텔의 접근성, 부대시설, 객실시설, 식음료, 서비스, 그리고 가격대비 만족도 측면에서의 상세 이슈를 발굴하여 제시하였다.

제안 방법론은 학술적, 실무적 차원에서 다음과 같은 기여를 갖는 것으로 판단한다. 우선 학술적 측면에서 제안 방법론은 하나의 텍스트 문서를 평가기준 별로 재구조화하는 방안을 제시했다는 의의를 갖는다. 물론 문서 전체가 아닌 문장 단위 또는 속성 단위의 분석에 대한 필요성 및 방안이 이미 소개되어 있지만, 각각 분석 결과의 신뢰성과 분석의 어려움으로 인해 널리 활용되고 있지 못하는 측면이 있다. 이에 비해 제안 방법론은 기본적으로 문장 단위의 분석을 사용하되 평가기준 별 용어사전을 통해 문장을 선별적으로 사용함으로써 분석의 용이성과 결과의 신뢰성 측면에서 우수성을 나타낼 것으로 기대한다.

이와 같은 학술적 기여 외에 본 연구의 기여는 실무적 측면에서 더욱 크게 나타날 것으로 기대한다. 많은 사이트를 통해 다양한 상품 및 서비스에 대한 정보를 얻을 수 있지만, 오히려 방대한 양의 정보로 인해 정작 사용자가 궁금해 하는 부분에 대한 정보를 얻기는 더욱 어려워진 측면이 있다. 이에 본 연구는 각 상품 및 서비스의 주요 평가기준을 정의하고 각 평가기준 별 주요 이슈를 요약하여 제시함으로써, 사용자가 필요로 하는 정보를 편리하게 습득할 수 있는 방안을 제시하였다. 제안 방법론은 호텔 정보 사이트를 중심으로 소개되었지만 여타의 여가활동 정보 사이트, 나아가서는 사용자의 리뷰를 다루는 모든 사이트의 개선에 적용될 수 있을 것으로 기대한다. 특히 제안 방법론은 기존의 리뷰를 평가기준 별로 자동으로 구분하는 과정을 포함하고 있기 때문에, 별도의 평가기준을 갖고 있지 않은 기존 사이트의 리뷰 제공 체계 개선에도 크게 기여할 수 있을 것이다.

이와 같은 기여에도 불구하고, 본 연구에서 제안하는 방법론은 향후 다음과 같은 측면에서 보완될 필요가 있다. 우선 제안 방법론의 성능에 대한 검증이 이루어질 필요가 있다. 연구 주제의 특성상 제안 방법론의 성능은 정확도 측면이 아닌 효과성 측면에서 이루어지는 것이 바람직하다. 따라서 설문조사 등을 통해 제안 방법론을 적용한 사이트 만족도를 측정함으로써, 본 방법론의 효과를 평가할 필요가 있다. 또한 보다 많은 사이트에 대한 추가 실험을 진행하여 방법론을 보완할 필요가 있다. 실제로 제안 방법론은 실제 사이트에 대한 실험을 수정하는 과정에서 꾸준히 보완되고 구체화되었다. 예를 들어 사용자의 오입력으로 인해 평가기준과 리뷰의 내용이 부합하지 않는 경우가 많다는 점, 리뷰에 비속어가 많이 존재한다는 점, 그리고 리뷰의 양이 충분하지 않을 뿐 아니라 특정 인기 상품에 편중되어 있다는 점 등이 실험 과정에서 파악되었다. 향후 더욱 많은 사이트에 대한 실험을 통해 다양한 현상을 파악하고, 이를 극복하는 방식으로 방법론을 더욱 견고하게 보강할 필요가 있다. 마지막으로 본 연구의 평가기준 별 용어사전 구축은 연구원의 수작업으로 이루어졌다. 물론 용어사

전의 정확성 향상을 위해 복수의 연구원이 동일 용어를 식별하는 방식을 채택하였지만, 이 과정에서 연구원의 주관이 개입되었을 가능성을 완전히 배제하기는 어렵다. 따라서 향후 특정 용어를 식별하여 평가기준 별 용어사전을 자동으로 구축하기 위한 후속 연구가 반드시 필요하다.

참고문헌(References)

Archak, N., A. Ghose and P. G. Ipeirotis, "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science*, Vol.57, No.8(2011), 1485~1509.

Bjørkelund, E., T. H. Burnett and K. Nørvåg, "A Study of Opinion Mining and Visualization of Hotel Reviews," *In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS '12)*, 2012.

Buneman, P., "Semistructured data," In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '97)*, 1997.

Chae, S. H., J. I. Lim and J. Y. Kang, "A Comparative Analysis of Social Commerce and Open Market Using User Reviews in Korean Mobile Commerce," *Journal of Intelligence and Information Systems*, Vol.21, No.4(2015), 53~77.

Choi, J. U., H. J. Ryu, D. B. Yu, N. R. Kim and Y. H. Kim, "System Design for Analysis and Evaluation of E-commerce Products Using Review Sentiment Word Analysis," *KIISE Transactions on Computing Practices*, Vol.22, No.5(2016), 209~217.

Gamon, M., A. Aue, S. Corston-Oliver and E. Ringger, "Pulse: Mining Customer Opinions from Free Text," *In Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis (IDA '05)*, 2005.

Jeon, B. K. and H. C. Ahn, "A Collaborative Filtering System Combined with Users' Review Mining : Application to the Recommendation of Smartphone Apps," *Journal of Intelligence and Information Systems*, Vol.21, No.2(2015), 1~18.

Kim, J. Y. and D. S. Kim, "A Study on the Method for Extracting the Purpose-Specific Customized Information from Online Product Reviews based on Text Mining," *The Journal of Society for e-Business Studies*, Vol.21, No.2(2016), 151~161.

Kim, K. H., "Design and Implementation Online Customer Reviews Analysis System based on Dependency Network Model," *The Journal of the Korea Contents Association*, Vol.10, No.11(2010), 30~37.

Lee, S. H., J. Cui and J. W. Kim, "Sentiment analysis on movie review through building modified sentiment dictionary by movie genre," *Journal of Intelligence and Information Systems*, Vol.22, No.2(2016), 97~113.

Lee, Y. J., J. H. Ji, G. Woo and H. G. Cho, "TRIB: A Clustering and Visualization System for Responding Comments on Blogs," *The KIPS Transactions : Part D*, Vol.16, No.5(2009), 817~824.

Lee, Y. J., I. J. Jung and G. Woo, "Extracting and Visualizing Dispute comments and Relations on Internet Forum Site," *The Journal of the*

Korea Contents Association, Vol.12, No.2 (2012), 40~51.

Liu, B., M. Hu and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *In Proceedings of the 14th International Conference on World Wide Web (WWW '05)*, 2005, 342~351.

Liu, C. and N. Kim, "Methodology for Improving the Reliability of the Rating System for Leisure Activity Information Sites : Focusing on a Movie Information Site," *Journal of Tourism and Leisure Research*, Vol.27, No.7(2015), 187~200.

Mun, S. M., G. N. Kim, G. C. Choi and K. W. Lee, "Movie Recommended System base on Analysis for the User Review utilizing Ontology Visualization," *Design Convergence Study*, Vol.15, No.2(2016), 347~368.

Scaffidi, C., K. Bierhoff, E. Chang, M. Felker, H. Ng and C. Jin, "Red Opal: Product-Feature Scoring from Reviews," *In Proceedings of the 8th ACM Conference on Electronic Commerce (EC '07)*, 2007.

Yang, J. Y., J. S. Myung and S. G. Lee, "A Product Review Summarization System Using a Scoring of Features," *The Journal of Society for e-Business Studies Symposium and other publications*, Society for e-Business Studies, 2008.

Yang, J. Y., J. S. Myung and S. G. Lee, "A Sentiment Classification Method Using Context. Information in Product Review Summarization," *Journal of KISS: Databases*, Vol.36, No.4(2009), 254~262.

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.

Yeon. H. B., S. J. Yoo, H. S. Jang, D. I. Han and Y. Jang, "Design and Implementation of a Web Crawling System for the Reviews of Korean Restaurants in the U.S.," *Korea Computer Congress Symposium*, Vol.2013, No.6(2013), 283~285.

Abstract

Methodology for Identifying Issues of User Reviews from the Perspective of Evaluation Criteria: Focus on a Hotel Information Site

Sungho Byun*·Donghoon Lee*·Namgyu Kim**

As a result of the growth of Internet data and the rapid development of Internet technology, "big data" analysis has gained prominence as a major approach for evaluating and mining enormous data for various purposes. Especially, in recent years, people tend to share their experiences related to their leisure activities while also reviewing others' inputs concerning their activities. Therefore, by referring to others' leisure activity-related experiences, they are able to gather information that might guarantee them better leisure activities in the future. This phenomenon has appeared throughout many aspects of leisure activities such as movies, traveling, accommodation, and dining. Apart from blogs and social networking sites, many other websites provide a wealth of information related to leisure activities. Most of these websites provide information of each product in various formats depending on different purposes and perspectives. Generally, most of the websites provide the average ratings and detailed reviews of users who actually used products/services, and these ratings and reviews can actually support the decision of potential customers in purchasing the same products/services. However, the existing websites offering information on leisure activities only provide the rating and review based on one stage of a set of evaluation criteria. Therefore, to identify the main issue for each evaluation criterion as well as the characteristics of specific elements comprising each criterion, users have to read a large number of reviews. In particular, as most of the users search for the characteristics of the detailed elements for one or more specific evaluation criteria based on their priorities, they must spend a great deal of time and effort to obtain the desired information by reading more reviews and understanding the contents of such reviews.

Although some websites break down the evaluation criteria and direct the user to input their reviews according to different levels of criteria, there exist excessive amounts of input sections that make the whole

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
School of Management Information Systems, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

process inconvenient for the users. Further, problems may arise if a user does not follow the instructions for the input sections or fill in the wrong input sections. Finally, treating the evaluation criteria breakdown as a realistic alternative is difficult, because identifying all the detailed criteria for each evaluation criterion is a challenging task. For example, if a review about a certain hotel has been written, people tend to only write one-stage reviews for various components such as accessibility, rooms, services, or food. These might be the reviews for most frequently asked questions, such as distance between the nearest subway station or condition of the bathroom, but they still lack detailed information for these questions. In addition, in case a breakdown of the evaluation criteria was provided along with various input sections, the user might only fill in the evaluation criterion for accessibility or fill in the wrong information such as information regarding rooms in the evaluation criteria for accessibility. Thus, the reliability of the segmented review will be greatly reduced.

In this study, we propose an approach to overcome the limitations of the existing leisure activity information websites, namely, (1) the reliability of reviews for each evaluation criteria and (2) the difficulty of identifying the detailed contents that make up the evaluation criteria. In our proposed methodology, we first identify the review content and construct the lexicon for each evaluation criterion by using the terms that are frequently used for each criterion. Next, the sentences in the review documents containing the terms in the constructed lexicon are decomposed into review units, which are then reconstructed by using the evaluation criteria. Finally, the issues of the constructed review units by evaluation criteria are derived and the summary results are provided. Apart from the derived issues, the review units are also provided. Therefore, this approach aims to help users save on time and effort, because they will only be reading the relevant information they need for each evaluation criterion rather than go through the entire text of review.

Our proposed methodology is based on the topic modeling, which is being actively used in text analysis. The review is decomposed into sentence units rather than considering the whole review as a document unit. After being decomposed into individual review units, the review units are reorganized according to each evaluation criterion and then used in the subsequent analysis. This work largely differs from the existing topic modeling-based studies. In this paper, we collected 423 reviews from hotel information websites and decomposed these reviews into 4,860 review units. We then reorganized the review units according to six different evaluation criteria. By applying these review units in our methodology, the analysis results can be introduced, and the utility of proposed methodology can be demonstrated.

Key Words : Big Data, Review Analysis, Text Mining, Topic Modeling

Received : August 17, 2016 Revised : September 19, 2016 Accepted : September 24, 2016
Publication Type : Regular Paper(Fast-track) Corresponding Author : Namgyu Kim

저 자 소 개

변 성 호
현재 국민대학교 비즈니스IT전문대학원에서 석사과정에 재학 중이다. 한성대학교 컴퓨터공학 학사 학위를 취득하였으며, 주요 관심분야는 텍스트 마이닝 및 토픽 모델링이다.

이 동 훈
현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이다.
한국방송통신대학교 컴퓨터과학과에서 학사 학위를 취득하고, 국민대학교 비즈니스IT전문대학원에서 비즈니스IT를 전공하여 석사 학위를 취득하였다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 온톨로지 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. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국CRM학회 이사, 한국IT서비스학회지 편집위원, JITAM 편집위원, 한국인터넷정보학회논문지 편집위원을 역임하였으며, 한국생산성본부 TOPCIT 개발사업 자문위원으로 활동 중이다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 데이터베이스 설계 등이다.