

# 오피니언 마이닝을 통한 브랜드 클러스터링: 자동차 산업 사례연구

황현석  
한림대학교 경영학과

## Clustering Corporate Brands based on Opinion Mining: A Case Study of the Automobile Industry

Hyun-Seok Hwang

Dept. of Business Administration, Hallym University

**요약** 인터넷의 등장으로 고객들은 자신의 의견이나 생각을 인터넷 공간에서 다른 사람들과 공유하기 시작하였다. 이에 따라 기업은 인터넷에서 수집된 데이터를 이용하여 기업에 활용할 수 있는 유의미한 결과를 찾으려는 노력을 하고 있다. 과거 설문조사를 기반으로 고객의 브랜드에 대한 태도나 만족도, 충성도 등을 분석하던 방식에서 소셜 네트워크 서비스(Social Network Service) 등에서 추출된 빅데이터를 이용하여 분석하려는 시도가 이루어지고 있다. 본 연구에서는 SNS에서 수집된 결과를 활용하여 브랜드간의 군집을 발견하는 프레임워크를 제시하고자 한다. 또한 제시된 프레임워크의 실무적용 가능성을 살펴보기 위해 자동차 산업에 대한 사례연구를 수행하였다. 두 개의 브랜드 이름이 소셜 미디어에서 동시에 언급되는 빈도가 높을수록 고객이 두 브랜드를 유사하게 인식한다는 가정 하에 자동차 브랜드 사이의 유사성을 측정하고 거리의 개념으로 변환한 후 다차원 척도법을 이용하여 3차원 상에 표시하였다. 또한 자동차 브랜드에 대한 고객의 인식을 파악하기 위해 유사한 브랜드간의 군집을 도출하고 각 군집을 특징을 기술하였다. 아울러 연구의 한계점과 향후 연구방향을 제시하였다.

**Abstract** Since the Internet provides a way of expressing and sharing Internet users' mindsets, corporate marketers want to acquire measurable and actionable insights from web data. In the past, companies used to analyze the attitude, satisfaction, and loyalty of consumers toward their brands using survey data, whereas nowadays this is done using the big data extracted from Social Network Services. In this study, we propose a framework for clustering brand names using the social metrics gathered on social media. We also conduct a case study of the automobile industry to verify the feasibility of the proposed framework. We calculate the brand name distance for each pair of brand names based on the total number of times that they are mentioned together. These distances are used to project the brand name onto a 3-dimensional space using multidimensional scaling. After the projection, we found the clusters of brand names and identified the characteristics of each cluster. Furthermore, we concluded this paper with a discussion of the limitations and future directions of this research.

**Keywords** : Automobile Industry, Brand clustering, Multidimensional Scaling, Opinion Mining, Social Media

### 1. 서론

지난 수년간 기업들은 기업의 브랜드 가치를 증대시

키기 위해 많은 노력을 시도해왔다. 특히 브랜드 포지셔닝을 위해 경쟁 브랜드와 비교한 자사의 브랜드 이미지를 분석하였으며 고객관계관리(CRM: Customer Relationship

본 연구는 한림대학교 학술연구비 지원(HRF-201501-008)에 의해 연구되었음.

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 방송통신정책연구센터(CPRC)지원사업의 연구결과로 수행되었음. (IITP-2016-R08801610080001002)

\*Corresponding Author : Hyun-Seok Hwang(Hallym Univ.)

Tel: +82-33-248-1835 email: hshwang@hallym.ac.kr

Received July 29, 2016

Revised (1st September 19, 2016, 2nd October 11, 2016)

Accepted November 10, 2016

Published November 30, 2016

Management)를 통해 고객 수익성(Customer Profitability)을 평가하고 고객 세그먼트에 따른 차별화된 고객관계관리를 전개해 오고 있다. 전통적인 고객관계관리 채널인 DM(Direct Mail)을 비롯하여 e-Mail, SMS(Short Message Service) 등의 통신망을 이용한 채널이 널리 사용되었는데 최근 몇 년 동안 페이스북(Facebook), 트위터(Twitter), 링크드인(LinkedIn)을 포함한 소셜 네트워크 서비스(Social Network Service)의 확산으로 고객관계를 위한 채널이 확장되고 있다. 기업 주도의 마케팅 패러다임에서는 대중 매체를 이용한 무차별적인 마케팅을 전개하였으나 점차 타겟그룹을 정하거나 1대1 맞춤형 마케팅 방식으로 발전하였다. 최근에는 기업주도 마케팅 이외에도 소셜 미디어 사용자가 포스팅한 글을 통해 제품이나 서비스에 대한 평가가 확산되어 구전 마케팅(Viral marketing)으로 이어지는 경우가 증가하고 있다. 과거에는 아마존(<http://www.amazon.com>)과 같은 전자상거래 사이트의 구매 후기가 기업의 상품 판매 페이지 내에서 공유되어졌다면 지금은 소셜 네트워크 서비스 상에서 자발적인 구매 후기나 브랜드에 대한 호감, 서비스 이용 경험 등에 대한 게시물이 공유되고 있다. 소셜 네트워크 서비스는 전통적인 채널과 달리 정보를 전달하는 방식이 다대다 형태이고 네트워크에 연결된 다수의 사용자에게 기하급수적으로 전달되는 특징이 있으며 네트워크로 연결된 사람과의 관계에 따라 신뢰할 수 있는 정보로 인식되는 경향이 있어 입소문 마케팅에 적합한 도구로 그 가능성을 인정받고 있다.

따라서 기업은 고객의 객관적이며 솔직한 평가를 설문지나 인터뷰에서 얻기보다 소셜 미디어에 남겨진 글을 통해 얻을 수 있다고 판단하고 소셜 미디어의 의견을 분석하는 오피니언 마이닝(Opinion mining)을 시도하고 있다.

기업 측면에서는 설문지나 인터뷰를 통한 방법보다 오피니언 마이닝을 통해 얻을 수 있는 장점이 많은데 i) 신속하게 고객의 소리를 들을 수 있다는 점, ii) 수집된 샘플의 수가 설문지나 인터뷰에 비해 충분이 크다는 점, iii) 설문지에서 수집하기 힘든 솔직한 표현을 수집할 수 있다는 점, iv) 텍스트 마이닝을 통해 감성적 표현을 분석할 수 있다는 점 등이 있다. 이와 동시에 단점 또한 가지고 있는데 대표적인 단점은 i) 소셜미디어의 특성상 응답자의 인구 통계학적인 분포를 파악하기 힘들고, ii) 텍스트 마이닝을 통해 감성 키워드(Sentiment words)를

분석할 수 있으나 문맥(Context)을 분석하는데는 아직 한계가 존재하며, iii) 기존의 통계적인 분석 기법을 적용하기 어렵다는 점이다.

본 연구에서는 아직은 학문적으로는 초기 단계에 있는 오피니언 마이닝을 마케팅 분야에 활용하는 프레임워크를 제시하고자 한다. 연구의 대상은 기업의 브랜드 이름으로 정하고 트위터나 마이크로 블로그(MicroBlog)에 언급된 브랜드 이름을 분석하여 서로 유사하다고 인식되는 정도를 도출하고 브랜드간 유사도를 거리의 개념으로 변환 후 다차원 척도법을 활용하여 3차원 공간에 프로젝션하고자 한다. 프로젝션의 브랜드들을 적절한 수의 군집으로 나누는 군집분석을 실시하고 각 군집이 갖는 특징을 분석하고자 한다. 제시된 프레임워크의 실무 적용 가능성을 확인하기 위해 자동차 산업의 브랜드 이름 18개를 정하여 사례연구를 수행하였다.

## 2. 관련 연구

브랜드와 관련된 연구는 많이 수행되어 왔으나 소셜 미디어 데이터를 이용한 연구는 최근에 활발히 연구되기 시작한 분야이다. 특히 선행연구들은 본 연구의 주제와 관련된 브랜드를 세분화(Segmentation)하여 살펴보는 연구보다 시장을 세분화(Market segmentation)하고 세분화된 시장에 맞는 브랜드를 포지셔닝하는 연구가 대다수를 이루고 있다. 이와 같은 접근법은 브랜드를 가진 회사가 현재의 세분화된 시장 중 목표시장을 정하여 포지셔닝을 하기에는 적합하다. 그러나 포지셔닝 결과가 소비자 관점에서 브랜드를 어떻게 인식하는가는 전혀 반영이 되어 있지 않는다는 단점을 가지고 있다. 따라서 다수의 소비자가 브랜드들을 어떻게 마음속에 포지셔닝하고 있는지를 이해하기 위해 브랜드와 관련된 소셜 미디어 데이터와 브랜드 관련 감성분석 관련 연구 및 브랜드 군집화와 관련된 연구를 살펴보기로 한다.

관련 연구는 크게 <Table 1>과 같이 정리될 수 있다. 첫 번째 연구방향은 단일 브랜드와 관련된 연구로 주로 단일 브랜드의 시장 내 포지셔닝과 가치측정을 위한 연구가 주를 이루고 있다. 두 번째 연구방향은 다수의 브랜드를 종합하여 연구하는 것으로 브랜드를 비교하거나 브랜드간 상대적인 위치를 정하는 것으로 주요 목적으로 하고 있다.

**Table 1.** Classification of Brand researches with sentiment analysis

Analyzing Focus	Characteristics	Approach	References
Individual Brand Analysis	Analyzing Single Brand for positioning or valuation	Supervised Learning	[1-3]
		Sentiment Polarizing	[4-10]
		Combining	[11-17]
Aggregate Brand Analysis	Analyzing Multiple Brands for perceiving relative position of each brand	Comparing	[18-19]
		Clustering	[20-27]

첫 번째 연구는 다시 3가지 방향으로 나눌 수 있는데 단일 브랜드에 대해 i) 감독학습(Supervised learning)이라는 기계학습 방법 통해 브랜드에 대한 평가를 예측하는 접근법, ii) 브랜드와 관련된 글에서 감성단어(Sentiment word)를 추출하고 긍정/중립/부정과 같은 극성(Polarity) 단어의 빈도나 가중합을 통해 브랜드에 대한 사용자의 감성을 판단하는 접근법, iii) 앞선 2가지의 방법을 혼합하여 브랜드를 분석하는 접근법이 있다.

두 번째 연구는 크게 2가지로 구분될 수 있는데 i) 여러 브랜드들의 특징을 복수 개의 변수를 통해 나타내고 브랜드별 변수값을 비교함으로써 각 브랜드별 특징을 도출하는 연구, ii) 여러 브랜드를 그룹화하고 그룹화된 브랜드에서 다시 특징을 도출하는 연구가 시도되었다.

### 2.1 감독학습에 의한 단일브랜드 연구

감독학습은 종속변수와 종속변수에 영향을 주는 다수의 독립변수들 간의 관계를 모형을 통해 예측하는 기법이다. 본 연구와 관련된 감독학습 방향은 브랜드 관련 문서의 감성단어나 키워드를 이용하여 문서의 극성(Polarity)을 예측하는 것이다. Pang, Lee & Vaithyanathan는 영화 감상평 데이터로 3가지 기계학습 방법 (Naive Bayes, Maximum entropy classification, Support vector machine)을 적용하여 주제분류(Topic classification)과 감성분류(Sentiment classification)의 정확성을 사람의 분류성능과 비교하였다[1]. 주제분류에서는 기계학습이 크게 우수한 성능을 보였으나 주제분류에서는 사람의 성능보다 크게 우수한 성능을 보여 주지는 못하였다.

Wang의 연구에서는 스마트폰 제품에 대한 전문가와 일반인의 리뷰를 분석하여 리뷰주제에 대한 감성분석을

할 수 있는 확률적인 모델을 제시하였다[2].

Shi & Li는 호텔 이용자의 리뷰를 대상으로 Support vector machine을 이용하여 감성의 극성을 분류하는 분석을 제안하였으며, 그 결과 TF(Term Frequency) 보다 TF-IDF(Term Frequency - Inverse Document Frequency)가 극성을 예측하는데 더 유용함을 밝혔다[3].

### 2.2 감성분석에 의한 극성도출에 관한 연구

감성분석에 의한 극성도출은 특정 브랜드와 관련된 문서에서 감성단어를 추출하고 감성단어들의 극성별 빈도나 가중합 등을 이용하여 문서 전체의 극성을 정하는 연구방법이다[4-6].

Bifet, Holmes & Pfahringer는 실시간으로 트위터에 게시되는 멘션을 분석하여 감성분석을 가능하게 하는 시스템을 제시하였으며 2010년 토요타 자동차에서 발생한 대규모 리콜 사태에 대한 감성분석을 예로 제시하여 브랜드에 대한 감성분석 가능성을 제시하였다[7].

de Vries, Gensler & Leeflang은 소셜 미디어에 개설된 특정 브랜드에 대한 팬 페이지 11곳에 게시되는 글과 브랜드의 대중성(popularity)을 분석하였다[8]. 연구결과 브랜드 팬 페이지 상단에 게시된 특정 브랜드에 대한 글은 대중성을 향상시키는데 긍정적인 영향을 주며 게시글이 활기차고 상호작용성을 가지고 있는 경우 ‘좋아요’를 누르는 경향이 있었으며 게시글에 대한 긍정적인 댓글이 많을수록 ‘좋아요’를 누르는 수 또한 증가함을 발견하였다.

Tedeschi & Benedetto는 브랜드에 대한 고객의 감성을 실시간으로 모니터링하기 위한 클라우드 기반의 시스템을 제안하고 Server 디자인, 서버 내 데이터베이스 스키마, 감성분석에서 극성판단 기준을 제시하였다[9].

Adeborna는 문서에서 주제분석 알고리즘을 제시하고 항공사 서비스 등급을 판단하는 사례연구를 수행하였다[10].

### 2.3 하이브리드 방법에 의한 감성분석 연구

하이브리드 방법에 의한 감성분석 연구는 감독학습과 극성분석을 포함하여 다양한 연구방법을 결합한 후 브랜드에 대한 소비자의 감성을 파악하려는 시도가 주를 이루고 있다.

Khan Bashir & Qamar는 트위터에서 수집된 데이터를 이용하여 감성분석을 실시하였다[11]. 연구에서는

Enhanced Emoticon Classifier, Improved Polarity Classifier, SentiWordNet Classifier 등을 결합(hybrid)한 기법을 사용하여, 추출된 감성언어를 긍정, 부정, 중립으로 분류하는 정확성을 향상시키고자 하였다.

Ghiassi, Skinner & Zimbra는 단일브랜드를 평가한 게시글을 감성분석하기 위해 복잡성을 감소시키는 방법으로 n-gram을 사용하였으며 트위터에서 자주 사용되는 어휘만을 골라 어휘사전을 따로 정의하여 정확성을 향상시켰다[12].

Lu & Tsou의 연구, Kahn, Atique, Thakare의 연구 및 Zhang, Ghosh, Dekhil, Hsu & Liu의 연구에서는 감독학습에서 사용되는 데이터의 어휘사전을 정의하고, 일반적인 감독학습을 통해 감성에측을 수행하는 모델을 훈련시켰으며, 마지막으로 감독학습 모델과 구축된 감성어휘사전을 통한 감성분석 모델을 혼합하여 새로운 분류기(Classifier)를 생성하였다[13-15].

Filho, Avanco, Nunes & Pardo의 연구에서는 트위터 메시지 분석을 위한 2013년의 연구를 발전시켜 3개의 분류기(규칙기반 분류기, 어휘기반 분류기, 기계학습기반 분류기)를 혼합한 형태를 제안하였으며 이전 연구보다 약9% 향상된 결과를 확인하였다[16].

Mostafa의 연구에서는 동시 언급된 트위터의 어휘들에 대해 다차원척도법을 적용하여 2차원에 맵핑한 후 어휘들의 관계를 살펴보았으며 5개의 브랜드들을 개별적으로 감성어휘의 극성과 빈도를 통해 개별 브랜드에 대한 소비자의 감성을 파악하였다[17].

#### 2.4 복수 브랜드의 인식차이 연구

Arora, Kin & Neville는 5개의 유명 스마트폰 브랜드에 대한 소비자의 인식을 비교하기 위해 트위터 데이터에 나타나 소비자의 감성을 분석하였다[18]. 각 브랜드에 대한 개별적인 감성 점수를 도출하고 이를 지역별, 기간별로 비교하였으며 배터리 성능에 대한 감성과 스마트폰 OS별에 대한 감성도 비교하였다[18].

Lihui & Tianshu는 지역 브랜드에 대한 가치평가 결과를 집적효과(aggregation effect)와 확산효과(diffusion effect)를 통해 설명하고자 하였다[19].

#### 2.5 브랜드 군집화에 대한 연구

브랜드 군집화에 대한 연구는 여러 브랜드의 특성을 파악하고 파악된 특성에 따라 유사 브랜드를 서로 군집

화하는 연구를 말한다.

Mazanec은 9개 여행사를 평가할 수 있는 여러 변수를 정의하고 여행사별로 변수값을 측정된 후에 자기조직화지도(SOM: Self-Organizing Map)을 이용하여 2가지 차원(Reliability vs. Youth)으로 축소시켜 맵핑하였다[20].

Jun & Park (2013)의 연구에서는 구글이나 네이버가 제공하는 소비자의 웹검색 트래픽을 활용해서 소비자가 생각하는 제품 포지션을 가시화할 수 있는 방법을 제안하였다[21]. 브랜드 간의 관계를 확인하기 위해, 동시 검색 트래픽 정보를 활용하여 네트워크 모델링의 방법을 사용한 시스템을 제안하고 있으며, 이를 통해 소비자들이 제품 간의 유사성을 어떻게 인지하고 형성하며, 새로운 혁신 제품 카테고리 내에서 제품 브랜드들이 소비자의 마음 속에서 어떻게 자리 잡고 있는지의 브랜드 포지셔닝을 확인할 수 있는 방법론을 제안하였으며 태블릿 PC 분야에 제안된 방법론을 적용하여 검증하고자 하였다.

G-R Kim의 연구에서는 5개(스타벅스, 커피빈, 할리스, 엔제리너스, 탐앤탐스) 커피전문점 이용고객을 대상으로 시장을 세분화 하는데 있어서 추구편익(benefits sought) 변수를 중심으로 시장을 세분화하고 포지셔닝 맵(positioning map)을 이용하여 각각의 커피브랜드의 상대적인 위치와 경쟁관계를 파악하고자 하였다[22]. 커피전문점의 평가에 사용되는 항목을 요인분석하고 도출된 요인 값을 이용하여 군집분석을 실시하였다.

Cho의 연구에서는 10개의 해외리조트 브랜드간의 포지셔닝을 위해 ‘리조트 안전성’, ‘리조트 접근성’, ‘리조트 시설’, ‘고객 맞춤형 서비스’, ‘리조트 차별성’의 5개의 요인을 도출하고 다차원 척도법을 이용하여 결과를 도식화하여 표현하였다[23].

Lee의 연구에서는 Cho의 연구와 유사하게 11개 수입 와인사의 브랜드를 12개의 변수를 통해 평가하고 이를 다차원 척도법을 이용하여 분석하였다[24].

Lim, Yoo & Ko,의 연구에서는 고가-저가 구매자의 세분화를 위해 고가제품과 저가제품을 구매하는 심리적인 요인을 분석하였다[25]. 500명의 소비자를 대상으로 확인적 요인분석과 퍼지군집화 기법을 적용하여 5개의 고객 군집(Deal hunters, Sale-prone shoppers, Active luxe-bargain shoppers, Royal shoppers)을 도출하였으며 파악된 군집의 특성을 분석하기 위해 분산분석(ANOVA)을 사용하였다.

Yoon & Yim의 연구에서는 포지셔닝 분석을 통해 소

설 네트워크 서비스 시장의 포지셔닝 맵을 제시하고, 소셜 네트워크 서비스 속성들이 이용자들의 서비스 선호도에 미치는 영향을 분석하였다[26]. 정보탐색, 상호작용, 정체성 확인, 편리성, 습관성, 여가/오락 등 6개 소셜 네트워크 서비스 속성별로 싸이월드, 트위터, 페이스북, 미투데이, 카카오톡의 유사성을 분석하였다. 또한 사용자 인식이 소셜 네트워크 서비스 선호도의 속성들에 대한 미치는 영향을 확인하기 위하여 다중회귀분석을 실시하였다.

Kim의 연구에서는 웹사이트 평가요소를 도출하여 개별 웹사이트에 대한 평가를 실시한 후 그 결과를 이용하여 군집화를 시도하였다[27].

### 3. 연구 프레임워크

본 연구에서는 소셜 미디어 데이터를 활용하여 오피니언 마이닝을 실시한 후 브랜드 클러스터링을 수행하는 연구 프레임워크를 다음과 같이 제시하고자 한다. 연구 프레임워크는 <Figure 1>과 같이 6개의 단계로 이루어져 있다.

첫 번째 단계에서는 브랜드 이름을 선택하는 것으로 연구의 주제를 정하는 단계에 해당된다. 연구주제로서 적절한 브랜드는 i) 소셜 미디어에서 여러 차례 언급될 수 있는 정도로 대중에게 알려져 있어야 하며, 대중에게 알려진 브랜드라고 하더라도 ii) 대중에게 제품이나 서비스가 전달되어 사용경험을 제공할 수 있는 브랜드이어야 한다. 또한 iii) 적절한 수의 경쟁 브랜드를 갖추고 있어야 한다. 브랜드의 개수가 너무 많은 경우 포지셔닝을 파악하기가 힘들고 분석된 결과를 해석하기도 힘든 문제점이 있고, 반대로 개수가 적은 경우는 실무적용 가능성이 낮고 군집에 속한 브랜드가 수가 작아서 군집 내의 브랜드 속성의 분산이 커서 소수의 브랜드 속성에 의해 결과가 바뀌게 되는 문제점이 있다. 본 연구에서 제시하는 적절한 브랜드 개수는 15개~30개 정도로 분석의 결과로 도출되는 군집의 개수가 적절하며 도출된 군집의 특성을 분석하기에도 적절한 수로 판단된다. 이러한 요건을 모두 갖춘 브랜드로는 자동차 산업의 배기량에 따른 차종이나 화장품 산업의 브랜드 등을 고려해 볼 수 있다.

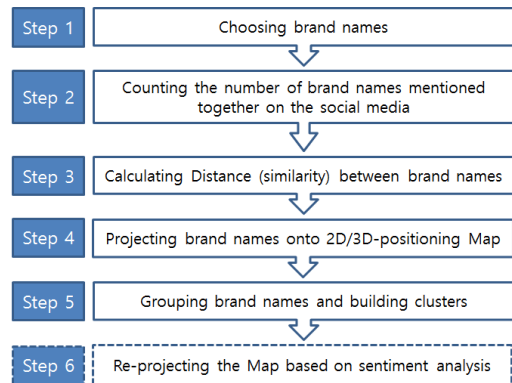


Fig. 1. Research framework

두 번째 단계는 이전 단계에서 정해진 브랜드 이름을 언급한 소셜 미디어의 게시 글을 검색하여 2x2 행렬을 생성시키는 단계이다. 행렬의 행과 열은 각각 첫 번째 단계에서 정한 브랜드의 이름들로 구성된다. 일반적인 설문분석과 상이하게 오피니언 마이닝의 경우 수집 (Crawling) 소프트웨어를 사용하여 많은 양의 문서에서 검색어에 해당되는 내용을 찾고 저장하게 되는데 본 연구에서는 두 개 이상의 브랜드 이름이 하나의 글에 언급된 수를 카운트 한다. 본 연구에서는 두 번째 단계의 결과로 생성된 행렬을 *Co-mention Matrix*로 명명하기로 한다. *Co-mention Matrix C*는 다음과 같은 성분( $C_{ij}$ )으로 구성된다.

행렬의 성분 중 분자를 먼저 살펴보면 대각 행렬은 0으로 정의하며 대각 행렬이 아닌 경우  $i$ 번째 브랜드 이름과  $j$ 번째 브랜드 이름이 같이 언급된 글의 수( $n$ )로 정의하기로 한다. 만일 같이 언급된 글이 없다면 1로 정의한다. 예를 들어 트위터를 검색하는 경우 검색 소프트웨어에 따라 반환되는 검색결과 수에 제한이 있어서 검색어에 따라 추출된 글에 브랜드이름 두 개가 같이 언급되지 않는 글도 존재한다. 이 경우 1로 정의하기로 한다. 그리고 특정 브랜드에 대한 언급이 많은 경우 해당 브랜드와의 거리가 지나치게 작아지는 문제점을 없애기 위해 데이터를 정규화한다. 정규화는 집계의 결과에 각 행 합과 열의 합을 더한 후 나누어 계산한다.

$$C_{ij} = \begin{cases} 0 & , \quad where \ i = j \\ \frac{1}{\sum_i C_{ij} + \sum_j C_{ij}} & , \quad where \ i \neq j, \ n = 0 \\ \frac{n}{\sum_i C_{ij} + \sum_j C_{ij}} & , \quad where \ i \neq j, \ n > 0 \end{cases}$$

세 번째 단계는 Co-mention Matrix로 부터 브랜드 사이의 거리를 계산하는 단계이다. 브랜드 간 거리의 개념은 향후 브랜드 이름을 적절한 차원의 공간에 프로젝션시키기 위해 필요하며 Co-mention Matrix로부터  $i$ 번째 브랜드와  $j$ 번째 브랜드 간 거리( $d_{ij}$ )를 계산하는 기본적인 가정은 다음과 같다.

**Co-mention Matrix에서 언급된 횟수가 많을수록 고객은 두 브랜드간 거리를 가깝다고 인식한다.**

Co-mention Matrix에서 파생된 거리행렬  $D$ 는 다음과 같이 정의 할 수 있다.

$$d_{ij} = \begin{cases} \frac{1}{C_{ij}}, & \text{where } i \neq j \\ 0, & \text{where } i = j \end{cases}$$

$D = \{ d_{ij} \}$ ,  $d_{ij}$ 는  $i$ 번째 브랜드와  $j$ 번째 브랜드의 인지된 거리

거리 계산 공식의 의미는 대각행렬의 요소는 같은 브랜드 이름 간 거리가므로 0으로 정하고 대각행렬이 아닌 요소들의 값은 Co-mention Matrix 값의 역수로 변환하는 것이다.

네 번째 단계에서는 다차원 척도법을 이용하여 브랜드 이름을 2차원이나 3차원 공간에 프로젝션 시키는 단계이다. 다차원 척도법은 여러 개체에 대한 평가를 이용하여 개체 사이의 공간적 거리를 정하는 방법이다. 공간적 거리는 유사성의 반대되는 개념으로 공간적 거리가 정해지면 위의 공간상에 개체를 위치시켜 시각적으로 표현하게 된다. 이렇게 표현된 그림을 지각도(Perceptual map) 또는 공간도(Spatial map) 라고 부른다. 지각도에 사용된 축은 각 개체의 거리를 정하는 새로운 기준이며 개체가 가지고 있는 변수들로부터 새롭게 추출된 값이다. 이 축을 차원 (Dimension) 이라고 부른다. 차원은 크게 두 가지 종류로 나눌 수 있는데 ‘크기’, ‘길이’와 같은 객관적이고 정량화된 측정방법이 존재하여 평가자에 따른 차이가 생길 수 없는 객관적 차원 (Objective dimension) 이 있고 ‘아름답다’, ‘만족스럽다’와 같이 개인의 견해나 판단이 개입될 수 있는 주관적 차원 (Subjective dimension) 이 있다. 다차원 척도법은 객관적 차원과 주관적 차원 모두를 가질 수 있다.

다섯 번째 단계에서는 브랜드들의 군집을 파악하는 단계로 프로젝션된 브랜드의 2차원과 3차원의 공간 좌

표를 이용하여 군집분석을 실시한다. 군집분석은 주어진 데이터를 일정한 기준에 의해 여러 개의 배타적인 집단으로 그룹화하는 방법이다. 본 단계에서는 적절한 군집간의 수를 정하기 위해 군집의 수를 변화시키면서 군집내 거리를 최소화하면서 동시에 군집간 거리를 최대화하는 군집 수를 찾는 방법을 택하였다.

마지막 단계는 프로젝션의 결과에 대한 신뢰성을 향상시키기 위해 감성분석(Sentiment analysis)를 추가로 실시하는 단계이다. 감성분석은 텍스트 마이닝을 이용하여 특정 키워드를 포함한 문서 내에 감정을 표현하는 단어들을 추가로 추출하여 특정 주제에 대한 문서의 감성적 경향을 분석하는 기법으로 긍정어, 부정어, 중립어로나누어 각각의 빈도를 측정한다. 두 개 이상의 브랜드 이름이 동시에 언급된 글이나 멘션에서 감성분석을 실시하여 긍정어와 부정어의 개수를 통해 두 브랜드의 거리를 재평가 하거나 단어들이 가진 의미망(Semantic network)에서 의미상의 거리를 측정하여 이를 거리 개념을 대체할 수도 있다. 예를 들어 두 브랜드 이름이 동시에 언급된 빈도가 많은 경우라도 부정적인 단어가 동시에 많다면 두 브랜드 사이의 거리를 더 멀게 평가하는 방법이 사용될 수 있다.

#### 4. 사례 연구

본 연구에서 제시된 프레임워크의 실무적용 가능성을 살펴보기 위해 사례연구를 수행하였다. 사례연구의 대상은 자동차 산업으로 하였으며 자동차의 이름을 브랜드 이름으로 설정하였다. 자동차 브랜드는 3장에서 살펴본 바와 같이 브랜드 이름을 정하기 용이하고 마이크로 블로그나 트위터에 그 평가가 언급되는 경우가 많으며 15개 이상의 수의 브랜드 명을 가지고 있어서 적절한 분석의 대상으로 판단하였다. 자동차 브랜드 이름이 언급된 마이크로 블로그의 글과 트위터의 글을 30일 동안 수집하였으며 Co-mention matrix를 생성하기 위해 특정 브랜드 이름이 언급된 글이나 멘션에 다른 브랜드 이름이 언급된 경우가 있는지를 조사하여 글이나 트위터 멘션의 수를 카운트하였다. <Table 2>은 Co-mention matrix C의 일부분을 제시한 것이다. 표에서도 알 수 있듯이 대각선의 값은 0으로 설정되었으며 대각 행렬을 중심으로 대칭의 형태를 보이며 데이터의 분산 값이 큰 것을 알 수 있다.

Table 2. Co-mention matrix C

Brand name	i40	K3	K5	K7	Grandeur	...	Sportage
i40	0.00000	0.00009	0.02159	0.03522	0.05742	...	0.02224
K3	0.00009	0.00000	0.03873	0.04594	0.00013	...	0.03852
K5	0.02159	0.03873	0.00000	0.04580	0.00332	...	0.04907
K7	0.03522	0.04594	0.04580	0.00000	0.00385	...	0.02871
Grandeur	0.05742	0.00013	0.00332	0.00385	0.00000	...	0.00205
...	...	...	...	...	...	...	...
Sportage	0.02224	0.03852	0.04907	0.02871	0.00205	...	0.00000

Co-mention matrix에 3장에서 제시된 거리 계산 공식을 적용하여 두 브랜드 이름간의 거리를 계산한 결과를 <Table 3>에 제시하였다.

Table 3. Distance matrix D

Brand name	i40	K3	K5	K7	Grandeur	...	Sportage
i40	0.0	10752.0	46.3	28.4	17.4	...	9667.0
K3	10752.0	0.0	25.8	21.8	7840.0	...	10941.0
K5	46.3	25.8	0.0	21.8	301.6	...	53.0
K7	28.4	21.8	21.8	0.0	259.4	...	30.7
Grandeur	17.4	7840.0	301.6	259.4	0.0	...	6755.0
...	...	...	...	...	...	...	...
Sportage	9667.0	10941.0	53.0	30.7	6755.0	...	0.0

본 연구에서는 브랜드간 거리로 계산된 행렬을 이용하여 3차원 공간에 브랜드 이름을 프로젝션 시키기 위한 방법으로 다차원척도법을 사용하였다. <Figure 2>는 3차원 공간에 18개의 브랜드를 프로젝션 시킨 결과이다.

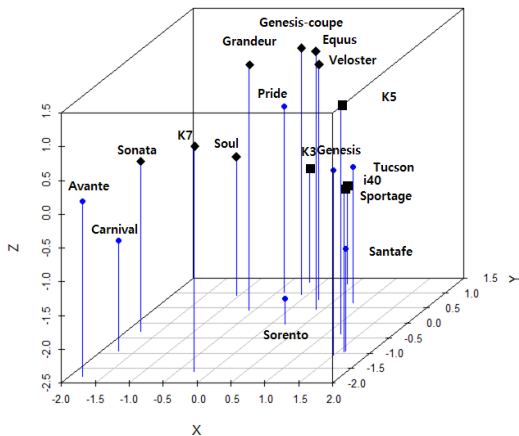


Fig. 2. Projected results of brand names

다차원 척도법에서 도출된 3차원 좌표와 결과를 살펴 보면 <Table 4>과 같다.

Table 4. Coordinate value of brand names

No	Brand name	X	Y	Z	Cluster
1	i40	.8140	.6780	-.4860	1(■)
2	K3	-.2020	1.3499	-.7999	1(■)
3	K5	1.2093	-.3536	.8898	1(■)
4	K7	-.2586	-1.6092	.8296	3(◆)
5	Grandeur	-.5776	.4188	1.1346	3(◆)
6	Veloster	.2506	.7916	.9664	3(◆)
7	Sportage	1.5989	-.9620	-.0919	1(■)
8	Santafe	1.5965	-.9192	-.9957	2(●)
9	Sonata	-1.7885	-.2818	.0138	3(◆)
10	Sorento	.2142	-.0437	-2.1168	2(●)
11	Soul	-1.0503	.9351	-.4515	3(◆)
12	Avante	-1.8246	-1.7659	.0893	2(●)
13	Equus	.3970	.4472	1.3230	3(◆)
14	Genesis	1.4874	-1.0582	.2355	2(●)
15	Genesis coupe	-.0923	.9463	1.1422	3(◆)
16	Carnival	-1.7534	-.9258	-.8659	2(●)
17	Tucson	.3774	1.3086	-1.0614	2(●)
18	Pride	-.3981	1.0439	.2449	2(●)

1Stress = .18, RSQ = .82

다차원 척도법의 Stress 값은 .18로 좋지 않은 편이나 회귀분석의 R<sup>2</sup>와 유사한 RSQ 값은 .82로 높게 산출되었다.

다음 단계는 비계층 군집분석 방법인 k-평균 군집분석 기법을 이용하여 브랜드간의 군집을 도출하는 과정이다. 군집분석의 주요 결정사항은 군집의 수를 결정하는 것으로 군집 간 거리를 최대화하면서 군집 내 거리를 최소화하는 새로운 지표  $f(n)$ 을 설정하여 <Table 5>와 같이 분석하였다. 군집의 수  $f(n)$ 를 변화시키면서 군집 간 거리(Intercluster Distance)와 군집 내 거리(Intracluster Distance)를 계산하고 이들의 비율을 계산하였는데 군집이 수가 증가할수록 군집간 거리가 군집의 수의 제곱에 비례하는 것을 보정하기 위해 군집 수에 따른 군집 간 거리 조합의 수  $n(n-1)/2$ 를 곱하여 보정하였다.

Table 5. Intercluster and Intracluster distance

Cluster (n)	Intercluster Distance (D1)	Intracluster Distance (D2)	$f(n) = D_2/D_1 \times n(n-1)/2$
2	2.33	24.87	10.66
3	6.83	19.98	1.03
4	14.43	17.18	5.04
5	25.45	12.45	20.45
6	38.84	10.52	55.39

$f(n)$ 은 작은 값일수록 적절할 수의 군집임을 나타내는데 본 연구에서는 군집의 수가 3개일 때 가장 적절한 것으로 나타나 군집의 수를 3개로 정하였다. 군집을 3개로 하여 분석하였을 때 각 브랜드가 해당되는 군집은 <Table 4>과 <Figure 3>에 각각 ■, ●, ◆로 나타내었다.

각 군집에 속한 브랜드의 특성을 살펴보면 1번째 군집(■)은 다른 군집보다 상대적으로 저렴하고 연비가 좋은 차량으로 구성되어 있으며 2번째 군집(●)은 가장 인기있는 차종으로, 대중적이며 가족들이 함께 탈 수 있는 다수의 브랜드로 구성되어 있다. 3번째 군집(◆)은 타 브랜드에 비해서 상대적으로 높은 가격의 차량으로 배기량이 크거나 개성이 강한 사람들이 선호할 수 있는 차량으로 구성되어 있다.

마지막 단계는 감성분석을 통해 거리의 개념을 더 정확하게 측정하는 단계로 브랜드 이름과 함께 감성단어의 극성을 통해 긍정의 감성단어가 많은 경우 거리를 더 가깝게, 부정의 감성단어가 많은 경우 거리를 더 멀게 보정하는 것이다. 이 단계는 본 연구의 범위를 넘어서는 것으로 판단하여 향후 연구에서 추가로 실시하기로 한다. 감성분석을 통한 거리 보정의 기본적인 공식을 제안하면 다음과 같다.

$$\hat{d}_{ij} = d_{ij} \times \frac{n_{ij}^-}{n_{ij}^+}$$

$\hat{d}_{ij}$ 은 감성분석을 고려한  $i$  와  $j$  브랜드간 거리이며  $n_{ij}^+$ 와  $n_{ij}^-$ 는 각각  $i$  와  $j$  브랜드가 언급된 글에서 분석한 긍정 감성단어 수와 부정 감성단어 수를 뜻한다.  $n_{ij}^+$ 와  $n_{ij}^-$  수가 동일한 경우 중립적인 내용으로 판단하여 거리를 보정하지 않으며 긍정 감성단어 수 대비 부정 감성단어 수의 비율만큼 거리를 증가시키는 공식이다.

## 5. 결론과 향후 연구 방향

본 연구에서는 소셜 미디어에서 언급된 브랜드 이름을 분석하여 브랜드 포지셔닝에 활용할 수 있는 프레임워크를 제시하고 자동차 산업에 적용하여 사례연구를 수행하였다. 본 연구의 결과 소셜 미디어에 언급된 빅데이터를 분석함으로써 소비자의 브랜드에 대한 인식을 찾아내고 군집을 도출할 수 있음을 알 수 있었다.

본 연구의 학문적인 함의는 브랜드 포지셔닝에 소셜 미디어 데이터의 활용 가능성과 활용 프레임워크를 제시했다는 측면이다. 감성분석만을 이용하여 브랜드에 대한 긍정/부정적인 감성을 분석하는 방법이 많이 제시되었으나 본 연구에서는 타 브랜드와의 인지된 거리를 찾아내고 군집화할 수 있는 방법을 제안하려고 하였다.

한편 실무적인 측면의 함의는 마케터에게 과거 설문이나 인터뷰로 수집된 데이터 분석과 차별화되는 소셜 미디어 데이터의 활용 가능성을 제시했다는 점이다. 기업은 자사의 브랜드 포지셔닝을 위해 다양한 기법을 활용할 수 있으나 그 출발점은 현재 시점에서 고객이 브랜드에 대해 가진 인식을 살펴보는 일이며 그 결과를 경쟁 브랜드와 시각적으로 비교할 수 있다면 그 상대적인 위치를 파악하기에 수월한 방법이 될 수 있다. 타 브랜드와의 거리를 도출하여 시각화하고 군집분석을 통해 군집을 찾아냄으로써 새로운 브랜드 포지셔닝을 찾는 데 도움을 줄 수 있을 것으로 판단된다.

본 연구의 특성상 소셜 미디어 데이터를 이용한 초기 단계의 연구로 볼 수 있으며 데이터의 수집과 분석에서 다음과 같은 한계점을 지니며 동시에 향후 연구를 통해 한계점의 보완이 필요하다.

첫 번째 한계점은 두 개의 브랜드가 동시에 언급된 횟수를 나타내는 Co-mention 행렬을 더 정확하게 측정할 필요가 있는 점이다. 현재의 30일간의 데이터를 수집하여 40000여 개의 글을 분석해 본 결과 두 브랜드 이름간의 언급이 충분하지 않은 경우가 있다. 따라서 현재보다 3~5배 정도 데이터를 수집 후 분석하여 결과의 신뢰성을 향상시키는 연구가 필요하다. 두 번째 한계점은 감성분석을 통해 파악된 긍정어와 부정어의 개수를 이용하여 두 브랜드의 거리를 재평가 하거나 단어들이 가진 강도(Intensity)나 단어 간 거리를 측정하여 현재의 거리 개념을 정교하게 측정하는 연구가 필요하다. 감성분석이 결합되는 경우 거리 측정의 정확성을 향상시키고 동시에



브랜드에 대한 긍정/부정/중립의 오피니언의 극성을 파악할 수 있는 장점을 지닐 수 있을 것으로 판단된다. 이를 위해 특정 산업의 감성단어를 보강하여 감성분석의 정확도를 높이는 일이 선행될 필요가 있다. 예를 들어 일반적인 감성분석에서 ‘크다’라는 단어는 중립적인 단어이지만 차량의 실내공간과 결합되면 ‘크다’라는 단어는 긍정적인 단어가 된다.

## References

- [1] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proc. of the ACL-02 Conference on Empirical Methods in Natural Language Processing, pp. 79-86, 2002.  
DOI: <http://dx.doi.org/10.3115/1118693.1118704>
- [2] W. Wang, "Sentiment analysis of online product reviews with semi-supervised topic sentiment mixture model", Proc. of 7th International Conference on Fuzzy Systems and Knowledge Discovery, vol. 5, pp. 2385-2389, 2010.
- [3] H.-X. Shi, X.-J. Li, "A sentiment analysis model for hotel reviews based on supervised learning", Proc. of International Conference on Machine Learning and Cybernetics, vol. 3, pp. 950-954, 2011.
- [4] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424, 2002.
- [5] M. Hu, B. Liu, "Mining and Summarizing Customer Reviews", Proc. of International conference on Knowledge discovery and data mining, pp. 168-177, 2004.  
DOI: <http://dx.doi.org/10.1145/1014052.1014073>
- [6] B. Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, pp. 168, 2012.
- [7] A. Bifet, G. Holmes, B. Pfahringer, "MOA-Tweet Reader: real-time analysis in twitter Streaming data," Lecture Notes in Computer Science, vol. 6926, pp. 46-60, 2011.  
DOI: [http://dx.doi.org/10.1007/978-3-642-24477-3\\_7](http://dx.doi.org/10.1007/978-3-642-24477-3_7)
- [8] L. de Vries, S. Gensler, P. S. H. Leeftang, "Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing", Journal of interactive marketing, vol. 26, no. 3, pp. 83-91, 2012.  
DOI: <http://dx.doi.org/10.1016/j.intmar.2012.01.003>
- [9] A. Tedeschi, F. Benedetto, "A cloud-based big data sentiment analysis application for enterprises' brand monitoring in social media streams", IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow, pp. 186-191, 2015.  
DOI: <http://dx.doi.org/10.1109/rtis.2015.7325096>
- [10] E. Adeborna, A. Funk, "An approach to sentiment analysis the case of airline quality rating", Proc. of the Pacific Asia Conference on Information Systems, pp. 363-368, 2014.
- [11] F. H. Khan, S. Bashir, U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme", Decision support systems, vol. 57, pp. 245-257, 2014.  
DOI: <http://dx.doi.org/10.1016/j.dss.2013.09.004>
- [12] M. Ghiassi, J. Skinner, D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network", Expert systems with applications, vol. 40, no. 16, pp. 6266-6282, 2013.  
DOI: <http://dx.doi.org/10.1016/j.eswa.2013.05.057>
- [13] B. Lu, B. Tsou, "Combining a large sentiment lexicon and machine for subjectivity classification", Proc. of International Conference on Machine Learning and Cybernetics, vol. 6, pp. 3311-3316, 2010.  
DOI: <http://dx.doi.org/10.1109/icmlc.2010.5580672>
- [14] A. Z. H. KHAN, M. ATIQU, V. M. THAKARE, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", International Journal of Electronics, Communication and Soft Computing Science and Engineering, Special Issue of IJECSCSE, pp. 89-91, 2015.
- [15] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis", <http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf>, 2011.
- [16] P. P. B. Filho, L. V. Avanco, M. d. G. V. Nunes, T. A. S. Pardo, "NILC\_USP: An Improved hybrid system for sentiment analysis in twitter message", Proc. of International Workshop on Semantic Evaluation, pp. 428-432, 2014.
- [17] M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments", Expert systems with applications, vol. 40, no. 10, pp. 4241-4251, 2013.  
DOI: <http://dx.doi.org/10.1016/j.eswa.2013.01.019>
- [18] D. Arora, F. L. Kin, Stephen W. Neville, "Consumers' Sentiment Analysis of Popular Phone Brands and Operating System Preference Using Twitter Data: A Feasibility Study", Proc. of International Conference on Advanced Information Networking and Applications, pp. 680-686, 2015.
- [19] S. Lihui, X. Tianshu, "On the Impact of Brand Cluster Effects of Forming of Place Brand: Take Wenzhou Cluster for Example", Proc. of International Conference on E-Business and E-Government, pp. 5351-5354, 2010.  
DOI: <http://dx.doi.org/10.1109/icee.2010.1339>
- [20] J. A. Mazanec, "Simultaneous positioning and segmentation analysis with topologically ordered feature maps: a tour operator example", Journal of retailing and consumer services vol. 6, no. 4, pp. 219-235, 1999.  
DOI: [http://dx.doi.org/10.1016/S0969-6989\(98\)00037-X](http://dx.doi.org/10.1016/S0969-6989(98)00037-X)
- [21] S.-P. Jun, D.-H. Park, "Intelligent Brand Positioning Visualization System Based on Web Search Traffic Information : Focusing on Tablet PC", Journal of intelligence and information systems, vol. 19, no. 3, pp. 93-111, 2013.  
DOI: <http://dx.doi.org/10.13088/jiis.2013.19.3.093>

- [22] G.-R. Kim, "The positionings of coffee brands and market segmentation based on perceived benefits", Master's thesis, Young Nam University, 2010.
- [23] H. S. Cho, "A Study on the positioning of Overseas Resorts based on Resort selection attributes by Multi-dimensional Scaling Method", Master's thesis, Kyunghee University, 2015.
- [24] Y.-Y. Lee, E. Yoo, J.-Y. Ko, "The Brand Positioning for Wine Importers in Korea : Focused on B2B", Journal of tourism and leisure research, vol. 24, no. 1, pp. 433-452. 2012.
- [25] C. M. Lim, R. Runyan, Y.-K. Kim, "Segmenting luxe-bargain shoppers using a fuzzy clustering method", International Journal of Retail & Distribution Management, vol. 41, no. 11/12, pp. 848-868, 2013.  
DOI: <http://dx.doi.org/10.1108/IJRDM-01-2013-0012>
- [26] J. Yoon, J. Yim, "The Market Positioning Analysis of Social Network Services", Korean journal of broadcasting, vol. 26, no. 3, pp. 416-457, 2012.
- [27] S. Kim, "A Classification of Luxury Fashion Brands' E-commerce Sites", Fashion business vol. 17, no. 6, pp. 125-140, 2013.

---

황 현 석(Hwang, Hyun Seok)

[정회원]



- 2004년 2월 : POSTECH, Management Information Systems, Ph. D.
- 2004년 3월 : 한림대학교 경영학부 교수
- 2004년 3월 : 한림 경영연구소 연구위원

<관심분야>

빅데이터, 데이터마이닝, 비디오게임