Journal of the Korean Institute of Industrial Engineers
Vol. 43, No. 1, pp. 062-071, February 2017.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2017.43.1.062

<Application Research>

여행 사이트 리뷰를 활용한 관광지 만족도 요인 추출 및 평가

조수현 · 김보섭 · 박민식 · 이기창 · 강필성†

고려대학교 산업경영공학부

Extraction of Satisfaction Factors and Evaluation of Tourist Attractions based on Travel Site Review Comments

Suhyoun Cho · Boseop Kim · Minsik Park · Gichang Lee · Pilsung Kang

School of Industrial Management Engineering, Korea University

In order to attract foreign tourists, it is important to understand what factors on domestic tour spots are critically considered and how they are evaluated after visit. However, most of the researches on tour business have collected information from tourists through survey on a small number of tourists, which leads to inaccurate and biased conclusion. In this paper, we suggest a data-driven methodology to figure out tourists' satisfaction factors and estimate sentiment scores on them. To do so, we collected review comments data from popular web site. Latent dirichlet allocation is employed to extract key factors and elastic net is used to estimate sentiment scores. Then, an aggregated evaluation score is generated by combining the factors and the sentiment scores per topics. Our proposed method can be used to recommend travel schedules with themes and discover new spots.

Keywords: Sentiment Analysis, Latent Dirichlet Allocation, Regression, Elastic Net, Tour Evaluation

1. 서 론

관광은 사회 · 문화와 경제 전반에 미치는 파급력이 크다. 사회 · 문화적 측면에서는 국가 간 친선도모, 문화교류는 물론 국가 홍보 효과까지 거둘 수 있으며(Jiang, 2015), 경제적인 측면을 보면 고용 창출, 외화 수입 가득률 증가, 내수시장 활성화 등을 꾀할 수 있다(Jee, 1999). 관광업이 가지는 몇 가지 부정적 효과(인플레이션, 국내 경제의 해외 의존도 심화, 계절·경기 변동에 따른 경제 취약성 증대(Ryu, 2003; Archer and Fletcher, 1996)에도 불구하고 저성장 국면으로 접어든 세계 많은 국가들은 침체된 내수시장을 살리기 위한 방편으로 해외 관광객 유치에 사활을 걸고 있다(Jiang, 2015).

<Figure 1>은 해마다 국내 유입 관광객 수가 증가하고 있으며 관광 수입 또한 그에 비례하여 늘고 있어 국내 관광업이 양적으로 팽창하고 있음을 보여준다. 하지만 국내 관광업이 질적으로도 발전하려면 해외 관광객들의 한국 여행 만족도 요인을 효과적으로 탐지할 방법론이 필요하다.

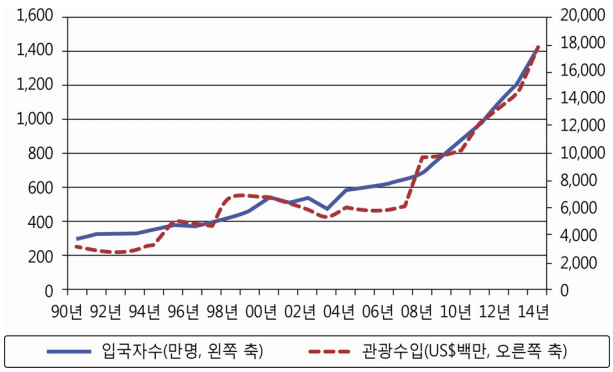

Figure 1. The number of Tourists and Tourist Income Increased Dramatically(Korea Tourism Organization)

제5회 산업융합 활성화 방안 및 사례연구 논문공모전 수상논문.
이 논문은 2016년도 정부(미래창조과학부 및 교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2014R1A1A1004648, No.2015R1A2A2A04007359, NRF-2016R1D1A1B03930729).
† 연락저자 : 강필성 교수, 02841, 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3383, Fax : 02-929-5888, E-mail : pilsung_kang@korea.ac.kr
2017년 1월 9일 접수; 2017년 2월 2일 수정본 접수; 2017년 2월 2일 게재 확정.

국내 방문 해외 관광객을 대상으로 한 여행 행태 조사는 대부분 설문지법을 토대로 하고 있다. 설문지법이란 연구자가 구성한 설문 문항을 토대로 조사 대상자의 취향, 의견 등을 조사하는 방식이다. 대표적인 것이 한국문화관광연구원이 국가적 차원에서 해마다 실시하고 있는 '외래관광객 실태조사'다. 하지만 설문지법을 토대로 한 조사는 설문 문항을 구성하는 과정에서 연구자의 관심이나 주관이 개입될 여지가 있다(Kim, 1995). 연구자가 집중적으로 연구하고자 하는 분야에 더 많은 수의 문항을 배분할 수 있고, 의도치 않게 원하는 방향으로 설문지의 문항이나 답을 구성할 위험이 있기 때문이다. 또한 조사 규모를 대폭 키우지 않는 한 경복궁 등 구체적인 관광명소 관련 정보까지 얻기가 쉽지 않다.

따라서 본 연구에서는 설문지법을 보완할 수 있는 방법론 또는 설문문항 구성을 위한 예비조사 방법론으로써 여행평가 사이트에서 관광지에 대한 평가 댓글과 평점을 활용하여 관광지의 세부 만족도 요인을 추출하고 국내 관광지에 대한 감정 점수를 산출하여 유형별 관광지 만족도 요인 점수를 산출하는 방법론을 제시하였다. 연구를 위해 여행 리뷰 사이트인 Tripadvisor (https://www.tripadvisor.com)에서 영어로 작성된 서울의 103개 명소에 대한 21,620개의 댓글과 평점을 수집하고 103개의 장소를 연구자가 3개의 유형으로 구분하였다. 이후 각 유형의 댓글을 bag of words(BOW)방식의 문서집합(corpus)을 구축하고 이를 잠재 디리클레 할당(Latent dirichlet allocation; LDA)에 적용하여 관광지 유형별 세부 만족도 요인을 도출하였다. 또한 감정분석을 실시하여 관광지 유형 별로 추출한 세부 만족도 요인을 정량적으로 평가하였다. 특히 벌점화 회귀분석(penalized analysis) 방법 중 하나인 엘라스틱 넷(elastic net)을 사용하여 감정사전을 구축하고 관광지를 대표하는 단어와 해당 감정점수를 연결해주어 관광지 유형 별 만족도 요인 점수를 계산하였다. 본 연구에서 제안하는 방법론을 통해 외국인들이 중요시하는 요인들을 파악하고, 해당 요인에 자원을 집중하여 더 많은 관광수입을 창출할 것으로 기대한다.

본 연구의 구성은 다음과 같다. 제 2장에서는 관광 만족도 요인에 대한 기준의 관련 연구들이 어떻게 이루어져 왔는지 간략히 소개한 후 제 3장에서는 연구에서 사용할 방법론과 제안하는 방법론을 소개한다. 제 4장에서는 제안하는 방법론의 실험 방법에 대해 서술한 후 제 5장에서 그 실험 결과를 제시한다. 마지막으로 제 6장에서 본 연구의 결론과 시사점을 서술한다.

2. 선행연구

2.1 관광객의 만족도 연구와 설문지법의 한계

설문지법에 근거하여 외국 관광객의 만족도를 조사한 연구들은 다음과 같다. Gargano and Grasso(2016)는 이탈리아 시칠리아 보르고(Sicilian Borghi)를 방문하는 관광객을 대상으로 만족도에 대한 설문조사를 실시하였다. 음식, 청결함, 시설, 친절함, 가격 등에 대해 5점 척도의 설문조사를 실시하여 정렬된 로짓 모델(ordered logit model)로 분석한 결과 현지 음식, 여행비용, 역사적 장소가 관광객의 만족도에 크게 기여하였다. Hui *et al*.(2007)은 싱가폴을 방문한 관광객을 대륙별로 나누어 만족도를 조사하였다. 숙박, 음식, 가격, 기후 등에 대해 만족도를 조사한 후 호텔링의 T 제곱 분석(Hotelling's T-squared statistics), 대응표본 T-검정(paired T-test), 단순 회귀(simple regression) 등을 사용하여 분석하였다. 그 결과 북미 지역의 관광객은 숙박과 음식을, 아시아인과 유럽인들은 명소, 오세아니아의 관광객은 문화를 가장 중요한 요인으로 평가하였다.

Kim(2007)은 서울에 장기체류하고 있는 외국인 200명을 대상으로 관광자원과 관광 서비스 만족도를 조사하였다. 사전에 연구자가 만족도 요인으로서 제시한 쇼핑, 여행지 접근성, 여행 안전도, 교통 등에 대해 5점 척도의 설문조사를 실시했다. 이후, t-test와 일원분산분석을 실시한 결과 쇼핑에 대한 만족도가 제일 높고 교통에 대한 만족도가 가장 낮았다. Jiang(2015)은 관광을 목적으로 한국을 방문한 중국 여성 관광객 125명을 대상으로 교통, 숙박, 쇼핑, 음식점 서비스, 여행지 환경 등의 만족도에 대해 5점 척도의 설문조사를 실시하였다. 이에 대해 기술 통계 분석, 상관관계 분석, 다중회귀분석을 실시한 결과 숙박, 음식 서비스, 입국 절차 등의 만족도가 높은 것으로 나타났다.

외국관광객에 대한 국가적 차원의 조사로는 한국문화관광연구원이 발간한 '외래관광객 실태조사'가 있다. 이 조사는 1만 명이 넘는 외국인을 대상으로 전국적인 범위의 여행행태를 해마다 꾸준히 실시하고 있기 때문에 신뢰성이 높다. 하지만 이와 같은 선행연구 방법들은 각 관광지의 세부적인 평가를 얻을 수 없고 설문조사의 문항을 구성하는데 있어서 연구자의 편향이 더해질 수 있다는 문제점이 있을 뿐더러 설문조사를 위해 제한된 수의 표본을 사용한다는 단점이 있다(Jiang, 2015).

2.2 리뷰 데이터를 통한 감정분석

감정분석(sentiment analysis)이란 어떤 대상에 대한 화자의 긍·부정적 태도 등을 추출해서 제시하는 텍스트 처리 기술이다(Pang and Lee, 2008). 분석 방법에는 크게 어휘 기반(lexicon-based)과 모델 기반(model-based) 방식으로 나뉜다.

먼저, 어휘 기반의 분석 방법은 사전에 구축한 범용적인 감정 사전을 이용하는 방식이다(Taboada *et al*., 2011). Hu and Liu (2004)는 문서 내에 등장한 긍·부정 단어의 개수를 세는 방법을, Liu(2012)는 사전기반, 문서집합 기반으로 단어의 극성(polarity)을 판별하는 방법을 제시하였다. 모델 기반 방식은 기계학습으로 자동 구축한 범용 감정사전을 활용한다(Pang *et al*., 2002). 특히, 모델 기반 방식의 감정 분석법 중 벌점화 회귀분석 방법인 엘라스틱넷(elastic net)을 활용한 방식은 개별 어휘의 감정점수로 연속적인 값을 할당할 수 있고 주요 변수를 선택할 수 있으며 통계적으로 유의미하다는 장점이 있다. Kim *et al*.(2015)은 엘라스틱넷(elastic net)으로 산출한 각 단어의 회귀계수를

해당 단어의 감정점수로 이용하여 감정사전을 구축한 후 극성 판별에 적용하였다.

그러나 같은 어휘라도 특정 분야에 따라 현저히 다른 의미로 쓰일 수 있는 점을 고려할 때 도메인에 관계없이 일률적으로 적용하는 범용사전의 정확도는 상대적으로 높지 않다. 관광객들의 관심 영역이 쇼핑, 먹거리, 정보 등처럼 제각기 다르다면 동일한 관광명소를 같은 어휘로 평가할 지라도 코멘트별로 다른 감정사전을 적용해야 한다.

본 연구는 이를 해결하기 위해 관광객들의 세부 만족도 요인별로 다른 감정 점수를 산출하는 방법론을 제시하였다. 외래 관광객의 한국 방문 장소가 대부분 서울이라는 점(Ministry of Culture, Sports, and Tourism, 2016)을 고려하여 Tripadvisor에 등재된 서울 명소에 대한 평점과 리뷰데이터를 수집하였다. 이를 토대로 엘라스틱넷을 이용하여 감정사전을 구축하였다. 동일 어휘일지라도 세부 평가요인에 따라 다른 감정 점수를 가질 수 있는 점을 고려하기 위해 LDA를 활용하였다.

3. 방법론

3.1 벌점화 회귀분석(Penalized Regression)

회귀 모델의 성능을 평가하는 지표는 다양하지만 새로운 데이터에 대한 예측력과 모델의 해석력이 일반적으로 중요한 기준이다(Zou and Hastie, 2005). 그러나 데이터의 관측치 수에 비해 변수의 수가 많아지면 변수들 사이의 강한 상관관계로 인해 다중공선성이 존재하거나 과적합이 발생하여 모델을 해석하기 어렵고 예측력이 떨어지는 문제가 있다. 이를 극복하기 위해 벌점화 회귀분석은 회귀식의 목적함수에 회귀계수에 대한 제약을 부여한다. 본 절에서는 단순회귀분석의 목적 식에 해당 방법론을 적용한 식 (1)을 토대로 본 연구에서 사용한 방법론을 설명한다.

$$\sum_{i=1}^{n}(Y_i-\beta_0-\sum_{j=1}^{P}\beta_i x_{ij})^2+\lambda_1\sum_{j=1}^{P}|\beta_j|+\lambda_2\sum_{j=1}^{P}|\beta_j|^2 \qquad (1)$$

벌점화 회귀 분석 중 엘라스틱 넷은 식 (1)의 λ_1과 λ_2의 합이 1이라는 제약조건 하에서 각 λ값을 0과 1사이로 조절함으로써 중요한 변수를 선택할 수 있고 다중공선성 문제를 해결할 수 있다. 본 연구에서는 엘라스틱넷을 적용하여 산출한 각 회귀계수를 개별 어휘의 감정 점수로 사용하였다. 이렇게 얻은 감정점수는 연속적인 값을 지니는데, 연속형 값은 극성을 이분법적으로 나눌 수 없는 자연언어의 특성을 상대적으로 잘 반영한다는 장점이 있다.

3.2 잠재 디리클레 할당(Latent dirichlet allocation; LDA)

LDA는 <Figure 2>처럼 문서가 주어졌을 때, 각 문서에 어떤 주제들이 어느 정도 분포해 있는지, 문서를 구성하는 단어들은 어떤 주제에 할당되는지를 나타내는 확률 모형이다. 아래 식 (2)에서 나타낸 바와 같이 문서를 구성하고 있는 주제, 주제의 비중, 주제에 할당된 단어 등을 추론하는 것이 LDA의 목적이다(Blei, 2003).

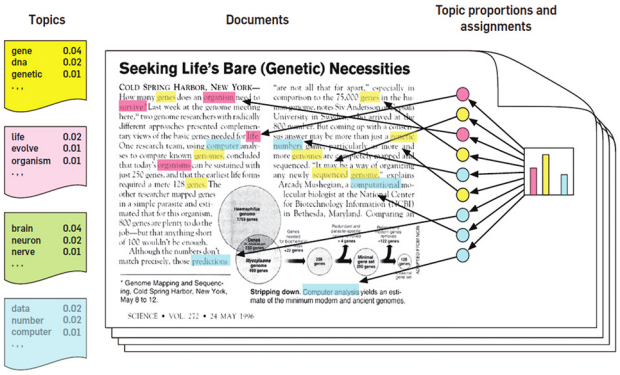

Figure 2. The Intuitions Behind Latent Dirichlet Allocation(Blei, 2012)

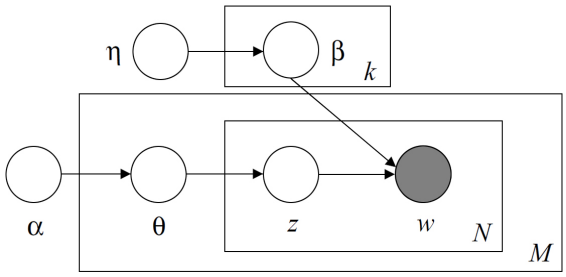

Figure 3. Graphical Model Representation of LDA(Blei, 2003)

$$p(\beta_{1:K},\theta_{1:M},z_{1:M},w_{1:M})=\prod_{i=1}^{K}p(\beta_i|\eta)\prod_{d=1}^{M}p(\theta_d|\alpha)\left(\prod_{n=1}^{N}p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:k},z_{d,n})\right) \qquad (2)$$

α : Dirichlet parameter, θ : 문서 당주제의 비중,
$z_{d,n}$: 단어 당주제의 할당, β_k : 주제별 특정단어가 생성될 확률,
η : 주제 hyperparameter, K: 주제의 개수,
M: 문서의 개수, N: 단어의 개수

<Figure 3>은 LDA의 문서 생성과정을 나타낸다. Dirichlet 파라미터 α가 주어졌을 때 문서의 개수 M개만큼 각 주제가 문서에 얼마큼 분포했는지를 파악하고 각 문서마다 문서를 구성하는 단어가 어느 주제에 해당되는지 알아본다. 또한, 주제에 대한 하이퍼 파라미터(hyper parameter) η가 주어졌을 때 각 주제마다 주제를 구성하는 단어의 확률 분포를 파악한 후 모두 종합하여 문서의 생성 확률을 구한다. 이러한 과정을 통해 LDA의 결과로 문서마다 각 주제가 얼마나 분포해 있는지 파악할 수 있다. 또한, 각 문서 내의 단어가 어느 주제에 해당되는지와 주제 당 단어의 확률 분포를 확인할 수 있다.

(1) LDA 기반의 가중 감정평가 방법론

감정어휘는 같은 단어라도 서로 다른 분야를 평가할 때 다른 의미·강도로 전달될 때가 많다. 따라서 본 연구에서는 LDA를 통해 얻을 수 있는 주제의 단어 분포를 유형 별 관광지의 만족도 요인의 점수를 계산할 때 활용하는 가중치로 활용하고, 회귀분석으로 계산한 단어의 감정점수와 문서 단어 행렬(document term matrix; DTM)을 조합하여 주제 별 감정점수를 산출하였다. 이를 도표로 나타내면 <Figure 4(a)>와 같다. 또한 <Figure 4(b)>에서는 리뷰의 전체적인 점수를 계산할 때 LDA 결과물인 문서의 주제 분포를 가중평균으로 반영하여 전체 점수를 계산한 것을 보여준다.

예를 들어 분식집과 레스토랑에 대해 평가할 때 '저렴하다'라는 단어를 맛과 가격이라는 평가요인에 사용할 수 있다. 일반적으로 '맛이 저렴하다'라는 문장보다는 '가격이 저렴하다'라는 문장이 실제로 더 많이 사용되고 자연스럽다. 따라서 '저렴하다'가 가격을 평가하는 데 쓰일 때 맛보다 높은 가중치를 부여한다. 또한 '저렴하다'가 레스토랑보다는 분식집을 평가할 때 사용될 가능성이 높다고 한다면 가격이라는 요인이 분식집 평가에 상대적으로 중요한 요소라는 점을 알 수 있게 된다. 본 연구에서 제안하는 방법론은 이처럼 주제의 단어 분포, 문서의 주제 분포를 모두 고려한다.

4. 실험설계

본 연구는 <Figure 5>에 따라 연구를 진행하였다. Tripadvisor에서 댓글과 평점을 수집하였고 댓글에 대해 전처리를 실시하였다. 이후 댓글로 문서-단어 행렬(document term matrix; DTM)을 구축하고 이를 이용하여 감정 사전을 만들고 세부 만족도 요인을 추출하였다. 최종적으로 감정 사전, 세부 만족도 요인 두 가지를 이용하여 각 명소의 평가 요인에 대한 점수를 산출하였다.

4.1 데이터 수집 및 전처리

본 연구에서는 Tripadvisor 사이트에서 서울에 있는 103개 여행지에 대한 리뷰로 1점에서 5점으로 기록된 평점을 21,620개를 수집하였다. 수집된 데이터는 리뷰를 작성한 외국인의 아이디, 위치, 날짜, 리뷰의 제목과 내용, 그리고 평점의 6개의 변수로 구성되어 있다. 수집된 댓글에 대해 형태소 분석(part-of-speech tagging; POS tagging)을 실시하여 품사가 명사, 형용사, 동사인 단어들만을 추출하였다. 이후 선택된 단어에 대표형 변환기법 중 하나인 Lemmatization을 실시하여 모든 댓글을 소문자로 변경하고 불필요한 구두점, 숫자, 공백을 제거하였다. 유의미한 분석을 위해 출현빈도가 10 미만이면서 문자열의 길이가 2개 미만인 단어는 제외하였다. 대표형으로 변환된 단어가 모인 문서집합을 BOW 방식으로 단어빈도(term frequency)가 행렬의 값이 되는 문서-단어 행렬을 구성하였고 이 문서-단어 행렬을 사전에 정의한 3가지 서울의 관광지 유형에 따라 분류하였다. 관광지의 유형과 관광지 유형 별 문서집합의 문서의 개수는 <Table 1>에 요약되어 있다.

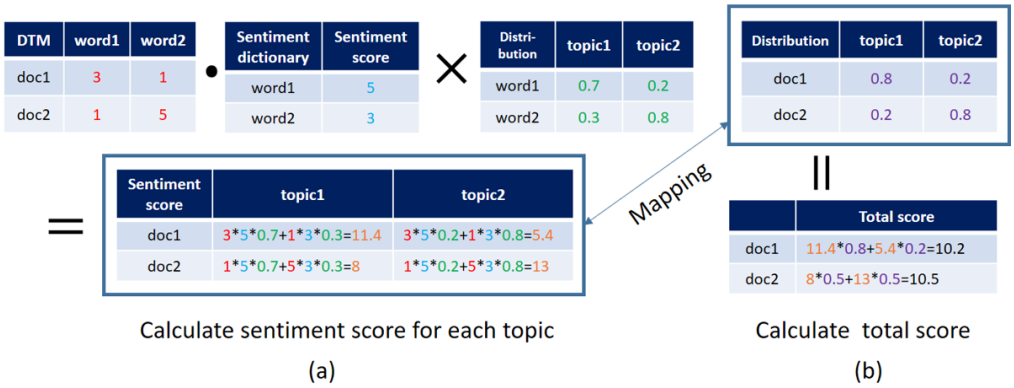

Figure 4. Calculating Total Score and Sentiment Score for Each Topic

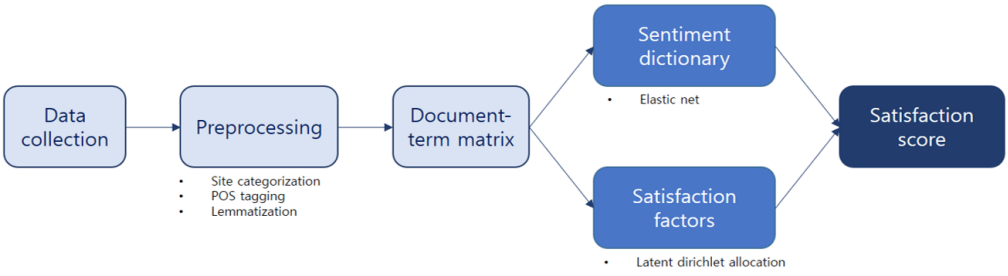

Figure 5. Framework of the Proposed Method

Table 1. Types of Tourist Attractions and the Number of Documents for Each Type

Type of tourist attractions	The number of documents	Place
Attractions & landmarks	14,443	N Seoul tower, Gyeongbokgung palace, Hangang park, Insadong etc.
Shopping & markets	3,589	Myeongdong shopping street, COEX mall, Times Square mall, Gwangjang market etc.
Museum & gallery	3,588	The war memorial of Korea, Trickeye museum, National museum of Korea etc.

4.2 유형 별 관광지의 감정사전 구축 및 만족도 요인 도출

제 4.1절에서 구축한 문서-단어 행렬 가운데 70%를 학습 데이터(training data)로, 나머지 30%를 검증데이터(validation data)로 구성하였다. 아래 식 (3)처럼 정의되는 평균절대오차(mean absolute error; MAE)를 지표로 활용하여 학습데이터에 10-fold cross validation 기법을 적용하여 엘라스틱넷을 학습하였다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{3}$$

엘라스틱넷으로 학습한 모델 가운데 검증데이터의 평점 예측성능이 가장 좋은 모델의 회귀계수를 개별 어휘의 감정점수로 정의하였다. 이를 바탕으로 리뷰의 긍·부정을 평가할 때는 리뷰에 등장하는 긍정과 부정 어휘에 해당 어휘의 회귀계수를 이용한 가중평균을 최종 감성 점수로 사용하였다. 구축된 감정사전의 성능을 평가하기 위해서 단순 정확도(accuracy), 재현율(recall), 정밀도(precision), 균형 정확도(balanced correction rate)를 활용하였다. 5점 만점의 리뷰 스코어의 실제 평균은 스코어의 중간 값인 3보다 크게 나타나기 때문에 본 연구에서는 4점을 기준으로 긍정과 부정 리뷰를 구분하였다. 실제 Tripadvisor 리뷰 결과와 본 연구에서 구축한 감정 사전을 바탕으로 추정된 긍·부정 결과를 바탕으로 <Table 2>와 같은 혼동 행렬(confusion matrix)이 구성될 수 있다. 이를 바탕으로 식 (4)~식 (7)을 이용하여 단순정확도, 재현율, 정밀도, 균형 정확도를 계산할 수 있다.

Table 2. Confusion Matrix Constructed Based on The Actual Review Scores and The Predicted Polarity by the Proposed Sentiment Dictionary

Confusion matrix		Predicted polarity by proposed sentiment dictionary	
		Positive	Negative
Actual review scores	Positive	a	b
	Negative	c	d

$$accuracy = \frac{(a+d)}{(a+b+c+d)} \tag{4}$$

$$recall = \frac{a}{(a+b)} \tag{5}$$

$$precision = \frac{a}{(a+c)} \tag{6}$$

$$balanced\ correction\ rate = \sqrt{\frac{a}{a+b}\cdot\frac{d}{c+d}} \tag{7}$$

유형 별 관광지의 문서-단어 행렬에 LDA를 적용하여 관광지 만족도 요인을 도출하였다. 본 연구에서는 LDA의 파라미터로 최대 5,000번의 반복 횟수(iteration), 5개의 토픽 수, 그리고 문서의 토픽분포를 결정하는 파라미터인 α를 0.1로 정하고 진행하였다. 이후 각 토픽별 상위 빈출 어휘들을 고려하여 적절한 이름을 부여한 후 이를 만족도 요인으로 정의하였다. 이 과정에서 단순히 토픽의 단어분포 뿐만 아니라 해당 토픽의 비중이 높게 나타나는 리뷰들을 함께 고려하여 보다 적절한 만족도 요인이 도출될 수 있도록 하였다.

4.3 유형 별 관광지 만족도 요인 점수 계산

LDA로 얻어낸 유형 별 관광지 만족도 요인에 <Figure 4>에서 제안한 방법론을 적용하여 관광지의 전체 만족도 점수를 계산하였다. 만족도 점수를 5점 척도로 변환하기 위해 식 (8)과 같은 변환 방법을 이용하였다.

$$z_i = \left(\frac{x_i - \min(x)}{\max(x) - \min(x)}\right) \times 4 + 1 \tag{8}$$

z_i = 5점 척도로 변환한 후 관광지의 전체 만족도 점수
x_i = 변환 전 관광지의 만족도 점수)

5. 실험 결과

5.1 유형 별 관광지의 감정사전

감정 사전을 구축한 결과 유형 별 관광지에 대한 긍정 및 부정적인 단어를 <Figure 6>과 같이 워드 클라우드로 나타낼 수 있다. 색깔이 진하고 글자의 크기가 클수록 긍정 혹은 부정의 강도가 강한 단어이다. 예컨대 명소·랜드마크 유형의 관광지는 기념품과 먹거리에 대한 외국인의 관심도가 높지만 관광지가 청결하지 않거나 지루할 경우 크게 실망하는 것을 확인할 수 있다. 쇼핑·시장에서도 해당 유형의 관광지에 대해 전반적으

로 만족하지만 점원의 태도가 무례하거나 가격이 비쌀 경우 부정적으로 평가하였다. 마지막으로 박물관 & 전시관 유형의 관광지에서는 문화, 유산과 같은 전시된 내용에 대해서는 긍정적으로 평가하지만 비싼 가격에 대해서는 실망하는 경향을 보였다.

감정사전 구축에 사용한 어휘의 수는 <Table 3>에 명시되어 있다. 전체 분석대상 어휘는 명소・랜드마크, 쇼핑・시장, 박물관・전시관 유형별로 각각 2,826, 1,319, 1,329개다. 이중 엘라스틱넷 회귀계수가 0이 아닌 유의미한 단어는 유형별로 각각 747, 249, 242개로 나타났다.

Table 3. The Number of Words of Sentiment Dictionary for Each Type of Tourist Attraction

	Attractions and landmarks	Shopping and markets	Museum and gallery
Total number of words	2,826	1,319	1,329
The number of words of sentiment dictionary	747	249	242

Table 4. The Performance of Sentiment Dictionary for Each Type of Tourist Attraction

	Attractions and landmarks	Shopping and markets	Museum and gallery
Accuracy	0.8246	0.7465	0.8717
Recall	0.8911	0.8028	0.9451
Precision	0.8999	0.8686	0.9149
Balanced correction rate	0.6444	0.6506	0.4561

각 관광지 유형별 감정사전의 성능은 <Table 4>에 요약되어 있다. 평가를 위해 단순 정확도, 재현율, 정밀도, 균형 정확도의 네 가지 지표를 사용한 결과 모든 유형의 감정사전에서 균형 정확도를 제외한 다른 여러 가지 지표는 감정사전으로서 타당한 성능을 확보했으나 균형 정확도의 성능은 상대적으로 높은 수준은 아닌 것으로 나타났다. 그 이유는 Tripadvisor에서 수집한 평점들이 <Figure 7>에 나타난 바와 같이 4점 또는 5점의 점수대에 많이 분포해 있는 불균형 데이터(unbalanced data)이기 때문이다.

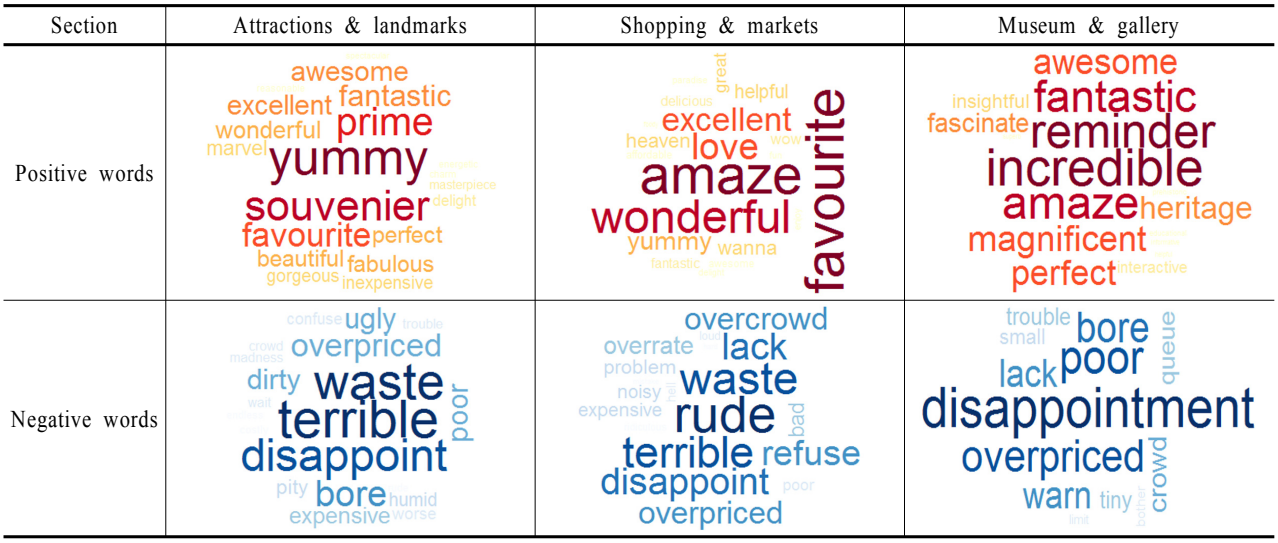

Figure 6. Word Clouds of Sentiment Dictionary for Each Type of Tourist Attraction

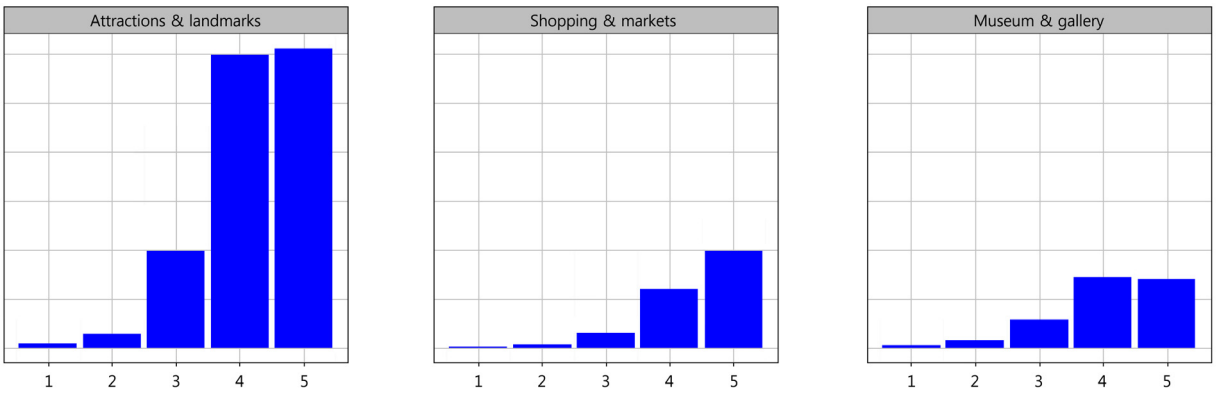

Figure 7. Histogram of Scores for Each Type of Tour Attractions

5.2 유형 별 관광지 만족도 요인

LDA를 적용하여 산출한 각 관광지 유형 별 만족도 요인과 요인들을 대표하는 단어는 다음 <Table 5>~<Table 7>과 같다.

Table 5. Satisfaction Factors and Words Representing Attractions and Landmarks

Souvenir shopping	Intricacy while climbing	Atmosphere	Traditional experience	Architecture
souvenir	ride	relax	tour	build
gift	hike	night	garden	architecture
sell	climb	lovely	costume	modern
price	crowd	peaceful	guard	structure
cheap	line	festival	ceremony	complex
market	queue	river	royal	wall
craft	long	romantic	historical	style
item	bus	wonderful	informative	home
unique	top	stroll	korea	room
local	hill	time	architecture	unique

Table 6. Satisfaction Factors and Words Representing Shopping and Market

Shopping at night	Clothes shopping	Accessibility	Cosmetic shopping	Food
night	shoe	find	cosmetic	seafood
evening	bag	locate	brand	fresh
late	accessory	metro	beauty	cook
pm	clothe	map	skin	taste
cosmetic	sock	convenient	skincare	delicious
shopper	bargain	open	mask	crab
open	shirt	hour	face	fry
clothe	design	end	sample	sashimi
brand	brand	exit	brand	octopus
vendor	cheap	leave	paradise	pancake

Table 7. Satisfaction Factors and Words Representing Museum and Gallery

War related information	Fun	Accessibility	Instructive	Art
tank	funny	free	education	collection
military	creative	walk	historical	display
plane	photo	time	informative	exhibition
ship	ice	subway	learn	gallery
vehicle	kid	city	guide	sculpture
conflict	friend	locate	culture	artist
soldier	trick	admission	understand	treasure
fight	eye	fee	bore	piece
weapon	illusion	enter	insight	design
monument	pose	close	explain	artifact

명소・랜드마크 유형 관광지들의 경우 해당 장소의 역사적 전통, 경치 및 분위기에 대한 언급이 많았으며 기념품 및 쇼핑에 대한 평가도 큰 비중을 차지하고 있는 것을 알 수 있다. 또한 다른 유형에서는 볼 수 없었던 장소의 복잡함에 대한 리뷰가 상당수 존재한다는 것을 확인할 수 있다. 쇼핑・시장 유형의 관광지들에 대해서는 목적에 맞게 합리적 쇼핑과 먹거리에 대한 평가가 주를 이루는 것을 알 수 있었으며, 특이하게 쇼핑 항목 중에서도 화장품과 관련된 리뷰가 많이 존재한다는 것을 확인할 수 있다. 또한, 실제 거래를 위해 판매자와 소비자가 의사소통을 해야 하는 쇼핑의 특성을 반영한 리뷰도 상당수 존재하는 것을 알 수 있다. 마지막으로 박물관・전시관 유형의 관광지들에 대해서는 목적에 맞게 역사와 문화, 전시물, 유익한 정보, 재미 등을 언급한 리뷰가 많았으며 특이사항으로는 교육적 목적으로 자녀를 동반하는 것에 대한 언급이 많았다는 것이다.

5.3 유형 별 관광지 만족도 요인 점수

제 4.3절에서 설명한 방식을 통하여 관광지별 전체 만족도 점수를 Tripadvisor에서 수집한 평점과 비교해 MAE로 측정한 성능은 <Table 8>과 같다. 명소・랜드마크에 대한 절대 오차가 1.18점으로 가장 낮았으며 쇼핑・시장 관광지들에 대한 오차는 1.69점으로 다소 높은 것으로 나타났다. 또한, 각 관광지 유형을 대표하는 장소들의 세부 만족도 점수와 전체 만족도 점수에 대해 상자그림(box plot)을 그린 결과는 다음 <Figure 8>~<Figure 10>과 같다.

Table 8. MAE for Each Type of Tourist Attractions

	Attractions and landmarks	Shopping and markets	Museum and gallery
MAE	1.1783	1.6914	1.3746

먼저, 관광지 유형별로 관광객이 중시하는 만족도 요인별로 점수의 분포가 상당한 차이가 있는 것을 확인할 수 있다. 명소・랜드마크의 경우 63빌딩과 남산 케이블카 모두 기념품 쇼핑, 복잡함, 분위기, 전통, 경치의 5가지 만족도 요인 중 기념품 쇼핑이 가장 낮은 만족도 점수를 받았다. 이외의 평가요인에 대해서는 만족도 점수의 분포가 비슷함을 확인할 수 있었다. 쇼핑・시장 유형의 대표적 장소인 이화여대 앞 쇼핑거리와 코엑스몰 모두 산출된 주요 만족도 요인 중에서 합리적인 쇼핑에 대한 점수는 높게 부여되었으나 저녁 여행 관점에는 상대적으로 낮은 점수가 부여된 것을 알 수 있다. 마지막으로 박물관 & 전시관 유형의 대표적 장소인 서대문 형무소와 대한민국 역사박물관의 경우 관광객들이 유익한 정보, 역사・문화, 전시 관점에서는 높은 평가를 내렸으나 자녀 동반 과점에서는 유보적인 입장을 취하며 재미 관점에서는 매우 평가가 낮게 분포하는 것을 확인할 수 있다.

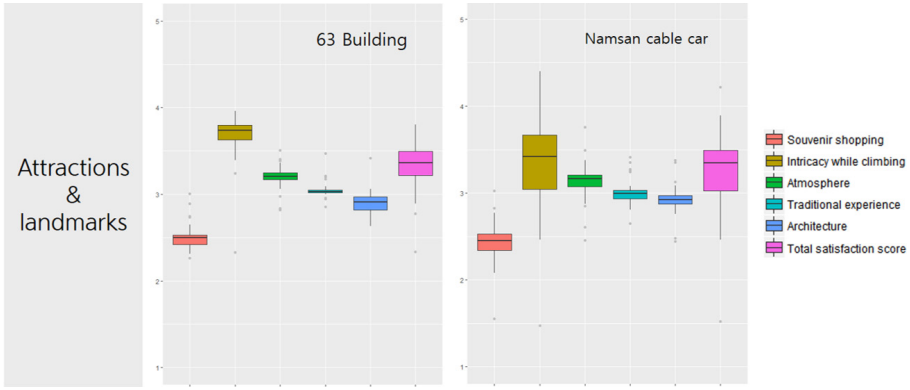

Figure 8. Box Plots of Each Satisfaction Score and Total Score for Representing Attractions and Landmarks

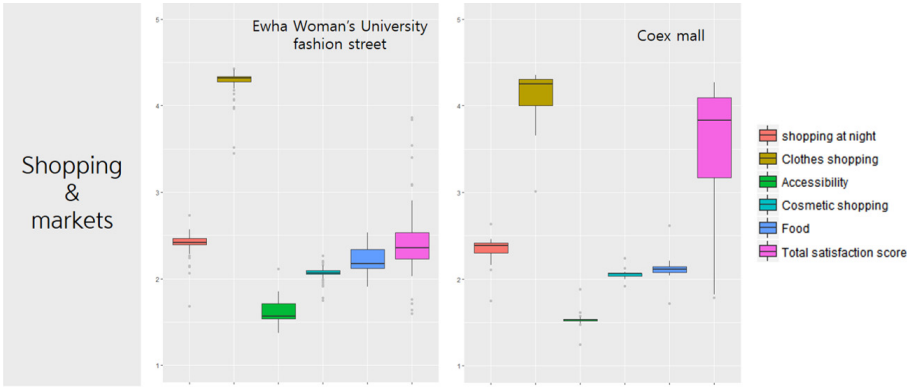

Figure 9. Box Plots of Each Satisfaction Score and Total Score for Representing Shopping and Markets

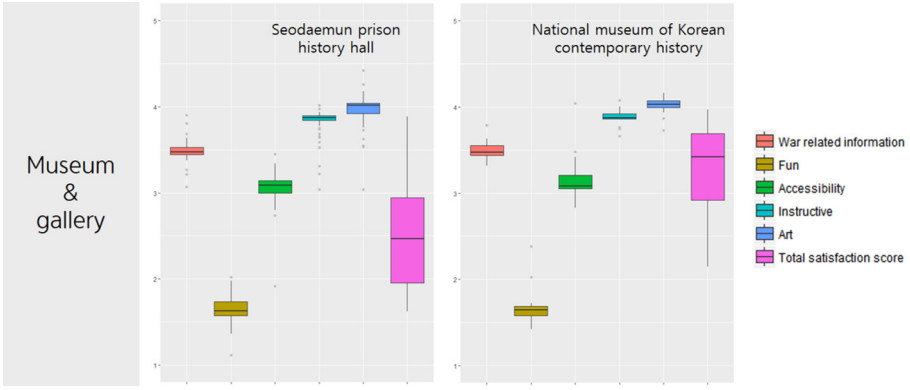

Figure 10. Box Plots of Each Satisfaction Score and Total Score for Rrepresenting Museum and Gallery

각 관광지 유형을 대표하는 장소가 세부 만족도 요인의 점수가 비슷한 양상을 보이기는 하지만 같은 유형에 속한다고 해서 모든 결과가 같은 것은 아니다. 예컨대 명소・랜드마크의 경우 63빌딩과 남산 케이블카가 같은 전체 만족도 점수를 갖더라도 남산 케이블카의 분산이 더 큰 것을 보아 남산 케이블카에 대한 관광객의 평가가 63빌딩보다 극명하게 갈리는 것을 확인할 수 있다. 복잡함이라는 세부 만족도 측면에서도 두 장소가 평균과 분산의 측면에서 확연한 차이가 난다. 63빌딩은 3점 후반대의 평균 점수를 가지며 그 분산은 남산 케이블카보다 작지만 남산 케이블카는 3점 초반대의 점수이면서 분산이 크다. 따라서 외국인 관광객이 복잡함 측면에서는 남산 케이블카를 63빌딩보다 다양하게 평가한다는 결론을 내릴 수 있다.

나머지 두 유형인 쇼핑・시장, 박물관・전시관에서는 각 유형을 대표하는 장소를 세부 요인을 기준으로 비교했을 때 평균과 분산이 큰 차이를 보이진 않지만 전체 만족도 측면에서는 평균과 분산이 확연하게 다름을 확인하였다. 이화여대 앞 쇼핑거리와 코엑스 몰을 비교했을 때, 의사소통, 합리적 쇼핑 등의 각 세부 만족도 요인 점수의 평균과 분산은 크게 차이가 나지 않지만 전체 만족도 점수는 큰 차이를 보이며 그 분산 또한 크다. 마

찬가지로 서대문 형무소와 대한민국 역사박물관 또한, 유익한 정보, 재미 등의 세부 만족도 점수의 평균과 분산은 거의 일치한 반면 전체 만족도 점수는 2점대와 3점대로 확연히 다르다. 이러한 결과로 볼 때, 관광지 유형이나 총 평점이 동일하더라도 세부 만족도 요인의 평균, 분산 등 그 분포가 다르게 나타나는 경우가 많으며 해당 관광지에 대한 단점을 보완하거나 새로운 관광 전략을 수립할 때 이러한 정보가 함께 고려가 되면 보다 효과적인 개선을 이루어 낼 수 있을 것으로 기대한다.

6. 결 론

본 연구는 국내 관광지를 유형별로 구분한 후 Tripadvisor의 리뷰, 평점을 토대로 엘라스틱넷 방법론을 통해 감정사전을 구축하였다. 또한, 토픽 모델링의 방법 중 하나인 LDA를 이용하여 유형 별 관광지의 만족도 요인을 도출하였고 감정사전과 만족도 요인으로 각 유형 별 관광지 만족도 요인 점수를 산출하는 방법론을 제시하였다. 그 결과 3가지 관광지 유형별로 세부 만족도 요인 점수의 분포가 비슷한 양상을 보이지만 각 유형의 장소들을 세밀히 관찰했을 때, 그 장소만의 특징 또한 갖고 있음을 확인하였다.

본 연구의 방법론은 해외 관광객들의 한국 여행 만족도 요인을 조사하기 위한 효과적인 수단으로서 각 관광지의 장단점을 파악하여 해당 관광지의 경쟁력을 높이기 위한 구체적인 투자 방안을 설립하는 근거로 사용될 수 있다. 본 연구 방법론과 컨텐츠 기반의 협업 필터링, 또는 사용자 기반의 협업필터링을 함께 사용한다면 고객 개인에게 맞춤화된 추천시스템을 구축할 수 있을 것으로 기대한다.

하지만 본 연구는 연구의 대상으로 선정한 관광지가 서울로 제한되어 있고 연구자가 사전에 관광지를 세 가지 유형으로 나누었다는 한계를 갖는다. 따라서 추후에 연구 대상을 지방까지 확대하고 군집화 등의 방법론을 이용하여 정량적으로 관광지 유형의 수를 결정한다면 좀 더 의미 있는 결과가 도출될 것으로 기대한다.

참고문헌

Archer, B. and Fletcher, J. (1996), The economic impact of tourism in the Seychelles, *Annals of tourism research*, **23**(1), 32-47.

Blei, D. M. (2012), Probabilistic topic models, *Communications of the ACM*, **55**(4), 77-84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), Latent dirichlet allocation, *Journal of machine Learning research*, **3**(Jan), 993-1022.

Hu, M. and Liu, B. (2004), Mining and summarizing customer reviews, *In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA.

Jee, B.-G., Lee, G.-H., and Kim, T.-G. (1999), Economy impact of tourism industry in Korea-Input/output analysis, *The Journal of the Korea Contents Association*, **11**(12), 884-892.

Jiang, W. (2015), *A study on the satisfaction factors for foreign tourism services*, Master's Thesis, Kookmin University.

Joachims, T. (1999), Making large scale support vector machine learning practical, *Advances in kernel methods*, 169-184.

Kim, B. M. (2007), *A study on foreign residents' satisfaction with the tourism services*, Master's Thesis, Hanyang University.

Kim, J. H. (1995), *A study on reliability analysis of questionnaire items*, Master's Thesis, Jeonju University.

Kim, S. B., Kwon, S. J., and Kim, J. T. (2015), Building sentiment dictionary and polarity classification of blog review By using elastic net, *In Proceedings of the 2015 Winter Conference of Korean Institute of Information Scientists and Engineers*, Pyeongchang, Korea.

Liu, B. (2012), Sentiment analysis and opinion mining, *Synthesis lectures on human language technologies*, **5**(1), 1-167.

Ministry of Culture, Sports, and Tourism (2016), *2015 International Visitor Survey*, Ministry of Culture, Sports, and Tourism, Korea.

Pang, B. and Lee, L. (2008), Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, **2**(1-2), 1-135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002), Thumbs up? : sentiment classification using machine learning techniques, *In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA.

Ryu, K.-H. (2003), The method for estimating socioeconomic impacts of tourism industry in Korea, Research Report, Korean Culture and Tourism Institute.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011), Lexicon-based methods for sentiment analysis, *Computational linguistics*, **37**(2), 267-307.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **67**(2), 301-320.

Gargano, R. and Grasso, F. (2016), Analysis of the Factor that Affect Tourist Satisfaction : A Case Study on "The Most Beautiful Sicilian Borghi."

Hui, T. K., Wan, D., and Ho, A. (2007), Tourists' satisfaction, recommendation and revisiting Singapore, *Tourism management*, **28**(4), 965-975.

<Appendix> Box Plots of Each Satisfaction Factor Score and Total Score for Different Tourist Attractions

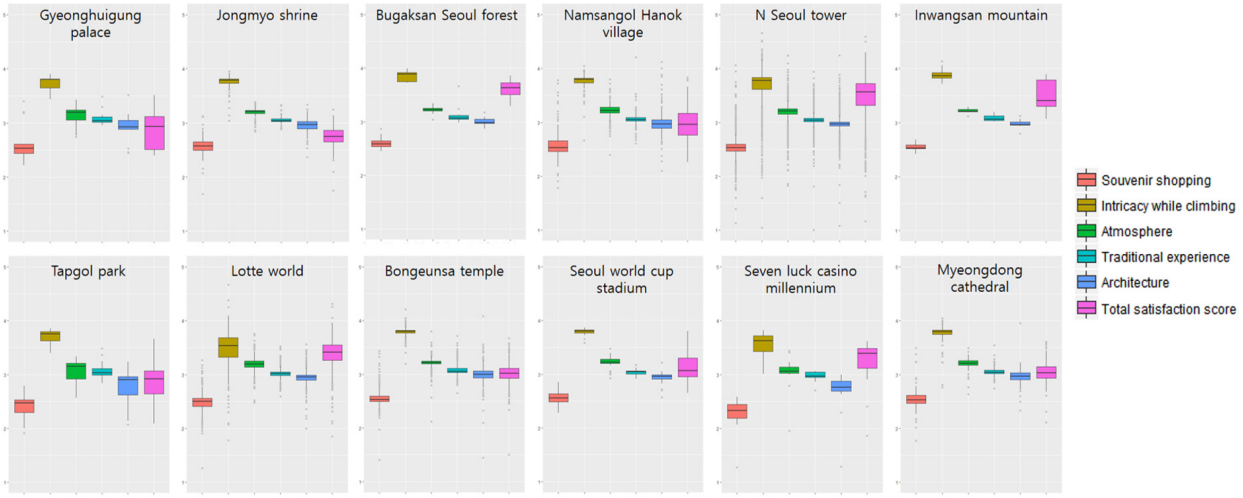

Figure A-1. Box plots of each satisfaction score and total score for attractions and landmarks

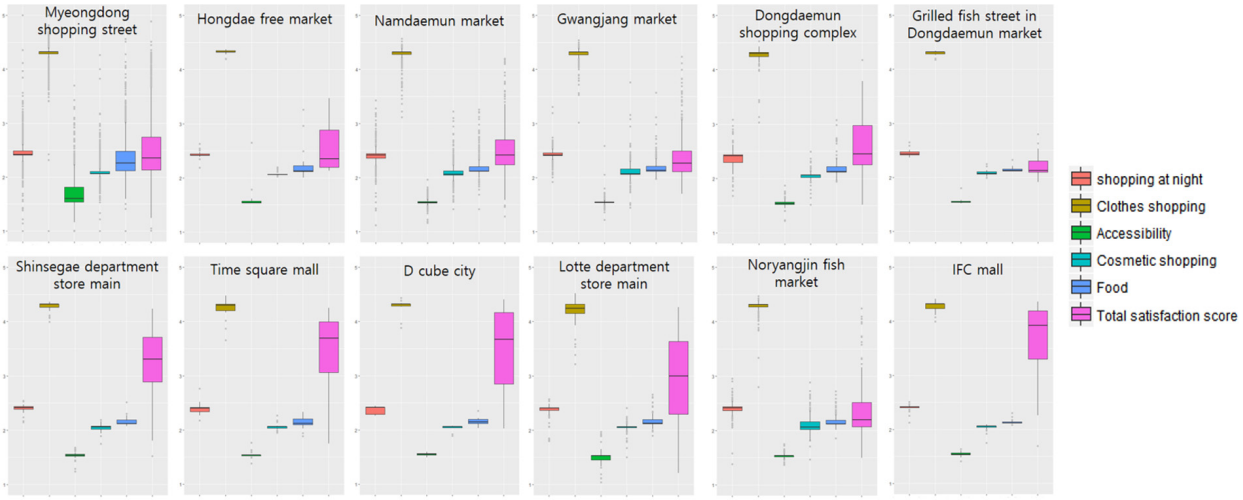

Figure A-2. Box Plots of Each Satisfaction Score and Total Score for Representing Shopping and Markets

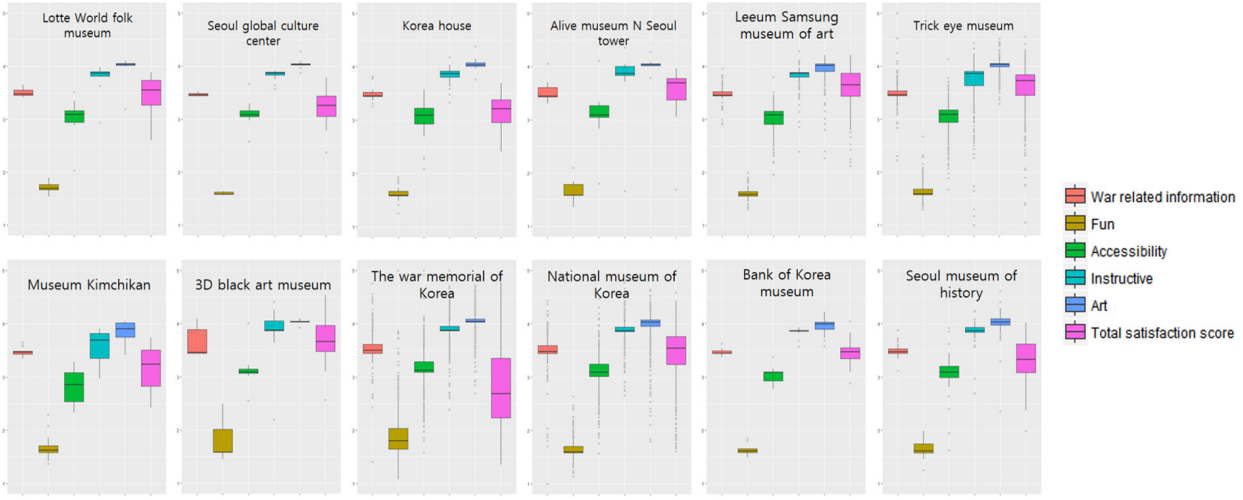

Figure A-3. Box Plots of Each Satisfaction Score and Total Score for Representing Museum and Gallery