*The Korean Journal of Applied Statistics* (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.1.075
**33**(1), 75–86

# Identification of sentiment keywords association-based hotel network of hotel review using mapper method in topological data analysis

Ye-Seul Jeon[a,1] · Jeong-Jae Kim[a]

[a]Department of Applied Statistics, Yonsei University; [b]Artificial Intelligence Lab, Daumsoft Inc.

(Received November 21, 2019; Revised December 27, 2019; Accepted January 2, 2020)

**Abstract**

Hotel review data can extract various information that includes purchasing factors that lead to consumption, advantages, and disadvantages for hotels. In particular, the sentiment keyword of the review data helps consumers understand the pros and cons of hotels. However, it is not efficient for consumers to read a large number of reviews. Therefore, it is necessary to offer a summary review to customers. In this study, we suggest providing summary information on sentiment keywords association as well as a network of hotels based on sentiment keywords. Based on a sentiment keyword dictionary, the extracted sentiment keywords associations construct the hotel network through topological data analysis based mapper. This hotel network allows a consumer to find some hotels associated with specific sentiment keywords as well as recommends the same related hotels. This summary information provides users with a summarized emotional assessment of hotels and helps hotel marketing teams understand consumers' perceptions of their hotel.

Keywords: topological data analysis, mapper, sentiment analysis, sentiment keywords association-based hotel network

## 1. 서론

최근 여행에 대한 관심이 높아지면서 여행과 밀접하게 연결되어 있는 숙박업에 대한 관심도 높아지고 있다. 소비자들은 숙박을 예약하기 위해 다양한 숙박 리뷰 사이트를 통해 해당 숙박에 대한 정보를 간접적으로 경험하고 파악한다. 이러한 리뷰 사이트는 구매 욕구가 존재하는 잠재적 소비자들이 실질적으로 소비가 이뤄진 내용에 대한 후기를 살펴봄으로써, 해당 상품의 긍정 및 부정적인 측면을 간접적으로 이해할 수 있도록 도움을 준다. 또한, 상품을 관리하고 판매하는 숙박업소 관계자도 소비자가 인식하고 경험한 자사의 상품에 대한 평가를 받아볼 수 있다는 측면에서 중요한 역할을 한다. 리뷰 데이터에서는 소비를 이끈 구매 요인, 상품에 대한 장점, 재구매 여부, 소비자의 인식 또는 상품의 한계 및 단점 등 다양한 정보를 추출할 수 있다. 특히, 상품에 대한 평가 및 반응을 살펴보기 위해 리뷰에 나타난 감성 키워드의 분포를 살펴볼 수 있다. 리뷰 데이터의 감성 키워드는 소비자들이 상품에 관해 이야기하고 있는 주요 내용을 파악하는 데 도움을 준다. 즉, 상품에 대한 긍정 또는 부정 반응이 상품에 대한 소비자들의 평

[1]Corresponding author: Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea. E-mail: jeon9677@yonsei.ac.kr

가를 대표적으로 요약한다. 하지만, 리뷰 데이터가 많은 경우에는 많은 양의 감성 키워드를 직접 살펴볼 수 없기 때문에, 감성 키워드의 긍정 및 부정에 대한 분포를 통해 상품에 대한 감성 정보를 요약할 수 있다. 즉, A라는 상품에 대해 긍정이 80%이고 부정이 20%라면, 소비자들은 대체로 A라는 상품에 대해 긍정적인 반응을 나타내고 있다는 것을 파악할 수 있다. 하지만, 이러한 정보는 전반적인 반응을 살펴볼 수 있지만, 구체적인 정보를 파악하기에는 여전히 한계가 존재한다. 즉, 어떠한 측면에서 긍정적인지 부정적인지를 파악하기 어렵다. 이를 위해, 감성 키워드의 관계망을 통해 상품에 대해 주로 언급된 감성 키워드들이 어떻게 연결이 되어있고, 어떠한 키워드가 대표성을 가지는지 파악할 수 있다. Ha 등 (2013)은 영화 리뷰데이터에서 감성 키워드의 분포 지도를 구축하여 유사한 영화를 추천한다. 감성 키워드의 분포 지도를 구축하기 위해 주성분 분석(principal component analysis; PCA)과 다차원척도법(multidimensional scaling; MDS)을 사용하며, 영화 추천을 위해 추천 분석 기법이 추가로 수행된다. 하지만, 구축된 감성 키워드 분포 지도만으로는 해당 감성 키워드가 어떤 영화에 속해 있고, 해당 영화가 어떤 감성 키워드로 다른 영화와 연결이 되어 있는지를 알 수 없기에, 별도의 추천 기법을 수행해야 한다는 점에서 다음과 같은 한계가 존재한다. 1) 주성분 분석과 다차원척도 분석 기법은 분석가의 주관적인 해석이 포함된다. 즉, 의미 있는 영역을 파악해 나름의 군집 분석을 수행해야 한다. 2) 해당 감성 키워드가 어느 영화에 해당하는지에 대한 내용을 파악하기 위해서, 다시 원 리뷰 데이터를 살펴봐야 한다.

이러한 한계를 보완하기 위해 본 연구에서는 감성 요약의 기법으로 군집 분석과 관계망(network)을 동시에 구축 할 수 있는 방법인 위상학적 데이터 분석 기반의 맵퍼(topological data analysis based mapper)를 제안한다. 맵퍼(mapper)는 복잡한 고차원 데이터를 2차원의 관계망 형태로 구축함으로써 고차원의 원 공간의 위상학적 구조를 간접적으로 파악할 수 있게 해준다는 장점이 있다. 이때, 데이터의 본래 가지고 있는 위상학적 구조를 잘 보존한다는 점에서 정보 손실을 최소화할 수 있으며, 눈으로 쉽게 관계들을 파악할 수 있는 관계망을 통해 정보를 전달한다는 점에서도 유의하다. 기존의 감성 키워드 분석들이 “놀라다”, “겁나다”, “감동적이다” 와 같은 단순한 감성 키워드 기반으로 분석을 했다면, 본 연구에서는 “가격”, “가성비”, “시설”과 같은 분석 대상과 “저렴함”, “좋음”, “조용함”과 같은 감성 대상과 긍부정이 정의된 감성 키워드를 기반으로 호텔 간의 관계망을 구축하고자 한다. 구축된 감성 키워드 기반 호텔 관계망으로 호텔 간의 관계들을 살펴볼 수 있으며, 연결된 호텔들이 어떠한 감성 키워드들로 연결되어 있는지 살펴볼 수 있다. 이를 통해 소비자들은 호텔에 대한 감성 평가를 한눈에 파악할 수 있으며, 동시에 해당 감성에 따라 연결된 유사한 호텔들을 추가로 탐색해볼 수 도 있다. 또한, 호텔 마케팅 및 전략 기획팀에서는 요약된 감성 정보들을 통해 소비자들의 인식을 파악할 수 있다. 본 논문의 구성은 다음과 같다. 2절에서는 제안방안의 단계별 분석 기법을 소개하고, 3절에서는 구축된 감성 키워드 관계망에 대한 실험 결과 및 분석을 한다. 마지막으로 4절에서는 결론 및 향후 연구에 관해 기술한다.

## 2. 제안방안

본 논문에서는 위상학적 데이터 분석 기법 중 하나인 맵퍼를 활용해 호텔 리뷰 데이터의 감성 키워드 호텔 관계망 구축을 제안한다. 위상학적 데이터 분석은 크게 퍼시스턴트 호몰로지(persistent homology)기법 Zomorodian과 Carlsson (2005)과 맵퍼 Singh 등 (2007) 방법으로 나눌 수 있다. 두 방법론 모두 데이터의 위상학적 구조를 파악하기 위한 방법론이며, 전자는 데이터의 전반적인 구조를 복원하고 다른 물체와 비교하기 위해 주로 사용되며, 후자의 방법은 데이터의 전반적인 구조를 시각적으로 파악하기 쉬운 2차원의 관계망의 형태로 제공된다.

본 논문에서는 맵퍼 기법을 통해 관계망의 시각화에 중점을 둔다. 맵퍼의 기본 원리는 고차원의 데이터로부터 $k$개의 심플릭스(k-simplex)를 구성함으로써 단순 조합 구성요소에서 위상 공간을 간단하게 표현

**Table 2.1.** Natural language processing result with aspect-sentiment keywords

| 리뷰 | 화장실 냄새는 만족스럽다. 하지만, 욕실 용품은 부실했습니다. |
|---|---|
| 문장 분리 | 문장 1 : 화장실 냄새는 만족스럽다.<br>문장 2 : 하지만, 욕실 용품은 부실했습니다. |
| 자연어 처리 결과 | 문장 1 : [화장실], [냄새+는] [만족스럽+다]<br>문장 2 : [하지만], [욕실], [용품+은], [부실하+었+습니다] |
| 주요 키워드 추출 | 문장 1 : 화장실, 냄새, 화장실냄새, 만족스럽다, 냄새만족스럽다, 화장실냄새만족스럽다<br>문장 2 : 욕실, 용품, 욕실용품, 용품부실하다, 욕실용품부실하다 |

한다. 즉, 위상 공간의 연속적인 형상을 처리하는 복잡성을 상대적으로 간단한 점, 선, 세모, 삼각뿔 등의 조합 및 계산하여 2차원의 공간으로 표현할 수 있다. 이는 시각적으로 관계망을 쉽게 파악하기 위함이며, 해당 저차원으로도 충분히 고차원 원 공간의 위상학적 구조를 파악할 수 있다. 기존의 고차원의 데이터를 2차원으로 축소하는 주성분 분석이나 다차원척도법 등과 달리 각 차원이 가지는 의미를 주관적으로 해석 및 판단이 필요 없다. 본 연구는 각 차원이 가지는 의미보다는 2차원의 공간에서 관계망의 연결성에 대해 중점을 둔다. 이를 위해 맵퍼에서는 크게 두 가지 단계를 수행한다. 먼저, 위상 공간에 존재하는 고차원의 데이터 집합을 실숫값 함수를 이용해 분석 가능한 실수의 공간인 2차원 공간으로 투사한다. 실수 값 함수는 데이터의 밀도, 너비 등의 기하학적 구조를 나타낼 수 있는 함수이거나 기존에 사용될 수 있는 2차원 공간으로 투사시키기 위한 차원축소 함수 등이 사용된다. 다음 단계는 실수 함수로 재정의된 데이터 집합을 부분적 군집화 기법을 통해 기존 데이터 집합내의 세부 집합들을 생성하고 이들 간의 상호관계를 살펴봄으로써 데이터 내의 관계들을 파악한다. 이를 통해 생성된 부분적 군집이 하나의 노드(node)가 되고, 노드 간의 구성하고 있는 속성이 교집합이 존재하는 경우 간선(edge)이 생성되어 관계망이 형성된다. 위상학적 데이터 분석의 맵퍼를 활용하여 감성 키워드와 호텔 간의 관계 그리고 호텔들 간의 관계를 파악하기 위해 다음과 같은 절차로 연구를 진행한다. 먼저, 데이터 셋을 구축하는 단계로 호텔 리뷰 데이터에서 사전에 구축된 감성 사전을 기반으로 감성 키워드를 추출한다. 추출된 감성 키워드 기반으로 호텔 감성 키워드 임베딩(embedding) 방식을 적용한다. 다음으로, 생성된 호텔 감성 키워드 임베딩을 위상학적 데이터 분석의 맵퍼 입력값으로 사용하여 감성 키워드 기반 호텔 관계망을 구축한다. 마지막으로, 구축된 관계망을 통해서 위상학적 분석 및 감성 키워드 기반의 호텔 분석을 한다.

## 2.1. 데이터 전처리

**2.1.1. 자연어 처리를 통한 주요 키워드 추출** 본 논문에서는 리뷰 데이터에서 명사 위주의 키워드가 아닌 감성 대상과 분석 대상을 갖는 주요 키워드를 추출한다. 호텔 리뷰 데이터는 소셜 데이터 특성상 사람이 사용하는 구어체 특성이 있어, 문장 분리 및 띄어쓰기 교정과 같은 기술이 요구되는 자연어 처리를 해야 한다. 이를 위해 다음소프트의 자연어 처리 분석 기술(Text Mining Engine Version 2; TM2) (http://www.daumsoft.com/contextualFinder.html)을 활용하여 리뷰 데이터의 원문을 소셜 데이터 특성에 맞게 자연어 처리하여 주요 키워드를 추출하며, Table 2.1이 그 예시이다. 먼저, 리뷰의 원문에서 문장 부호를 기반으로 문장 분리를 적용하며, 문장 부호가 생략된 경우는 종결 어미를 이용하여 분리한다. 다음으로 규칙 및 통계기반으로 띄어쓰기 오류를 교정한다. 다음으로 형태소, 고유명사, 복합명사 및 불용어 사전과 같은 형태소 사전들과 품사 전이 확률, 어휘 문맥 통계 지식과 같은 언어 통계 데이터를 기반으로 형태소 분석을 한다. 마지막으로, 그룹어 및 유사어가 처리된 개체명 및 명사, 서술어를 주요 키워드로 추출한다.

**2.1.2. 감성 사전 기반 감성 키워드 추출** 사용자는 특정 호텔을 경험하고, 가격, 시설, 서비스, 위치 등과 같은 각 요소에 대해 리뷰를 작성한다. 사용자들이 개재한 의견이나 감정 등의 패턴으로부터 각 요소에 대한 긍정 또는 부정의 감성 정보를 추출할 수 있다. 본 방법론에서는 추출된 주요 키워드 가운데 분석 대상 및 감성 대상이 존재하는 키워드를 추출하기 위해서 호텔 도메인 지식자들이 호텔 도메인의 감성 사전을 구축한다. 각 키워드는 분석 대상과 감성 대상이 정의된다. 분석 대상은 "가격/가성비", "객실", "위치/교통", "직원/서비스" 등에 해당하며, "객실"는 세부적으로 "침구", "냉난방", "어메니티", "욕실" 등으로 구성된다. 감성 대상은 분석 대상에 따라 존재하는 감성으로, "객실"의 "냉난방"의 경우는 "온도", "청결", "품질" 등의 감성이 존재할 수 있다. "온도"의 경우는 세부적으로 "더움", "추움", "시원함", "적당함", "따뜻함" 등으로 구성된다. 예를 들어, "화장실냄새만족하다" 키워드는 분석 대상은 "객실"의"욕실"이며, 감성 대상은 "청결도"의 "청결함"에 해당하며, 감성은 "긍정"에 해당한다. 또한 "욕실용품부실하다" 키워드의 분석 대상은 "객실"의 "어메니티"이며, 감성 대상은 "품질"의 "좋지않음"에 해당한다. 구축된 감성 사전을 기반으로 Table 3.1에서는 추출된 키워드 가운데 "화장실냄새만족하다"와 "욕실용품부실하다"가 감성 키워드로 추출된다. 구축된 호텔 도메인의 감성 사전은 분석 대상과 감성 대상으로 범주화되어 있어, 호텔의 리뷰를 요약하는데 용이하다. 예를 들어, 분석 대상 "객실"에서 분석하는 경우, 해당 범주를 갖는 키워드들의 긍부정 감성 분포에 따라 "객실"의 긍정성 또는 부정성을 확인할 수 있다.

**2.1.3. 호텔 단위 감성 키워드 임베딩** 최근, 리뷰데이터를 임베딩하는 방식은 단어 사전을 기반으로 하여 각 단어가 발현한 경우를 1로, 아닌 경우 0으로 표현하는 원-핫 인코딩(one-hot encoding) 또는 뉴럴 네트워크(neural network) 기반의 단어 임베딩 모델 중 하나인 Word2Vec을 활용한다. 감성 키워드를 활용하여 임베딩 하는 경우, 리뷰마다 추출된 감성 키워드 개수가 작아 중심단어가 주어졌을 때 주변 단어가 나타날 확률을 학습하는 Word2Vec의 Skip-gram 임베딩 방식은 적합하지 않다. 또한, 원-핫 인코딩 방식은 감성 키워드 단어 사전의 크기가 6,531개로, 평균 1.4개가 값이 1이고, 대부분이 0인 희소 행렬(sparse matrix) 형태를 보이고 있어, 리뷰 단위의 임베딩은 적합하지 않다. 본 논문에서는 리뷰 단위의 원-핫 인코딩 방식을 적용하고, 이를 호텔 단위로 각 감성 키워드별 평균을 한 호텔 단위 감성 키워드 임베딩을 제안한다. 호텔 리뷰 데이터 $D$는 다음과 같이 표현한다. 특정 호텔 $h_i$는 $n$개의 리뷰 $r \in R$를 가지며, 특정 리뷰 $r_i$는 2.1.1절과 2.1.2절 방식을 통해 원문에서 추출된 $m$개의 감성 키워드 $k \in K$로 구성된다. 전체 호텔 리뷰 대상 감성 키워드 집합을 통해 감성 키워드 사전 $b =< k_1, k_2, \ldots, k_m >$를 구축한다. 특정 호텔 $h_i$의 감성 키워드 임베딩 방식 $w_{h_i}$은 감성 키워드 사전 $b$를 집합으로 하여, 호텔 $h_i$의 리뷰들에서 나타난 각 감성 키워드의 발현 빈도수 평균값을 사용한다. 예를 들어, 호텔 $h_i$의 리뷰가 $R = \{r_1 =< k_1, k_2 >,\ r_2 =< k_2, k_3 >\}$이고, 감성 키워드 사전이 $b =< k_1, k_2, k_3, k_4 >$와 같다면, 호텔 $h_i$의 감성 키워드 임베딩 $w_{h_i}$은 $< 0.5, 1, 0.5, 0 >$과 같다.

### 2.2. 맵퍼를 통한 감성 키워드 관계망 구축

앞서 설명된 호텔 감성 키워드 임베딩을 맵퍼의 입력으로 활용하여 감성 키워드 관계망을 구축한다. 맵퍼는 필터링(filtering), 분할(partitioning), 군집(clustering), 시각화(visualization) 순으로 구성되며 Figure 2.1과 같다. 필터링 단계는 실수함수를 통해 데이터의 위상 공간에서 실수 공간으로 투사시키며, 분할 단계는 실수 공간의 데이터 집합 내에서 부분적 군집 분석을 수행한다. 군집 단계에서는 부분적 군집된 노드들 기반으로 교차성을 수행하여 노드 간의 연결을 수행한다. 마지막으로 시각화 단계를 통해서 연결된 노드들 기반으로 관계망을 구축한다.

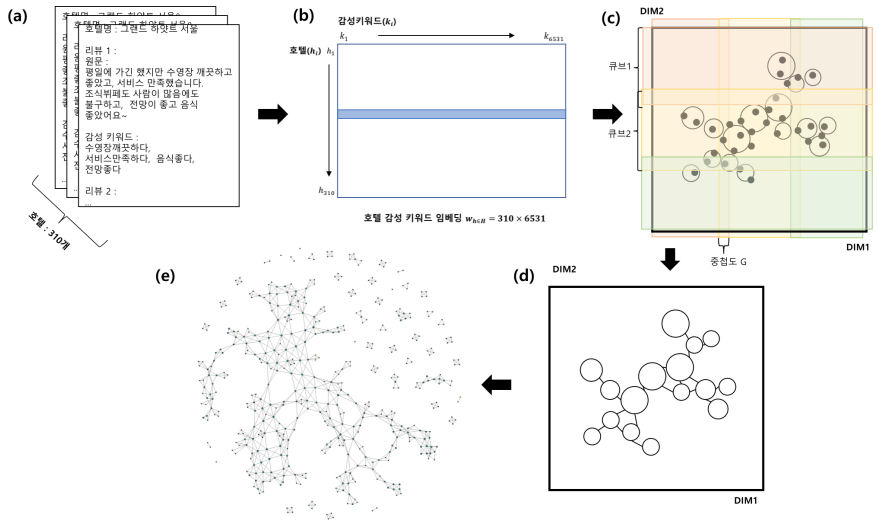



**Figure 2.1.** A framework of the proposed methodology.

**2.2.1. 필터링 단계** 맵퍼에서의 필터링 단계는 실수함수를 통해 위상공간에 존재하는 데이터 집합을 실수 공간으로 투사 및 좌표화 하는 단계이다. 실수함수는 필터 함수 또는 렌즈라고 하며, 데이터의 특성에 따라 연구자가 실수함수를 지정할 수 있다. 만약, 데이터의 기하학적 구조를 주로 살펴보고자 한다면 기하학적 특성을 실숫값으로 산출할 수 있는 함수를 필터 함수로 지정하면 된다. 예를 들어, 도넛과 컵이 원 하나의 모양을 가지고 있다는 측면에서 원의 개수를 산출 할 수 있는 필터 함수를 지정한다면 두 물체는 위상학적 구조에서 같은 물체로 인식된다. 이처럼 필터 함수를 무엇으로 설정할 것인가에 따라 데이터의 위상학적 구조를 여러 방면에서 살펴볼 수 있다. 특히 위상 공간 간의 거리를 계산할 수 있고 동시에 데이터로부터 기하학적 구조의 정보를 추출할 수 있는 보편적인 필터 함수로는 가우시안 커널을 이용한 밀도추정 기법이라든지 중심성으로부터 얼만 큼 벗어나 있는지를 추정하는 Eccentricity 기법들을 사용할 수 있다 Singh 등 (2007). 또한 이러한 필터 함수는 고차원 데이터를 저차원 공간으로 투사시키는 차원축소 기법으로도 사용할 수 있다 Guo와 Banerjee (2007).

본 연구에서는 고차원 데이터의 공간으로부터 저차원의 공간으로 투사하는 차원 축소 기법을 필터 함수로 설정해 실수공간의 데이터 집합을 형성한다. 차원축소 함수로 전통적인 선형적 차원축소 기법인 주성분 분석과 고차원 공간에서의 거리를 2차원 공간에서의 계산 가능한 거리로 산출하는 다차원척도법 등이 존재한다. 본 연구에서는 비선형 차원 축소 방법으로 대표적인 확률적 임베딩 기법(stochastic neighborhood estimation; SNE)을 선택하였다. 확률적 임베딩 기법은 확률적으로 고차원 원 공간의 이웃 간의 거리를 최대한 보존하여 학습하여, 저차원 공간으로 매핑하면서 데이터의 지역적 구조 정보를 유지한다. 고차원의 원공간과 저차원 공간에서의 서로 다른 두 지점$(i, j)$를 선택할 확률은 $p$와 $q$와 식 (2.1)과 같다.

$$p_{j|i} = \frac{e^{-\frac{|x_i - x_j|^2}{2\sigma_i^2}}}{\sum_k e^{-\frac{|x_i - x_k|^2}{2\sigma_i^2}}}, \qquad q_{j|i} = \frac{e^{-|y_i - y_j|^2}}{\sum_k e^{-|y_i - y_k|^2}}. \tag{2.1}$$

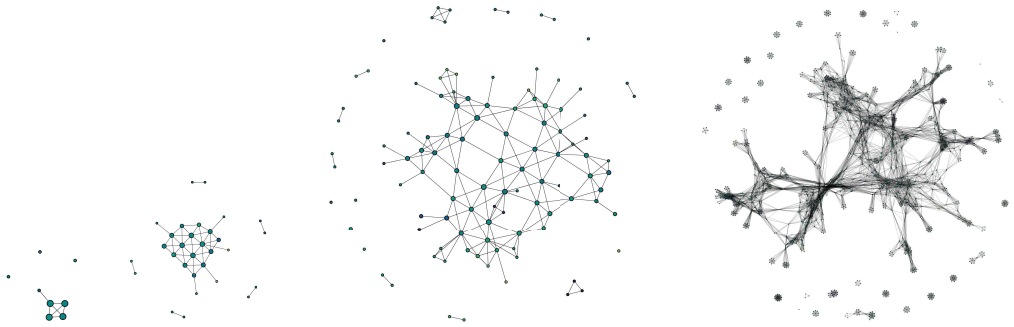
두 확률 분포를 측정하는 쿨백–라이블러 발산(Kullback-Leibler divergence)을 통해 $p$와 $q$의 분포가 완전히 다르면 1, 동일하면 0의 값을 갖게 된다. 확률적 임베딩 기법은 식 (2.1)를 최소화하여 학습한다.

$$\text{cost} = \sum_i \text{KL}(P_i||Q_j) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}. \tag{2.2}$$

가우시안 분포를 전제하는 확률적 임베딩 기법는 멀리 떨어져 있는 두 지점이 선택될 확률이 낮아 거리의 정도에 대한 학습이 잘 이루어지지 않는다. 이를 해결하기 위해서 Maaten과 Hinton (2008)은 확률적 임베딩 기법에 가우시안 분포가 아닌 스튜던트의 분포가 접목된 $t$-분포 확률적 임베딩 기법($t$-distribution stochastic neighborhood estimation; $t$-SNE)를 제안하였고, 본 논문에서는 필터 함수로 $t$-분포 확률적 임베딩 기법을 적용하였다. 또한 비정형 데이터인 텍스트 데이터의 경우, 선형성을 띠지 않아 $t$-분포 확률적 임베딩 기법의 방법을 통해 분석하는 것이 정보 손실을 최소화할 수 있다.

**2.2.2. 분할 단계** 앞선 필터링 단계는 고차원의 위상학적 정보를 실수의 공간으로 투사 시켜 분석 가능한 실수 공간으로 데이터를 변환시켰다면, 분할 단계에서는 실수의 공간에서 위상학적 구조를 추론하기 위해 전제 데이터로부터 부분 집합들을 생성하는 단계이다. 이 단계가 기존의 차원축소 기법과 달리 원 데이터로부터 부분 집합 데이터를 추출하고 다음 단계에서의 부분 집합 내에서 군집 분석을 수행한다는 점에서 차이가 존재한다. 분할 단계는 원 데이터로부터 부분 집합들을 생성하기 위해 두 개의 초매개변수가 필요하다. 두 개의 초 매개변수는 전체 데이터 집합을 어느 정도의 해상도에서 바라볼 것인지 혹은 부분 데이터 집합 간의 공유하고 있는 중첩도를 어느 정도로 설정할 것인지에 대한 부분을 결정해준다. 첫 번째 초매개변수 해당도(resolution; $R$)는 전체 데이터 집합을 어느 정도의 해상도에서 바라볼 것인지 결정하며, 두 번째 초매개변수 중첩도(gain; $G$)는 부분적 집합 간의 공유되는 중첩도에 대해 결정한다. Figure 2.1(c)에서 해상도 $R$은 $x$축과 $y$축에 적용되는 큐브의 개수에 해당하며, 중첩도 $G$는 각 큐브 간의 중첩되는 확률에 해당한다. 해상도 $R$의 값이 높을수록 다수의 큐브가 생성되므로 데이터들이 개별적인 부분 집합으로 구성되어, 데이터의 지역성을 관찰할 수 있다. 반면에, 해상도 $R$이 낮을수록 큐브 내에 다수의 데이터가 군집화되기 때문에 데이터의 중심성을 살펴볼 수 있다. 중첩도 $G$의 값이 높을수록 큐브들간의 공유하고 있는 면적이 넓어져 큐브 간의 공유되는 노드들이 많아져 연결성이 많아진다. 반면에, 중첩도 $G$가 낮을수록 연결성이 적은 단순한 형태의 관계망이 형성된다. 본 연구에서는 다양한 실험을 통해서 $R$과 $G$의 값을 변화시키면서, 관계망의 구조가 크게 변화하지 않는 구간을 찾아 $R$과 $G$를 각각 20과 0.5로 설정하였다.

**2.2.3. 군집 및 시각화 단계** 앞서 분할 단계를 통해 데이터의 부분 집합을 구했다면 군집 단계는 부분 집합 내에서 군집 분석을 수행함으로써 부분 집합 내에서의 구조를 살펴본다. 이를 부분적 군집 분석이라고도 한다. 부분적 군집 분석은 계층적 군집 분석 중 하나인 계층적 병합군집화(agglomerative clustering)을 통해 유사도가 높은 군집을 합치면서 군집 개수를 줄여간다. 각 큐브 안에서의 데이터 개수가 일정하지 않음으로 군집 개수를 지정하기 어려우므로 상향식의 계층적 병합군집화 방식으로 군집 개수를 줄여간다. 군집 간의 거리를 측정하는 방식에서는 군집 간의 데이터의 모든 조합에 대한 최솟값(single), 평균(average) 또는 최댓값(max)을 적용한다. 거리 척도로는 유클리드 거리, 맨해튼 거리 또는 코사인 거리를 적용한다. Figure 2.1(d)와 같이 노드들 간의 공통적인 속성 즉, 교집합이 존재하는 경우, 그 군집은 다른 군집과 연결되어 있다고 여겨 하나의 간선이 연결된다. 이때 하나의 군집은 다른 군집들과 여러 개의 간선이 연결될 수 있다. 예를 들어, 군집화된 호텔들은 하나의 노드이며, 각 군집이 하나 이상의 호텔을 공유하고 있으면 간선으로 연결된다. 이러한 단계를 통해 기존의 연구와 달리, 군집 분석과 동시에 연결성을 수행하게 된다. 앞선 연결성으로 점, 선, 삼각형, 삼각 뿔 등의 $k$개의 심플릭스



(a) $R = 2$ (b) $R = 4$ (c) $R = 8$ (d) $R = 32$

**Figure 3.1.** Sentiment keywords networks according to a different resolution value, $R$.

가 구축이 되며 이러한 형태를 통해 원래 데이터의 위상학적 구조를 파악할 수 있다. 이러한 단계를 거쳐 Figure 2.1(e)와 같은 호텔 연결망을 최종적으로 구축할 수 있다.

## 3. 실험 결과 및 분석

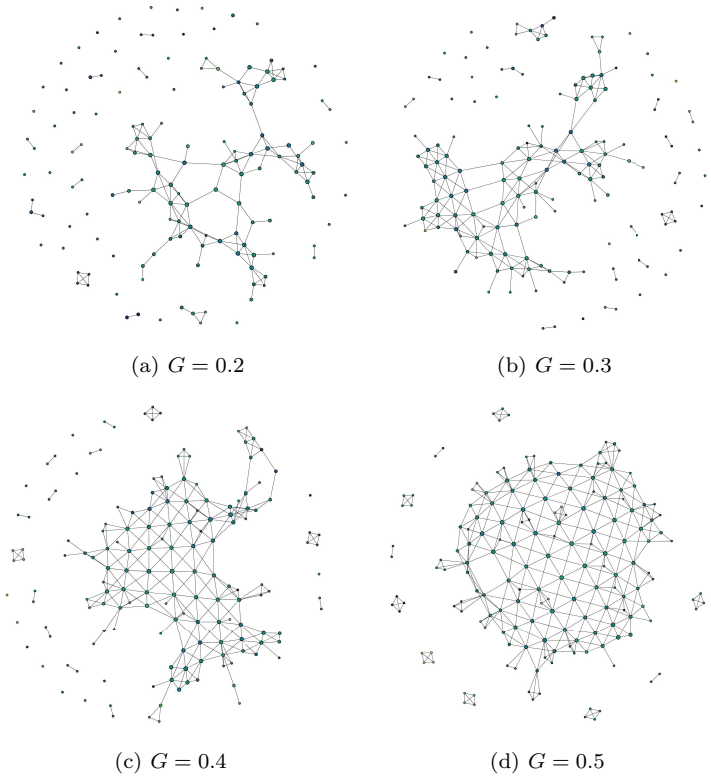
### 3.1. 실험 환경

제안 방안에서 활용된 호텔 리뷰 데이터 셋은 대표적인 호텔 예약 업체 중 하나인 데일리 호텔(https://m.dailyhotel.co.kr/)의 리뷰 정보와 호텔 정보를 크롤링하여 수집하였다. 수집된 리뷰는 총 216,267개이며, 호텔의 개수는 7,202개이다. 수집된 호텔의 등급은 5–1성급, 비즈니스, 콘도 등의 10개의 등급으로 나뉘며, 본 논문에서는 시설 및 서비스 등에 대한 요소별 평가가 유의미한 3성급 이상의 호텔 대상의 리뷰들에 대해 감성 키워드 관계망을 구축한다. 호텔 리뷰를 바탕으로 구축된 호텔 도메인 감성 사전은 총 23,745개로, 실제 리뷰에서 추출된 감성 키워드의 개수는 6,531개이다. 비교 모델인 비계량형 다차원척도법은 소프트웨어 scikit-learn(https://scikit-learn.org/)을 통해서 구현하였으며, 제안방안에서의 맵퍼는 소프트웨어 KeplerMapper(https://kepler-mapper.scikit-tda.org/)를 활용하여 구현하였다.

### 3.2. 위상학적 데이터 분석

**3.2.1. 맵퍼에 대한 분석** 분할 단계에서는 해상도 $R$와 중첩도 $G$에 대한 매개변수 탐색을 통해서 감성 키워드 관계망의 위상학적 분석을 한다. Figure 3.1은 해상도 $R$에 따른 감성 키워드 관계망이다. Figure 3.1(a)는 해상도 $R$의 값은 2로, 크게 3개의 군집을 형성한다. 2개의 작은 군집은 각각 호텔 $h_{144}$과 $h_{194}$ 호텔이며, 대표 감성 키워드는 각각 “리모델링깨끗한다”와 “직원분많다”이며, 군집화되지 않은 이상점이다. 나머지 하나의 큰 클터스터는 5개의 작은 군집을 하고 있으며, 각각 230, 226, 216, 210 그리고 1개의 호텔로 구성되며, 대표 감성 키워드는 각각 “소음들리다”, “용품청결하다”, “호텔가고싶다”, “분위기아쉽다”와 “조식뷔페만족스럽다”이다. Figure 3.1 (b)–(d)는 해상도 $R$이 높아짐에 따라 군집의 지역성을 확대해서 해석 가능하다.

Figure 3.2는 중첩도 $G$에 따른 감성 키워드 관계망이다. 중첩도 $G$는 Partitioning 단계에서 각 큐브 간의 중첩도를 의미한다. 중첩도 $G$값이 낮을수록 큐브간의 간격이 멀어져 Figure 3.2(a)와 같이 군집 간의 간선의 수가 적었지만, Figure 3.2(d)와 같이 중첩도 $G$값이 높을수록 군집 간의 간선이 다수 연결된



(a) $G = 0.2$ (b) $G = 0.3$

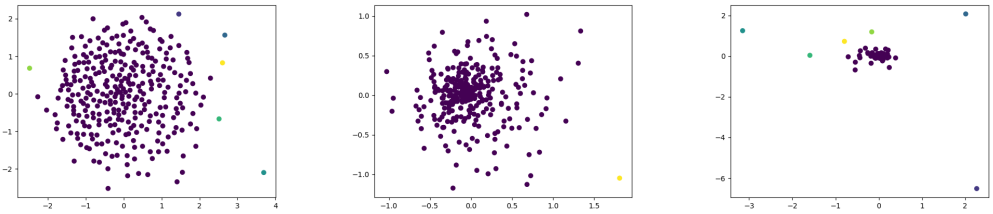
(c) $G = 0.4$ (d) $G = 0.5$

**Figure 3.2.** Sentiment keywords networks according to a different overlap value, $G$.

다. 위의 실험을 바탕으로, 해상도 $R$은 축소를 통해서 각 군집의 대표성을 갖는 요약 정보들을 추출할 수 있으며, 확대를 통해서 각 군집의 지역성에 대한 정보를 추출할 수 있다. 중첩도 $G$는 군집 간의 연결성에 초점이 맞춰져 있는 것을 관찰할 수 있다. 호텔 감성 키워드 임베딩은 필터링 단계를 통해서 2차원으로 차원 축소되어, 데이터에 적합한 거리 척도를 찾을 필요가 있다. 본 실험에서는 군집간의 거리 방식이 최소이면서 거리 척도가 맨해튼 거리가 적합함을 다양한 실험을 통해서 얻었다.

### 3.3. 실험 결과

**3.3.1. 기존 방안 분석** 호텔 리뷰 데이터는 비선형적 특성이 있어 분석하고자 하는 감성 키워드의 선형성을 전제로 하는 주성분 분석의 결과는 제 1주성분과 제 2주성분의 설명력이 합산 25% 밖에 되지 않는다. 반면에, 다차원척도법은 다차원 관측값, 개체들 간의 거리 또는 비유사성을 이용하여 개체들을 고차원의 원 공간으로부터 저차원의 공간상에 위치 시켜, 개체들의 상대적 위치 등을 통해 개체들의 사이 관계가 설명할 수 있다. 호텔 단위 감성 키워드 임베딩은 연속형 변수가 아닌 키워드별 도수 자료이기 때문에 Kruskal (1964)의 비계량형 다차원척도법을 적용한다. 다차원척도법에서 사용되는 개체 간의 공간상의 거리는 Jung 등 (2019)의 카이제곱 거리와 가중 유클리드 거리 포함 카이제곱 거리를 활용한다. 두 호텔의 감성 키워드 임베딩은 서로 독립적이기 때문에 각 감성 키워드마다 도수 관측치에 대한 카이제곱 통계량을 통해서 두 호텔의 상대적 거리를 계산한다. 호텔 단위 감성 키워드 임베딩간의 카이



(a) 카이제곱 거리 다차원척도법 (b) 카이제곱 거리 비계량형 다차원척도법 (c) 가중 유클리드 거리 포함 카이제곱 거리 비계량형 다차원척도법

**Figure 3.3.** Results of previous methodologies of data reduction.

제곱 거리 $d_{\mathrm{CD}}$는 식 (3.1)과 같이 표현한다.

$$d_{\mathrm{CD}} = d(X_{rl}, X_{sm}) = \left[ \sum_{t=1}^{q} \frac{(x_{rlt} - x_{smt})^2}{(x_{rlt} + x_{smt})} \right]^{\frac{1}{2}}. \tag{3.1}$$

이때, $X_{rl} = (x_{rl1}, x_{rl2}, \ldots, x_{rlq})^t$, $l = 1, \ldots, n_r$은 $r$번째 개체 $X_r$의 $l$번째 문서이며, 마찬가지로 $X_{sm} = (x_{sm1}, x_{sm2}, \ldots, x_{smq})^t$, $m = 1, \ldots, n_s$은 $s$번째 개체 $X_s$의 $m$번째 문서에 해당한다. 또 $q$는 호텔 단위 감성 키워드 임베딩의 길이로 6,531에 해당한다. 또한, 다차원척도법에서의 비유사성 측정을 위해서 가중 유클리드 거리를 포함 카이제곱 거리 $d_{\mathrm{WED}}$를 활용하며, 식 (3.2)와 같이 표현한다.

$$d_{\mathrm{WED}} = d(X_{rl}, X_{sm}) = \left[ \sum_{t=1}^{q} \frac{(\frac{x_{rlt}}{x_{rl\cdot}} - \frac{x_{smt}}{x_{sm\cdot}})^2}{(\frac{x_{rlt}}{x_{rl\cdot}} + \frac{x_{smt}}{x_{sm\cdot}})} \right]^{\frac{1}{2}}. \tag{3.2}$$

Figure 3.3은 다차원척도법과 비계량형 다차원척도법으로 호텔 단위 감성 키워드 임베딩의 고차원의 데이터를 2차원의 공간으로 투사한 후 비지도 학습 방식으로 병합 군집한 결과이다. Figure 3.3(a)는 호텔 단위 감성 키워드 임베딩 값에 공간상의 거리 척도로 카이제곱 거리를 사용한 다차원척도법의 결과이며, 호텔들의 분포가 하나의 원의 형태를 가지고 있어 뚜렷한 패턴이 나타나지 않았다. 또한, 군집 결과는 중심에 분포한 대부분의 호텔이 하나의 군집으로 묶여, 이상점에 해당하는 호텔들은 개별적인 군집을 형성한다. Figure 3.3(b)는 감성 키워드 임베딩의 도수 자료의 특성을 반영해 카이제곱 거리 기반 비계량형 다차원척도법을 적용한 결과이다. 분석 결과 앞서 분석한 다차원척도법보다 중심으로 밀집된 호텔들과 이외의 호텔들로 분포가 나뉘어 있는 것을 살펴볼 수 있다. 그런데도 군집 분석 결과는 하나의 호텔을 제외하고는 모든 호텔이 하나의 군집으로 형성되어 있는 것을 볼 수 있다. 즉, 이는 다수의 호텔을 구성하고 있는 감성 키워드의 정보들이 차원을 축소했을 때 정보 손실이 이루어져 차별적인 패턴들을 찾는 데 한계가 존재하는 것으로 볼 수 있다. Figure 3.3(c)는 호텔 단위 감성 키워드 임베팅 데이터의 0이 대부분인 희귀성(sparsity)을 고려해 가중 유클리드 포함 카이제곱 거리 기반 비계량형 다차원척도법을 적용한 결과이다. 호텔들의 리뷰를 살펴보면 중점적으로 주로 이야기되는 가격 및 시설에 대한 긍정 감성 키워드가 주로 분포되어 있고, 이외의 호텔별로 다른 속성의 감성 키워드들이 분포되어 있다. 앞선 방법들과 달리 속성이 차지하는 비율에 대한 가중치를 줌으로써 특정 호텔에서만 나타나는 감성 키워드들의 정보가 잘 보존되는 것을 확인할 수 있다. 하지만, 세 방법론 모두 개별적인 몇 개의 호텔들만 다른 군집으로 나타나고 대부분의 호텔이 하나의 군집으로 편향되어 있어 감성 정보를 추출하는 데 한계가 존재한다.

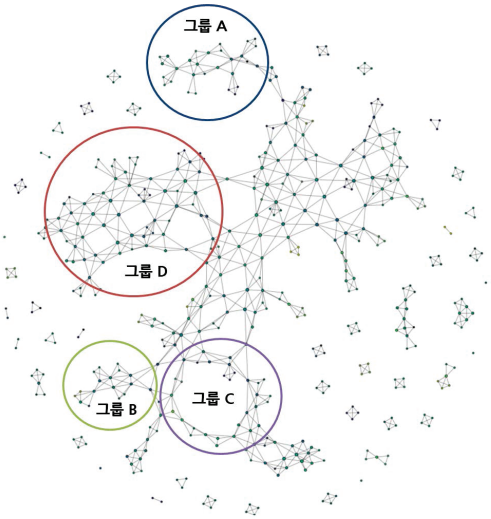



**Figure 3.4.** Hotel network based on sentiment keywords association using Mapper.

**3.3.2. 제안 방안 분석** Figure 3.4는 앞선 기존 방안의 실험과 같은 호텔 감성 단위 임베딩을 위상학적 데이터 분석 기반 맵퍼에 적용한 호텔 관계망이다. 맵퍼를 통해 구축된 호텔 연결망 결과, 다음과 같은 의미 있는 그룹들을 살펴볼 수 있다. 먼저, 중심 연결망으로부터 상단에 발현된 그룹 A 연결망을 살펴보면 주로 3성급의 호텔들이 묶여 있는 것을 볼 수 있다. 이는 부정적인 “히터안되다”, “호텔시설부족하다”, “직원마음에들지않다” 등의 감성 키워드 기반으로 3성급 호텔들이 묶여 있는 것을 볼 수 있다. 상단 부분이 주로 3성급 호텔에 대한 연결망이라면 반대 측에 속해 있는 하단 부분의 연결망 그룹 B와 그룹 C는 5성급 호텔에 대한 연결망들이 주로 구성되어 있는 것을 볼 수 있다. 그룹 B와 그룹 C는 같은 5성급 호텔에 대한 이야기이지만, 그룹 B는 “주차공간넓다”, “시설진짜좋다” 등의 긍정 감성 키워드가 관련되어 이야기되는 것을 볼 수 있다. 반면에 그룹 C는 같은 5성급 호텔이지만 “타올부족하다”, “침대별로이다”, “화장실청소안되다”, “침구얇다” 등의 부정 감성 키워드가 관련되는 것을 관찰할 수 있다. 호텔 연결망의 상단부분이 3성급 호텔이 주로 분포되어 있고 하단 부분의 연결망이 5성급 호텔이 주로 분포되어 있다면, 호텔 연결망의 중심 부분은 주로 4성급 호텔이 분포한다. 특히 그룹 D의 경우는 일부 3성급과 4성급 호텔이 섞여 있는 연결망들이 있고, 4성급만 존재하는 연결망이 존재한다. 일부 3성급과 4성급 호텔이 섞여 있는 연결망의 경우, “가격대비쓸만하다”, “거리편하다”, “방안깔끔하다”, “특가저렴하다” 등의 긍정 감성 키워드가 존재하는 것을 볼 수 있다. 이를 통해 4성급 중에서도 특가를 이용해 3성급 수준의 가격으로 사용해 만족도를 표현한 경우와 3성급 호텔 중에서도 가격대비 평가가 좋았던 호텔들임을 유추해볼 수 있다. 같은 그룹 D이지만, 4성급 호텔의 연결성을 살펴보면, 중앙 연결망의 대부분의 보편적인 4성급 호텔에 대한 긍정 반응과 달리 부정 반응들이 3성급 호텔들과 같이 연결되어 있는 그룹 D에 속해 있는 것을 볼 수 있다. 이러한 4성급 호텔들의 부정 감성 반응은 “화장실방음안되다”, “샤워시설 낙후되다”, “조식형편없다” 등의 이야기들로 보통 4성급 호텔들에 대한 긍정적인 감성 키워드가 주로 이룸에도 부정 감성 반응들은 따로 형성되는 것을 확인할 수 있다. 이를 통해, 3성급과 4성급 호텔이 연결되어 있는 경우, 4성급 호텔에서는 긍정적인 부분이 가격에 대한 부분 이외에 존재하지 않고 대부분에 부정적인 반응들이 나타나는 호텔들과 4성급 호텔 수준은 아니지만 3성급 중에서도 4성급 호텔만큼이나 긍정 반응이 높은 경우 하나의 연결망으로 연결되는 것을 확인해 볼 수 있었다. 이

**Table 3.1.** Representative Sentiment keywords and hotels of groups

| 그룹 | 그룹 대표 호텔 | 그룹 대표 감성 키워드 |
|---|---|---|
| 그룹 A | $h_{4789}, h_{4950}, h_{523}, h_{5193}$ 등의 3성급 호텔 | 허터안되다, 호텔시설부족하다, 직원마음에들지않다 하수구없다, 응대부족하다 |
| 그룹 B | $h_{4546}, h_{4551}, h_{4545}$ 등의 5성급 호텔 | 주차공간넓다, 시설진짜좋다, 잠자리만족하다, 위치깔끔하다 |
| 그룹 C | $h_{4348}, h_{38}, h_{7011}$ 등의 5성급 호텔 | 침구류쾌적하다, 뷰완벽하다 편의시설만족스럽다, 다음번가고싶다, 방음좋다 |
| 그룹 D | $h_{1632}, h_{1147}, h_{1204}$ 등의 3–4성급 호텔 | 방안깔끔하다, 특가저렴하다, 가격대비쓸만하다 |
| | $h_{1130}, h_{1133}, h_{7009}, h_{4404}$ 등의 4성급 호텔 | 화장실방음안되다, 탁자넓지않다, 놀거리가깝다 대중교통이용하기좋다, 거미줄가득하다, 샤워시설낙후되다 |

러한 호텔 연결망을 통해 호텔의 성급에 따라 연결망이 형성되는 것을 볼 수 있었고, 긍정 및 부정 감성 키워드가 리뷰에서 얼 만큼 발현되었냐에 따른 연결망임을 고려했을 때, 긍정 및 부정 감성 리뷰는 호텔의 성급에 따라 영향을 받는 것을 유추해볼 수 있다. 즉, 5성급일수록 주로 긍정적인 부분이 높고 3성급일 수록 주로 부정적인 부분이 높은 것을 확인할 수 있었다. 다만, 3성급 중에서도 4성급 만큼이나 만족도가 높은 경우, 그리고 4성급 중에서도 가격에 대한 만족도가 높은 경우에는 하나의 연결망으로 묶이는 것으로 보아, 4성급에서는 가격 이외에 대한 부분은 다른 긍정 반응이 높은 4성급 호텔 및 5성급 호텔에 대해서 큰 영향력을 미치는 요인이 아님을 알 수 있다.

## 4. 결론

본 논문에서는 호텔 리뷰데이터에 위상학적 데이터 분석의 맵퍼 방법론을 적용하여 감성 키워드기반 호텔 관계망을 구축한다. 맵퍼를 통해서 고정된 관점으로만 살펴보는 것이 아닌, 해상도 $R$ 및 중첩도 $G$ 등의 여러 방면에서 관계망을 구축한다는 점에서 다양한 분석이 가능하다. 또한, 구축된 관계망은 사용자뿐만 아니라 호텔 마케팅 및 전략 기획팀에 요약된 감성 정보들을 제공할 수 있다. 향후 연구에서는 리뷰데이터의 평점 정보를 활용하여 평점에 영향을 미치는 감성 키워드 등 평점에 따른 호텔의 관계망들을 살펴보고자 한다.

## References


Guo, W. and Banerjee, A. G. (2017). Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs, *Journal of Manufacturing Systems*, **43**, 225–234.

Ha, H. J., Kim, G. N., and Lee, K. W. (2013). A study on analysis of affective words in movie reviews and the situation of watching movies, *Design Convergence Study*, **12**, 17–32.

Jung, M. J., Shin, S. M., and Choi, Y. S. (2019). Creation and clustering of proximity data for text data analysis, *The Korean Journal of Applied Statistics*, **32**, 451–462.

Kruskal, J. (1964). Nonmetric multidimensional scaling: a numerical method, *Psychometrika*, **29**, 115–129.

Maaten, L. V. D. and Hinton, G. (2008). Visualizing data using t-SNE, *Journal of Machine Learning Research*, **9**, 2579–2605.

Singh, G., Memoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3D object recognition, *SPBG*.

Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology, *Discrete and Computational Geometry*, **33**, 249–274.

# Topological Data Analysis 기법을 활용한 호텔 리뷰데이터의 감성 키워드 기반 호텔 관계망 구축

전예슬[a,1] · 김정재[b]

[a]연세대학교 응용통계학과, [b]다음소프트 인공지능랩

(2019년 11월 21일 접수, 2019년 12월 27일 수정, 2020년 1월 2일 채택)

**요 약**

호텔 리뷰 데이터에는 소비를 이끈 구매 요인, 호텔에 대한 장점 및 단점 등 다양한 정보를 추출할 수 있다. 특히, 리뷰 데이터의 감성 키워드는 소비자들이 호텔에 관해 이야기하고 있는 평가 및 반응 등의 주요 내용을 파악하는 데 도움을 준다. 하지만 많은 양의 리뷰 데이터를 소비자가 직접 살펴보기에는 효율성이 떨어진다. 이를 위해 리뷰 데이터를 요약하는 기술이 요구된다. 본 연구에서는 기존의 감성 키워드 관계망을 구축하는 연구에 더 나아가, 이와 관련된 호텔에 대한 정보까지 동시에 제공하고자 한다. 이를 위해 호텔 도메인에 적합한 감성 키워드 사전을 구축하고, 이를 바탕으로 위상학적 데이터 분석 기반의 맵퍼(topological data analysis based mapper)를 통해서 감성 키워드 기반의 호텔 관계망을 구축한다. 구축된 관계망을 통해 유사한 감성을 기반으로 연결된 호텔들을 살펴볼 수 있으며 동시에, 호텔에 대한 감성 정보도 파악할 수 있다. 이러한 리뷰 요약 정보는 사용자들에게 호텔들에 대한 요약된 감성 평가를 제공하며, 호텔 마케팅 및 전략 기획팀에 분석 대상에 대한 소비자들의 인식을 파악할 수 있도록 돕는다.

주요용어: topological data analysis, mapper, sentiment analysis, sentiment keywords association-based hotel network

[1]교신저자: (03722) 서울시 서대문구 연세로 50, 연세대학교 응용통계학과. E-mail: jeon9677@yonsei.ac.kr