온라인 리뷰를 활용한 관광지 키워드 추출 기법 - 토픽 모델링과 Markov Chain

김명선*‡ , 이강우*, 임지원*, 홍순구*†
* 스마트거버넌스 연구센터
† 동아대학교 경영정보학과
‡ 동아대학교 컴퓨터공학과
mxxsxxkim@gmail.com, kangwooster@gmail.com, ji1e@naver.com,
shong@dau.ac.kr

Keyword Extraction Technique for Attractions using Online Reviews - Topic Modeling and Markov Chain

Kim, MyeongSeon*‡ , Lee, KangWoo*, Lim, JiWon*, Hong, Soon-Goo*†
* Smart Governance Research Center
† Dept. of MIS, Dong-A University
‡ Dept. of Computer Engineering, Dong-A University

요 약

관광 분야에서 온라인 리뷰의 중요성이 커지고 있다. 온라인 리뷰의 텍스트 데이터는 파악이 어렵다. 이에 본 연구에서는 특정 관광지에 대한 온라인 리뷰 텍스트 데이터가 나타내는 전반적인 의견을 직관적으로 도출하는 방법에 대해 알아보고자, 토픽 모델링과 Markov Chain을 시행했다. '해운대'에 대한 온라인 리뷰를 수집한 후, LDA와 BTM을 활용하여 주제를 도출하고, Markov Chain을 시각화하여 키워드 간의 관계와 전체적인 평가 내용을 확인했다. 사용된 기법은 각자 특징적인 결과를 제시했기 때문에 다양한 기법을 상보적으로 이용하기를 제안하였다.

1. 서론

관광 분야에서 온라인 리뷰는 구매 의사결정 프로세스의 일부로 중요성이 계속 커짐에 따라 여행자에게 미치는 영향력이 커지고 있다[1]. 별점과 같은 수치 데이터는 직관적이지만, 리뷰와 같은 텍스트 데이터는 방대하고 개념적이므로 평가에 대한 전체적인 파악이 힘들다.

따라서 본 연구에서는 관광지에 대한 온라인 텍스트 리뷰가 나타내는 전반적인 의견을 직관적으로 도출하는 방법에 대해 알아보고자 한다. 이를 위해 특정 관광지에 대한 온라인 리뷰를 수집하여 분석했다. 온라인 리뷰에 숨겨진 토픽을 추출하기 위해 LDA, BTM을 활용한 토픽 모델링을 시행하고, 단어 간의 관계를 알아내기 위해 Markov Chain을 시각화했다. 본 연구를 통해 관광지에 대한 의견을 도출하고, 연구 결과에 기반하여 온라인 리뷰 활용방안을 제안하고자 한다.

2. 연구 계획

2.1 데이터 수집 및 전처리

관광지에 대한 의견을 분석하기 위해 부산의 관광지 중 하나인 '해운대'에 대한 리뷰 데이터를 트립어드바이저[2]에서 수집했다. 리뷰 데이터는 한국어로 작성된 해운대에 관한 내용으로, 2011년 09월부터 2021년 03월까지 548개를 수집했다.

자연어 처리를 위해 KoNLPy 패키지의 mecab 형태소 분석기를 이용하여 실질적 의미를 가지는 명사, 형용사, 동사를 추출했다[3]. 이때 '웨스틴조선호텔', '마린시티', '센텀시티' 등의 21개의 지명과 '가성비', '버스킹', '태닝' 등의 9개의 미등록 단어는 사용자 사전으로 정의하여 하나의 형태소로 분석했다. '부산', '하다', '되다', '같다' 등의 무의미한 320개의 단어는 불용어로 처리했다. 추가로 기존의 리뷰 단위 데이터에서 문장 부호를 기준으로 분리한 문장 단위 데이터를 생성하여 동일한 전처리 과정을 시행했다. 문장 단위 데이터는 1,665개이고, 형태소 분석기로 추출된 단어는 9,432개, 중복 제거 후 1,862개이다.

2.2 토픽 모델링

관광지 리뷰에서 주요 주제를 추출하기 위해 토픽 모델 LDA과 BTM을 수행했다. LDA는 토픽에 대한 단어분포를 바탕으로 주어진 문서의 토픽을 추정하는 모델이고[4], BTM은 두 단어 간의 결합확률 분포(joint probability distribution)에 기초하여 문서의 topic을 추정하는 모형이다[5]. 일반적으로, LDA는 길이가 긴 문서에 적합한 반면, BTM은 twitter와 같은 짧은 문서에 적합하다. 토픽 개수는 Coherence Score가 가장 적합한 경우를 채택했다. Coherence Score는 주제 일관성으로, 토픽 모델링의 평가 기준 중 하나이다. LDA는 Coherence Score가 높을수록 성능이 좋은 c_v로, BTM은 0에 가까울수록 성능이 좋은 umass를 Coherence Score 계산에 사용했다.

2.3 마르코브 연쇄 (Markov Chain)

Markov Chain은 문장을 순서적으로 나열된 단어들의 천이확률로 표현하고, 높은 천이확률을 가지는 단어순서쌍을 추출함으로써 전반적 주체를 추정할 수 있다[6]. 문장 단위 데이터에서 Bigram을 활용하여 특정 단어 뒤에 오는 단어를 정리하고, 3번 이상 등장하는 단어 쌍을 그래프로 표현하여 Markov Chain을 시각적으로 구현했다.

3. 연구 결과

3.1 LDA (Latent Dirichlet Allocation)

데이터를 리뷰 단위로 분석한 결과, 최적 토픽의 개수는 7개로 나타났으며, Coherence Score(c_v)는 0.597이다. topic1은 조용, 공연, 명소, 관광객, 힐링 등의 단어가 상위로 나타났으며, topic2는 야경, 바닷바람, 성수기, 매력, 시원 등이 키워드로 등장하였다. 각 토픽의 상위 단어는 키워드는 관광지 명소로써 해운대에 대한 전반적인 의견을 담고 있다<표 1 참조>.

<표 1> LDA 분석 결과

번호	상위 단어
topic1	조용/공연/명소/관광객/힐링/백사장/붐비다/한적/관광
topic2	야경/바닷바람/성수기/매력/시원/상쾌/지저분/휴식/굿
topic3	봄/재미있다/가을/낭만/즐비/축제/성수기/예쁘다/쓰레기
topic4	아이/광안리/날씨/낮/비싸다/아침/음식/여유/축제
topic5	가을/힘들다/해안가/길다/해돋이/갈매기/버스/한적/평일
topic6	갈매기/젊음/먹거리/명소/아침/괜찮다/주차/파도/시원
topic7	밤바다/야경/도시/파도/맥주/오랜만/친구/버스킹/공연

3.2 BTM (Biterm Topic Modeling)

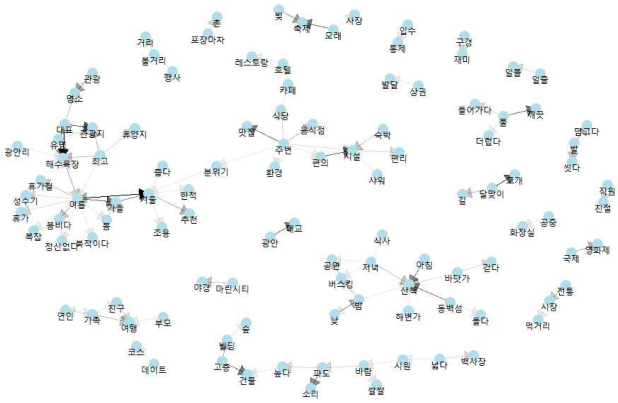
데이터를 문장 단위로 사용해 분석한 결과, 최적 토픽의 개수는 5개이며, Coherence Score(umass)는 -118이다. topic1은 센텀, 백화점, 벡스코, 신세계, 전통 등의 키워드가 등장하여 쇼핑 및 체험 관련 내용으로 추론할 수 있으며, topic2는 운동, 대회, 맞이, 부대시설, 한산 등의 단어가 상위 키워드로 나타났다<표 2 참조>.

<표 2> BTM 분석 결과

번호	상위 단어
topic1	센텀/백화점/벡스코/신세계/전통/스파/야외/쌀쌀/입구
topic2	운동/대회/맞이/부대시설/한산/부모/태양/비둘기/덥다
topic3	재미/상권/공원/막히다/서울/시민/상가/해안선/경험/쾌적
topic4	주차/바가지/다르다/스팟/전국/외지인/광활/발렛/더럽다
topic5	여름/정신없다/먹을거리/길다/피서객/활기/젊음/거닐다/유명

3.3 Markov Chain

문장 단위 데이터에서 Bigram을 이용하여 Markov Chain을 그래프로 나타낸 결과이다<그림 1 참고>. 간선으로 연결된 집단을 보면 관광, 계절, 주변 시설, 산책, 여행객 단위 등을 중심으로 한눈에 관광지에 대한 의견을 파악할 수 있다. 중심 단어들은 방향성을 지닌 간선을 통해 연결되고, 연쇄적 단어쌍을 통해 해운대에 대한 주제를 추정할 수 있다.



(그림 1) Markov Chain 시각화 그래프

4. 결론

본 논문에서는 3가지 텍스트 마이닝 방법-LDA, BTM 그리고 Markov Chain-을 활용해서, 특정 관광지에 대한 리뷰 분석을 시도하였다. 토픽 모델링은 온라인 리뷰에 숨겨진 주제를 키워드 집합으로 제시하여 해당 관광지의 토픽에 집중한다. 반면에 Markov Chain은 단어 간의 순서적 연쇄관계를 표

현한 그래프를 제시하여 온라인 리뷰가 가지는 전체적인 대략적 내용을 파악하기 용이하다.

토픽 모델링에 사용된 LDA와 BTM은 서로 다른 결과를 제시했다. LDA는 '관광지 해운대'에 대한 전형적 의견을, BTM은 관광지 해운대의 '세부적' 주제를 도출했다. 관광지 리뷰 특성상 다양한 내용을 동시에 언급하기 때문에 리뷰 단위로 분석한 LDA의 경우 토픽이 명료하게 구분되지 않았지만, 단문에 유리한 BTM의 문장 단위 결과를 보면 LDA보다 상대적으로 특징적인 토픽을 도출했다는 점에서 BTM을 보완적으로 사용할 수 있다.

본 연구에서는 온라인 텍스트 리뷰에서 전반적인 의견을 추출하기 위해 여러 텍스트 마이닝 방법을 시행했다. 본 연구에서 활용한 기법은 각자 다양한 결과를 제시했다. 따라서 특정 관광지에 대한 의견을 요약적으로 파악하기 위해 다양한 기법을 상보적으로 이용하기를 제안한다

ACKNOWLEDGMENT

이 논문은 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018S1A3A2075240)

참고문헌

[1] O'Connor, P. (2008). User-Generated Content and Travel: A Case Study on Tripadvisor.Com. *Information and Communication Technologies in Tourism 2008*, pp. 47-58.

[2] TripAdvisor, http://www.tripadvisor.co.kr/

[3] Park, E. L. & Cho, S. (2014). KoNLPy : Korean natural language processing in Python. *Annual Conference on Human and Language Technology 26*, pp. 133-136.

[4] Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research 3*, pp. 993-1022.

[5] Yan, X., Guo, J., Lan, Y. & Cheng, X. (2013). A Biterm Topic Model for Short Texts. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445-1456.

[6] Anderson, T. W. & Goodman, L. A. (1957). Statistical Inference about Markov Chains. *Annals of Mathematical Statistics 28*, pp. 89-110.