논문 투고일:2016.04.06, 논문 심사일:2016.04.06-2016.04.27, 게재 확정일:2016.04.27

온톨로지 시각화를 활용한 사용자 리뷰 분석 기반 영화 추천 시스템

Movie Recommended System base on Analysis for the User Review utilizing Ontology Visualization

주저자
문성민 (Mun, Seong Min)
아주대학교 라이프미디어 협동과정 통합디자인연구실 연구원

공동저자
김기남 (Kim, Gi Nam)
네이버 콘텐츠 비즈 개발부 연구원

최경철 (Choi, Gyeong cheol)
아주대학교 라이프미디어 협동과정 통합디자인연구실 연구원

교신저자
이경원 (Lee, Kyung Won)
아주대학교 미디어학과
kwlee@ajou.ac.kr

이 논문은 2012년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2012S1A5A2A01020132)

목차

(요약)

최근 소비자 구전정보에 대한 연구들은 소비자가 제품 구매 과정에서 다른 소비자의 구전에 의한 정보를 활용한다는 연구 결과를 시사하고 있다. 본 연구는 제품에 대한 소비자의 의견을 파악하고 활용할 수 있도록 오피니언 마이닝과 시각화를 통해 도움을 줄 수 있는 방법을 제안하고자 한다. 이를 위해 최근 들어 관람할 영화를 선택할 때 인터넷 상의 영화리뷰를 참고 하는 상황이 증가함을 고려하여 “영화” 도메인의 온톨로지를 구축하고 오피니언 마이닝을 수행하여 시각화 한 후 그 결과에 대해 논하고자 한다. 온톨로지를 구축하는 과정에서 평가요소에 대한 속성 분류뿐만 아니라 평가요소에 대한 서술어 사전을 구성하였다는 점에서 기존의 연구와 차별성이 있으며 분석 결과를 통해 이러한 방법이 오피니언 마이닝에 유효함을 증명하고자 한다. 연구를 통해 도출한 결과는 크게 세 가지로 나누어 볼 수 있다. 첫째, 본 연구에서는 기존에 구축된 온톨로지를 활용하지 않고 키워드 추출과 토픽모델링을 활용하여 영화 도메인에 대한 온톨로지를 구축하는 방법에 대해 서술하였다. 둘째, 개별 영화에 대한 시각화 분석을 시행하여 영화에 대한 관객의 종합적인 의견을 한눈에 파악할 수 있도록 하였다. 셋째, 제품에 대한 평가 결과에 따라 유사한 평가를 받은 제품끼리 군집화 되는 것을 발견하였으며 본 연구의 분석에 사용된 130개의 영화는 크게 3개의 집단으로 군집화 됨을 보였다.

(Abstract)

Recently, researches for the word of mouth(WOM) imply that consumers use WOM informations of products in their purchase process. This study suggests methods using opinion mining and visualization to understand consumers’ opinion of each goods and each markets. For this study we conduct research that includes developing domain ontology based on reviews confined to “movie” category because people who want to have watching movie refer other's movie reviews recently, and it is analyzed by opinion mining and visualization. It has differences comparing other researches as conducting attribution classification of evaluation factors and comprising verbal dictionary about evaluation factors when we conduct ontology process for analyzing. We want to prove through the result if research method will be valid. Results derived from this study can be largely divided into three. First, This research explains methods of developing domain ontology using keyword extraction and topic modeling. Second, We visualize reviews of each movie to understand overall audiences’ opinion about specific movies. Third, We find clusters that consist of products which evaluated similar assessments in accordance with the evaluation results for the product. Case study of this research largely shows three clusters containing 130 movies that are used according to audiences’opinion.

(Keyword)

Visualization, Movie review, Ontology, Opinion Mining, Case study

1. 서론

1-1. 연구배경 및 목적

웹 2.0 시대 이후로 인터넷 이용자로부터 온라인 구전정보가 생산됨에 따라 구전정보가 사회에 미치는 영향에 대한 연구도 활발하게 이루어지고 있다. 구전정보(WoM: Word of Mouth)가 주목받는 이유는 소비자의 제품 구매, 기업의 이미지 형성에 영향을 미치기 때문이다(윤영선, 2013). 소비자의 입장에게 구전정보는 제품 구매 결정에 도움이 될 수 있고, 기업의 경우, 구전의 내용이나 콘텐츠를 관리하면 물품에 대한 소비자의 반응 파악에 활용할 수 있기 때문에 인터넷 구전정보는 소비자와 기업 모두에게 유용한 자원으로 활용된다(이은영, 2008). 하지만 인터넷 구전정보를 인력으로 분석하기에는 많은 비용이 발생한다는 한계점이 존재하고 이를 해결하기 위한 방법으로 자동화된 시스템으로 구전정보를 분석하는 오피니언 마이닝이 사용된다.

감성분석(Sentiment analysis)으로도 불리는 오피니언 마이닝(Opinion mining)은 사람들이 가지고 있는 의견, 평가, 태도 그리고 그들이 사용한 제품, 서비스, 기관, 이슈, 이벤트, 토픽들에 대한 감정을 분석하는 것으로 정의된다.[1)]

오피니언 마이닝은 텍스트 마이닝의 한 분야로서 말뭉치 혹은 코퍼스(corpus)로 불리는 대규모 언어 데이터베이스를 기반으로 대상이 되는 문장을 분석하고 극성정보가 포함된 사전을 통해 작성자가 긍정적인 반응을 보였는지 부정적인 반응을 보였는지 판별한다. 국내에서는 1998년부터 시작된 국가규모의 코퍼스 구축 프로젝트인 '21세기 세종계획'이 2007년 완성되어(김시우, 2008) 이를 기반으로 오피니언 마이닝에 대한 연구가 활발하게 이루어지고 있다. 하지만 이러한 코퍼스 사전을 이용한 방법은 다음과 같은 한계점을 가지고 있다. 첫 번째로 관용어나 신조어 등 복잡한 표현을 분석하기 어렵다. 예를 들어 "스토리가 장난이 아니다." 또는 "캐스팅이 쩔어준다." 는 기존의 구문분석 방법으로는 의미를 해석하기 어렵다. 두 번째로 의미 표현을 찾기 위해 사전 전체를 탐색해야 하는 비용이 발생한다. 예를 들어 평가의 대상이 "인테리어" 인 경우에도 "맛"과 관련된 키워드인 "매콤하다", "달콤하다" 등 평가어로 등장할 확률이 전무한 평가 어휘도 탐색해야 하기 때문에 데이터 또는 사전의 크기가 증가할수록 처리시간도 증가할 수밖에 없다.

오피니언 마이닝의 가장 마지막 단계는 분석된 데이터를 시각화하는 단계이다. 그림 1 은 다음 소프트에서 서비스하는 '소셜메트릭스'로, 주제어와 관련된 연관 키워드들이 자아 연결망(Ego Network)으로 구성되어 있으며, 키워드의 속성에 따라 다른 색상으로 표현되는 것을 확인할 수 있다.[2)]

No.	연관어	분류	탐색건수
1	맛	속성	19,403
2	먹고싶다	심리	19,315
3	집	장소	17,239
4	피자	상품	15,983
5	생일	속성	14,035
6	케익	상품	13,685
7	형님	속성	12,637
8	맛있다	심리	12,612
9	고맙다	심리	12,564
10	맥주	상품	12,500

〈그림 1〉 주제어에 대한 연관키워드를 시각화한 '소셜메트릭스'

1) Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.

2) Social Metrics, http://insight.some.co.kr

이처럼 시각화는 데이터(Data), 정보(Information), 지식(Knowledge)을 사람이 인지할 수 있는 시각적인 형태로 변환하는 것으로 시각화를 적절히 사용하면 사용자의 정보 습득을 도울 수 있다.[3)]

본 연구는 위에서 제시한 배경을 바탕으로 온톨로지와 시각화를 활용한 오피니언마이닝 방법을 제안하고자 한다. 문장 분석 과정에 있어서 앞서 제시된 문제를 해결하기 위해 특징과 특징을 표현하는 표현어휘를 포함하는 온톨로지를 활용한다. 또한 기존의 방법에서는 불가능했던 복잡한 표현어휘를 분석하고 분석 대상에 따라 부분적인 평가표현을 탐색하게 함으로써 처리비용을 최소화 하는 방법을 제안하고자 한다. 이와 더불어 새롭게 제안하는 시각화 방법을 통해 효율적인 정보 습득이 이루어지도록 유도하는 것이 본 연구의 목적이다. 본 연구는 연구에 활용될 도메인으로 “영화”를 선정하였다. 영화 시장은 매년 성장하고 있으며 최근 미디어 기술의 발달로 영화 시장의 규모가 더 커지고 있다. 영화진흥위원회가 2014년에 발표한 영화 산업 결산 원고에 따르면 2014년 영화 시장의 매출액은 역대 최대 액수인 1조 6,641억 원을 기록, 2013년 대비 7.3% 증가했고 관객 수도 2억 명을 돌파하며 작년도 대비 약 0.8% 증가한 2억 1,506만 명에 달하였다.[4)] 이와 동시에 영화에 대한 소비자들의 의견도 계속 증가하는 추세이며 최근에는 영화에 대한 소비자들의 의견이 다른 소비자의 영화 관람 여부에 큰 영향을 미치고 있다. 따라서 본 연구는 “영화”에 대한 소비자들의 리뷰에 대해 온톨로지를 구축하고 이를 활용하여 오피니언 마이닝을 수행, 시각화 한 후 결과에 대해 논의하고자 한다.

1-2. 연구의 방법

본 연구는 이론 및 선행 연구의 고찰, 온톨로지의 구축, 온톨로지를 활용한 오피니언 마이닝, 시각화의 구현 및 결과 분석 그리고 결론의 순서로 이루어진다.

이론 및 선행연구의 고찰에서는 앞서 서술되었던 연구 배경 및 목적의 토대가 되는 오피니언 마이닝 및 시각화에 대한 선행연구를 살펴보고 각 연구에서 사용된 방법과 한계점에 대해 다루었다.

온톨로지 구축을 위해 리뷰 데이터를 형태소 분리하여 범주에 따라 평가요소가 되는 키워드와 서술어를 선별하고 묶는 작업을 진행한다. 예를 들어 “연기”와 “대사”는 “배우”를 평가하는 평가요소이며, “능숙하다”, “뛰어나다”는 이에 대한 서술어가 될 수 있다. 범주에 따른 서술어사전을 구축하고 예문을 통해 영화리뷰가 분석되는 과정을 설명하도록 한다. 그리고 실제 영화 리뷰데이터를 대상으로 오피니언 마이닝을 수행하고 결과에 대한 분석을 진행한다.

시각화 구현 단계에서는 기존의 시각화 방법들을 살펴보고 본 연구에서 수행한 오피니언 마이닝 결과에 적합한 시각화를 구현하여 결과에 대해 해석하고자 한다. 이와 더불어 본 연구에서 제안하는 시각화의 분석용이성을 측정하기 위해 사용자 분석을 통해 검증하도록 한다.

마지막으로 본 연구의 의의와 연구를 통해 도출한 결과, 그리고 본 연구에서의 한계점에 대해 논하고자 한다.

2. 이론 및 선행 연구의 고찰

2-1. 오피니언 마이닝 관련연구

오피니언 마이닝은 분석방법에 따라서 문장 자체에 대한 극성을 판별하거나 문장에 나타나는 특정

3) Nahum Gershon, Stephen G. Eick, Stuart Card, Information Visualization, ACM, 1998.

4) 2014년 한국 영화산업 결산, 영화진흥위원회 정책연구부, p.6, 2014.

평가요소에 대한 극성을 판별하는 것으로 구분된다.

문장 자체에 대한 극성을 판별하는 방법은 분석 하고자 하는 키워드가 포함된 문장에서 극성을 띄는 모든 단어를 추출하여 극성 값을 계산한다. 예를 들어 긍정의 극성을 띄는 단어가 10개, 부정의 극성을 띄는 단어가 1개라면 9만큼 긍정의 극성을 띄는 문장으로 판별하게 되는 원리다. 조하나의 연구에서는 이러한 방법을 통해 인터넷 뉴스 댓글에 나타나는 감정을 분석하여 여론조사 기관의 결과와 유사한 결과를 도출할 수 있음을 보였다(조하나 등, 2013). 하지만 문장 전체의 극성을 판별하는 방법의 경우 거시적인 범위의 여론 분석에는 적합하지만 세부적인 평가요소에 대해 어떠한 평가가 이루어졌는지에 대한 분석에는 적합하지 않다.

이러한 문제점을 해결하기 위해 명재석의 연구(명재석 등, 2008)에서는 문장에 나타나는 평가 요소와 그것에 대한 평가표현을 파악하여 극성을 판별하는 오피니언 마이닝 방법을 제안하였다.

2-2. 온톨로지와 오피니언 마이닝

2-2-1. 온톨로지의 정의

온톨로지(Ontology)는 그리스어로 '존재'를 뜻하는 ontos와 '단어'를 뜻하는 logos에서 유래한 것으로 알려져 있으며(정도헌, 2003), 특정 도메인을 개념화(Conceptualization)하기 위해 명시적으로 정형화한 명세서로서 온톨로지를 정의할 수 있다(Gruber, 1993). 따라서 온톨로지를 통해 어떠한 사물을 범주화하고 다른 사물들과의 관계 또한 명시할 수 있는 것이다. 예를 들어 "개미"를 온톨로지로 표현할 때 "개미"는 2개의 더듬이를 가지고 6개의 다리를 가진다. 이를 정형화된 형태로 다음과 같이 나타낼 수 있다.

```
<Class: 개미>
    <Property: 더듬이>
        <value: 2>
    <Property: 다리>
        <value: 6>
```

〈그림 2〉 '개미'에 대한 온톨로지 표현

온톨로지의 형태는 제한이 없으며 다만 용도에 따라 RDF(Resource Description Framework/웹상의 자원의 정보를 표현하기 위한 규격), OWL(Ontology Web Language/웹 온톨로지 언어), SWRL(Semantic Web Rule Language/웹 의미 언어 규칙) 등 다양한 언어로 표현된다. 하지만 모든 온톨로지는 일반적으로 다음과 같은 구성요소를 가지게 된다.

2-2-2. 온톨로지의 구성

범주(Class)는 사물 또는 개념의 범주에 해당한다. 앞서 예시에서 "개미"를 온톨로지로 표현하기 위해 범주로 지정하였다. 범주는 어떤 개념을 떠올렸을 때 그 개념을 포함하는 상위개념의 단어를 의미한다.

속성(Property)은 범주가 가지는 성질을 나타낸다. 위의 예시에서 "개미"가 가진 더듬이는 "개미"라는 범주의 속성에 해당한다. 마찬가지로 다리 역시 "개미"라는 범주의 속성에 해당한다. 속성은 또한 값을 포함하게 되는데 위의 예시에서는 개미의 더듬이는 2개이기 때문에 2의 값을 가지게 되고 개미의 다리는 6개 이므로 6의 값을 가지게 된다.

이러한 형태로 정형화된 온톨로지는 컴퓨터로 처리하기 용이하며 때문에 시맨틱웹 중심의 정보검색 분야와 함께 인공지능, 전자상거래 등 다양한 분야에서 활용되고 있으며 본 연구에서 진행하고자 하는 오피니언 마이닝 분야에서도 역시 활용되고 있다.

2-2-3. 온톨로지를 활용한 오피니언 마이닝

오피니언 마이닝 분야에서 온톨로지는 대부분 요소 기반 오피니언 마이닝에 대한 연구에서 활용되고 있다. 평가 대상이 가지고 있는 수많은 평가요소들 간의 관계를 표현하는데 온톨로지가 적절한 자료구조로 활용될 수 있기 때문이다. 예를 들어 "연기력"은 연기력 그 자체에 대한 평가가 될 수도 있지만 "연기력"이라는 평가요소를 가진 "배우"에 대한 평가가 될 수도 있다. 관련 연구로 Anaïs Cadilhac et al(2010)의 연구와 Larissa A. de Freitas et al.(2013)의 연구에서는 오피니언 마이닝을 위한 평가요소 추출에 온톨로지를 활용할 수 있음을 보였다.

2-3. 정보 시각화

2-3-1. 정보시각화의 정의

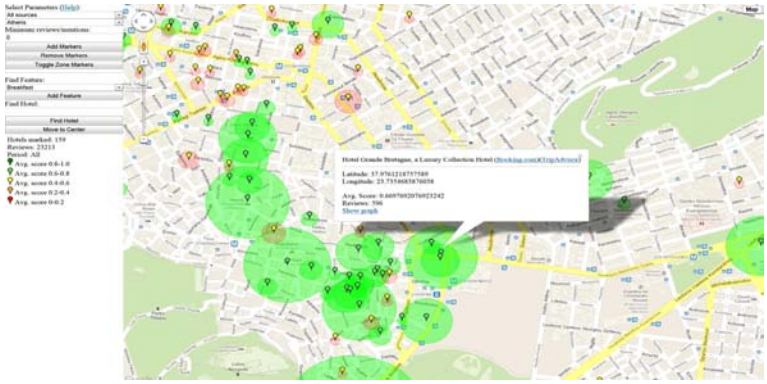
사람이 감각을 통해 얻는 정보 중 70%는 시각을 통해 습득하게 되는데 이는 청각 및 촉각에 비해 매초 100배 가까운 정보를 얻을 수 있기 때문이다. 때문에 인류는 선사시대부터 시각적인 방법을 통해 정보를 전달하고 분석해왔으며 막대그래프나 파이차트 등 다양한 시각적 그래프를 우리는 이미 익숙하게 사용하고 있다. 최근 빅데이터의 활성화와 함께 데이터마이닝을 위한 방법론으로서 정보시각화(Information Visualization)라는 용어로 많은 관심과 연구가 이루어지고 있다.

정보 시각화는 데이터에 대한 인지력 확장을 위해 전산처리, 상호작용, 시각적인 표현을 사용하는 것으로 정의할 수 있다.[5] 시각적인 디자인은 정보의 이해를 빠르게 전달할 수 있으며 시각화된 이미지의 패턴을 통해 의미 있는 관계를 찾을 수도 있다(최영화, 2012).

2-3-2. 오피니언 마이닝의 시각화

오피니언 마이닝에 대한 관심이 증가함에 따라 오피니언 마이닝의 결과를 인지적으로 쉽게 판단할 수 있도록 돕는 시각화에 대한 연구도 진행되고 있다. 허문열의 연구(허문열 등, 2007)에서는 여론조사 분석에 시각화를 활용하였다. 연구에 사용된 데이터가 오피니언 마이닝을 통한 데이터가 아닌 수치화된 정형데이터를 사용하였지만 시각화를 여론조사의 결과분석에 활용할 수 있음을 확인하였다. 이윤정의 연구(이윤정 등, 2009)에서는 인터넷 댓글에 나타나는 키워드에 따라 유사한 키워드를 가진 댓글끼리 군집화(Clustering)하고 그 결과를 시각화하는 시스템을 제안하였다. 이 연구에서는 시각화를 통해 유사한 의미를 가진 댓글을 분류하였지만 문맥적 의미를 고려하지 않았기 때문에 극성 정보를 확인할 수 없는 한계를 가지고 있다. Eivind Bjørkelund 의 연구(Eivind Bjørkelund 등, 2012)에서는 그림 3과 같이 호텔 예약 웹서비스인 TripAdvisor와 Booking.com의 투숙객 리뷰를 대상으로 오피니언 마이닝 시각화를 제안하였다. 그들은 평가요소에 대한 0과 1사이의 오피니언 스코어를 계산하여 최종 평균 스코어를 해당 호텔의 극성 스코어로 결정하고 이를 Google Map을 활용하여 그림 2 와 같이 지도상에 나타내는 방법을 제안하였다. 사용자는 호텔의 색상이 녹색인지 붉은색인지를 보고 해당 호텔에 대한 투숙객들의 감성평가를 확인할 수 있다. 하지만 이 시각화는 호텔의 어느 요소가 부정적인 평가를 받았는지 확인하기 어렵다는 한계를 가지고 있다.

5) Stuart K. Card, Jock D. Mackinlay, Ben Shneiderman, Readings in Infomation Visualization: Using Vision to Think, p 7, Morgan Kaufmann, 1999.



〈그림 3〉 Biokelung의 시각화 프로토타입

3. 연구 및 분석

3-1. 온톨로지 구축

3-1-1. 평가요소 선정

오피니언 마이닝을 위한 온톨로지를 구축하기 위해 우선은 온톨로지의 범주가 되는 범주와 각 범주와 관련된 속성들의 평가요소 키워드를 선별할 필요가 있다. Li Zhuang 은 영화리뷰를 분석하고 요약하는 그의 연구(Li Zhuang 등, 2006)에서 영화 구성요소의 키워드를 표 1 과 같이 선별하였다.[6]

〈표 1〉 Li Zhuang 연구에서의 영화 구성요소 키워드 분류

Element class	Feature words
Overall	film, movie
ScreenPlay	story, plot, script, storyline, dialogue, screenplay, ending, line, scene, tale
Character	character, characterization, role
Vision Effects	scene, filght-scene, action-scene, action-sequence, set, battle-scene, picture, scenery, setting, visual-effects, color, background, image
Music and Sound	music, score, song, sound, soundtrack, theme
Special Effect	special-effects, effect, CGI, SFX

하지만 이렇게 분류된 키워드는 한글로 작성된 데이터에는 적합하지 않다. 따라서 본 연구에서는 위에서 제시된 분류를 참고하여 새롭게 리뷰 데이터에서 색인어를 추출하고 평가요소가 될 수 있는 키워드를 선별, 분류하였으며 토픽 모델링을 활용하여 이를 검증하였다는 부분에서 기존연구와 차별성이 있다.

데이터는 NAVER 영화 서비스에서 찾을 수 있는 영화에 대한 140자평을 자체 제작한 크롤러(Crawler)를 사용하여 수집하였다.[7] 온톨로지 구축을 위해 사용될 표본 영화는 장르가 유사하여 발생하는 편향성과 리뷰의 개수 부족으로 발생할 수 있는 표본 수 부족으로 인한 일반화의 오류를 방지하기 위해 장르의 중복이 최대한 발생하지 않으면서 관객의 리뷰가 많이 포함된 영화를 표본으로 선정하고자 하였다. 두 가지 사항을 주요 기준으로 선별된 5편의 표본 영화는 표 2 와 같으며 해당 영화들로부터 총 18,518개의 다양한 리뷰가 수집되었다.

6) Zhuang, L., Jing, F., Zhu, X. Y. (2006, November). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, pp. 46, 2006.

7) NAVER 영화, http://movie.naver.com

〈표 2〉 키워드 추출을 위해 선정된 영화와 영화의 장르

영화이름	장르
링컨 차를 타는 변호사	범죄, 드라마, 스릴러
캐리비안의 해적: 낯선 조류	액션, 모험, 판타지
오싹한 연애	멜로/로맨드, 공포, 코미디
장화신은고양이	애니메이션, 모험, 코미디, 판타지
화이트: 저주의 멜로디	공포, 미스터리

다음으로 18,518개의 리뷰를 대상으로 색인어 추출 작업을 진행한 결과 총 12,639가지의 색인어가 등장하였고 이 중에서 출현 빈도 빈도가 높고 선행 연구의 결과를 참조하여 6개의 범주에 대한 평가요소에 속하는 키워드를 선정하였다. 선정된 키워드는 표 3 과 같다.

〈표 3〉 색인어 추출 작업을 통해 추출된 빈도가 높은 색인어

범주	속성
영화	영화(4374), 작품(101), 전체적(66), 스케일(48)
배우	연기(606), 배우(217), 캐릭터(155), 주인공(91), 연기력(84), 대사(55), 조연(48), 캐스팅(32), 목소리(24), 스타일(24), 인물(18)
감독	감독(157), 연출(63), 구성(57), 편집(10)
스토리	스토리(849), 내용(155), 소재(155), 전개(141), 이야기(127), 결말(122), 엔딩(55), 시나리오(55), 개연성(30), 줄거리(29), 설정(29)
영상	장면(293), 씬(36), 볼거리(46), 영상(38), 분위기(36), 표현(32), 화면(22)
음향	노래(136), 소리(135), 사운드(41), 음향(16), 멜로디(13), 곡(10)

다음으로 빈도에 따라 색출된 색인어에 대한 검증 과정으로 토픽모델링을 활용하였다. 토픽모델링 혹은 LDA(Latent Dirichlet Allocation)로 불리는 비정형데이터에 대한 일반 확률 모델은 어떤 확률분포와 그 파라미터가 있다고 가정 할 때, 그로부터 랜덤 프로세스에 따라 데이터를 생성하는 모델이다.[8] 또한 개별 문서 더 나아가 문서 컬렉션(Corpus)를 표현하는 방법을 찾기 위해 많이 사용된다. 토픽 모델링은 다양한 분야에서 활용되고 있으며 특히 주제 분류나 문서 간 유사도 계산을 할 때 많이 사용된다. 6가지 대표 분류어휘와 관련된 토픽모델링 결과는 다음의 표 4 와 같다.

〈표 4〉 선정된 키워드에 대한 상위 10개 토픽모델링 결과

토픽	키워드
영화	영화, 작품, 평가, 전체, 스케일, 주제, 예술, 완성도, 대중, 상업
배우	배우, 연기, 시나리오, 각본, 스토리, 조연, 주연, 연기력, 출연, 역할
감독	감독, 작품, 연출, 연출력, 흡입력, 편집, 봉준호, 박찬욱, 천재
스토리	스토리, 볼거리, 구성, 그래픽, 전개, 짜임새, 스케일, 내용, 영상미, 결말
영상	영상, 장면, 음악, 스토리, 배경, 영상미, 분위기, 화면, 촬영, 그래픽
음향	음향, 노래, 사운드, 소리, 멜로디, 곡, 효과, 감성, 자극, 청각

8) Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation", the Journal of machine Learning research, p. 999, 2003.

대표 분류어휘와 관련된 주제 관련 어휘를 토픽모델링을 통해 확인한 결과 빈도를 활용하여 색인된 색인어와 유사한 결과가 나오는 것을 확인하였다. 따라서 빈도가 높고 평가요소로서 적절한 키워드를 토픽 모델링의 결과와 비교하여 온톨로지를 구성할 키워드로 선정하였다. 최종으로 선정되어 온톨로지에 활용된 키워드는 표 5 와 같다.

〈표 5〉 온톨로지 구축을 위해 최종 선정된 키워드

범주	속성
영화	영화, 작품, 평가, 전체, 스케일, 주제, 예술, 완성도, 대중, 상업
배우	배우, 캐릭터, 연기력, 역할, 스타일, 대사, 등장인물, 캐스팅, 주인공, 연기, 인물, 출연, 조연, 사람, 대본
감독	감독, 편집, 연출, 연출력, 구성, 흡입력
스토리	스토리, 주제, 결말, 소재, 각본, 시나리오, 이야기, 내용, 원작, 전개, 줄거리, 설정, 짜임새, 스토리 텔링, 구상, 구조, 사건, 개요, 실화
영상	영상, 장면, 배경, 볼거리, 화면, 표현, 그래픽, 영상미, 분위기, 비주얼, 시퀀스, 분장
음향	사운드, 노래, 소리, 음향, 멜로디, 곡, 효과, 감성, 자극, 청각

3-1-2. 평가어휘 구축

본 연구에서는 평가요인이 되는 대상의 범주에 따라 별도의 서술어 사전을 구축함으로써 오피니언 마이닝 처리과정의 효율성을 증가시키고자 하였다. 이를 위해 상위 단계에서 구축된 각 범주별 평가요소 키워드와 함께 나타나는 서술어들을 추출하여 사전을 구축하였다. 예를 들어 “시나리오가 단순하다.”라는 문장은 형태소 분석과정에서 다음과 같이 분리된다.

시나리오/NNG(일반명사) 가/JKS(주격조사) 단순/NNG(일반명사)

하/XSV(동사파생접미사) 다/EFN(종결어미) ./SF(마침표, 물음표, 느낌표)

이 문장은 “스토리” 범주에 포함되는 “시나리오”라는 키워드를 포함하고 있기 때문에 유의미한 문장으로 취급된다. 그리고 이 문장에서 “시나리오”에 대해 “단순하다”라는 서술어가 등장하므로 이를 “스토리”에 대한 서술어사전에 추가하게 된다. 서술어사전은 그림 4와 같은 레코드의 집합으로 이루어진다.

〈그림 4〉 서술어사전의 레코드 형식

Class는 해당 서술어가 속하는 범주를 의미하며 문장 내에 이 범주에 포함되는 키워드가 있어야 해당 서술어가 탐색 대상으로 지정된다. 상위의 예문에서 이 범주는 “스토리” 가 된다.

Pdt는 본문에 나타나게 되는 서술어(Predicate)의 원형이다. 상위의 예문에서는 “단순하다” 가 된다.

TagPdt는 형태소 분리가 된 상태의 서술어를 의미한다. 문장을 분석하는 단계에서 형태소로 분리된 상태의 문장을 활용하기 때문에 서술어 또한 형태소로 분리된 형태가 필요하다. 상위의 예문에서는 “단순/NNG 하/XSV”가 된다.

Pol은 서술어가 가지는 극성(Polarity)을 의미하며, 선행연구에서 언급한 조하나의 연구를 참고하여 긍정일 경우 1, 부정일 경우 -1을 나타낸다.

본 연구에서는 오피니언 마이닝의 정확도 향상을 위해 은어와 관용어도 서술어사전에 포함하였다. 전현경은 관용어를 이루는 고정 단어 열을 복합단위형태소라 정의하고 형태소 분석 과정에서 하나의 단위로 인식하는 방법을 제안하였다(전현경 외, 1998). 본 연구에서도 이러한 방법을 사용하여 은어와 관용어를 하나의 단위로 인식하는 방법을 사용하였다. 상위에 제시된 과정을 통해 표 6 과 같은 서술어사전이 구축되었다.

〈표 6〉 범주에 따른 서술어 사전의 예

Class	Pdt(predicate)	TagPdt	Pol
영화	훌륭하다	훌륭/XR(어근) 하/XSV(동사파생접미사)	1
	형편없다	형편없/VA(형용사)	-1
배우	훌륭하다	훌륭/XR(어근) 하/XSV(동사파생접미사)	1
	호흡이잘맞다	호흡/NNG(일반명사) 이/JKS(주격조사) 잘/MAG(일반부사) 맞/VV(동사)	1
감독	훌륭하다	훌륭/XR(어근) 하/XSV(동사파생접미사)	1
	형편없다	형편없/VA(형용사)	-1

3-2. 영화 리뷰 오피니언 마이닝

이번 단계에서는 “연기는 훌륭하였지만 스토리는 따분했다.” 라는 예문으로 온톨로지를 활용한 오피니언 마이닝의 수행 과정을 설명하고자 한다. 문장은 형태소 분석 단계를 거쳐 표 7 과 같이 형태소로 분리된다.

〈표 7〉 형태소 분리 결과

- 연기/NNG(일반동사) 는/JX(보조사) 훌륭/XR(어근) 하/XSA(형용사파생접미사) 였/EPT(선어말어미) 지만/ECE(연결어미)
- 스토리/NNG(일반동사) 가/JKS(주격조사) 따분/XR(어근) 하/XSV(동사파생접미사) 었/EPT(선어말어미) 다/EFN(종결어미) ./SF(마침표, 물음표, 느낌표)

형태소 분석된 문장은 온톨로지의 각 범주에 포함된 키워드를 가지고 있는지 검사한다. 이 과정에서 온톨로지에 포함된 키워드를 가지고 있지 않은 경우 해당 문장은 무의미한 문장으로 분리되어 자동으로 다음 문장을 분석한다. 예문에서는 표 8 과 같이 “배우”의 속성에 해당하는 “연기”와 “스토리”의 속성에 포함되는 “스토리”가 평가요소인 주제어(SUB)로 처리되었다.

〈표 8〉 평가요소 탐색 결과

- **연기/SUB** 는/JX(보조사) 훌륭/XR(어근) 하/XSA(형용사파생접미사) 였/EPT(선어말어미) 지만/ECE(연결어미)
- **스토리/SUB** 가/JKS(주격조사) 따분/XR(어근) 하/XSV(동사파생접미사) 었/EPT(선어말어미) 다/EFN(종결어미) ./SF(마침표, 물음표, 느낌표)

평가요소가 발견된 문장은 유효한 문장으로서 해당 평가요소에 대한 서술어가 나타나는지 확인한다. 서술어는 발견된 평가요소의 범주에 속하는 것으로 제한되어 탐색한다. 예를 들어 발견된 평가요소인 “연기”는 “배우”범주에 속하므로 “배우”범주의 서술어사전 중 “훌륭하다”라는 서술어만 탐색된다. 반면에 “따분하다”는 “배우”범주에 속하는 서술어가 아니므로 탐색되지 않는다. 표 9 는 “배우” 범주의 서술어가 탐색된 결과다.

〈표 9〉 "배우"에 대한 평가표현 탐색결과

- **연기/SUB** 는/JX(보조사) **훌륭하다/PDT** 지만/ECE(연결어미) - **스토리/SUB** 가/JKS(주격조사) 따분/XR(어근) 하/XSV(동사파생접미사) 었/EPT(선어말어미) 다/EFN(종결어미) ./SF(마침표, 물음표, 느낌표)

주어진 예문은 복문이므로 두 번째 평가요소인 "스토리"에 대한 서술어 또한 탐색한다. 탐색 과정 중 이미 탐색된 "훌륭하다"는 /PDT 태그가 붙은 상태이므로 탐색 대상에서 제외된다. 두 번째 평가요소인 "스토리"에 대한 평가표현을 탐색한 결과는 표 10 과 같다.

〈표 10〉 "스토리"에 대한 평가표현 탐색결과

- **연기/SUB** 는/JX(보조사) **훌륭하다/PDT** 지만/ECE(연결어미) - **스토리/SUB** 가/JKS(주격조사) **따분하다/PDT** 다/EFN(종결어미) ./SF(마침표, 물음표, 느낌표)

발견된 서술어의 극성을 서술어사전에서 참고하여 리뷰를 요약한 결과는 표 11 과 같다.

〈표 11〉 리뷰 요약 결과

Class:배우 Property:연기 Predicate:훌륭하다 Polarity: 1 Class:스토리 Property:스토리 Predicate:따분하다 Polarity: -1

위와 같이 요약된 리뷰는 등장한 극성 값을 종합하여 평가요소와 범주의 최종 점수로 반영된다.

3-3. 시각화

오피니언 마이닝을 통한 제품의 긍부정 분석은 단일 제품에 대한 소비자들의 평가를 분석하기에 적합하지만 경쟁사 제품 등 같은 도메인의 제품들을 비교하며 분석하기 힘들다는 단점을 가지고 있다. 본 연구에서는 기존의 오피니언 마이닝 시각화에서 볼 수 있었던 제품에 대한 긍정과 부정에 대한 정보를 제공하는 동시에 위에 제시된 단점을 보완하기 위해 유사한 평가를 받은 제품 간 군집을 형성하고 각 제품에서 가장 많은 평가를 받은 요소를 확인함으로써 다른 제품들과 비교 분석할 수 있는 새로운 시각화를 제안하고자 한다. 이를 위해 우선 제품에 대한 긍부정 정보 시각화를 수행하였고 유사한 평가를 받은 제품끼리 군집화를 위해 군집 정보 시각화를 수행하였다. 마지막으로 인터페이스를 구성하여 사용자와의 인터렉션을 통해 다양한 시각화 분석이 이루어질 수 있도록 하였다.

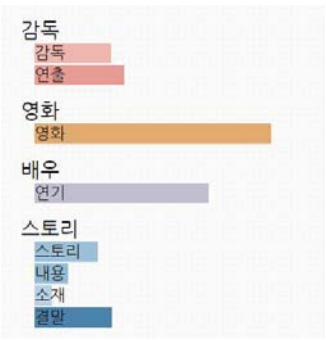
3-3-1. 정보 시각화

오피니언 마이닝에서 정보를 표현하는 요소로는 색상을 가장 많이 활용한다. 앞서 소개한 오피니언 마이닝 시각화 중 하나인 Bjørkelund et al. 의 호텔 리뷰 시각화의 경우에도 긍정적인 평가를 받은 호텔은 녹색, 부정적인 평가를 받은 호텔은 붉은색으로 표현하였다. 하지만 본 연구에서는 하나의 제품에 여러 개의 평가요소들이 존재하며 이러한 평가요소들을 구분하기 위해 평가요소를 포함하는 6개의 대표 범주에 그림 5의 좌측과 같이 색상을 부여하였다.

실제 리뷰에서 얼마나 언급되었는지 표현하기 위해 각각의 영역의 크기가 그림 5의 우측과 같이 조절되도록 하였다. 예를 들어 그림 5의 우측과 같은 경우 리뷰에서 스토리에 대한 평가가 가장 많이 이루어졌음을 확인할 수 있고 그 다음으로 영화자체에 대한 평가가 많이 이루어졌다는 것을 알 수

있다. 하지만 이와 같은 원형 그래프에서는 세부 평가요소에 대해 일일이 레이블링하여 나타내기 어렵기 때문에 그림 6과 같이 별도의 막대그래프를 통해 세부 평가요소에 대한 평가결과를 표현하였다.

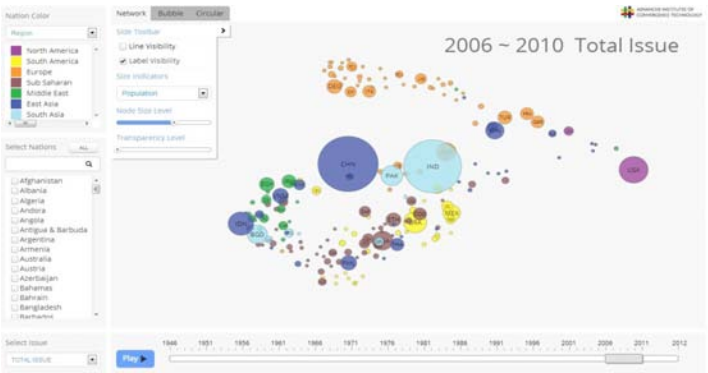
〈그림 5〉 평가 범주에 대한 색상부여(좌)와 평가 결과 시각화(우)



〈그림 6〉 세부 평가요소에 대한 막대그래프 시각화 예시

3-3-2. 군집 시각화

데이터의 수가 많은 경우에는 개별적으로 데이터의 특성을 확인하기가 어렵다. 따라서 데이터를 군집화 하고 군집간의 특성을 확인하여 데이터 사이의 관계를 파악한다. 이러한 시각화의 예로 Politiz UN(http://203.234.55.97/politiz/un/#)에서는 연도별로 UN가입국가들의 투표결과를 분석하여 그림 7과 같이 보여주며 이를 통해 사용자들은 시간에 따른 투표성향의 변화와 시간에 따른 국가 간 이해관계 변화를 파악 할 수 있다(Ginam Kim, 2013).



〈그림 7〉 유엔 표결데이터로 국가 간 군집을 표현한 Politiz UN

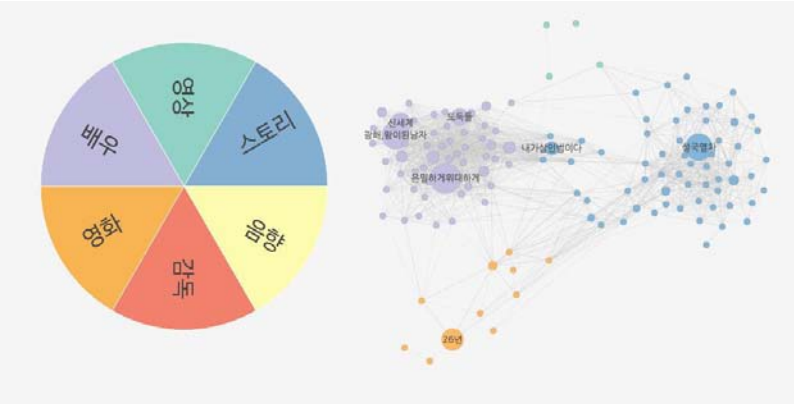
본 연구에서도 유사한 평가를 받은 제품끼리 군집을 형성하기 위해 오피니언 마이닝 결과를 바탕으

로 유사도를 계산하였다. 유사도 계산을 위한 각 제품의 벡터 값은 전체리뷰 중 평가요소별 비율이며 코사인 유사도(Cosine Similarity)로 계산하였다. A와 B의 벡터 값이 n개 주어졌을 때 코사인 유사도는 다음과 같이 구할 수 있다.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times (B_i)^2}$$

유사도의 결과는 대상이 n개 일 때 $n \times n$행렬이며, 각 원소는 두 대상의 유사도를 나타낸다. 예를 들어 영화별 유사도 행렬을 M이라 할 때, $M_{\text{설국열차,아이언맨}}$은 영화 “설국열차”와 “아이언맨”의 유사도를 나타내며 1에 가까울수록 유사한 평가를 받은 영화가 된다.

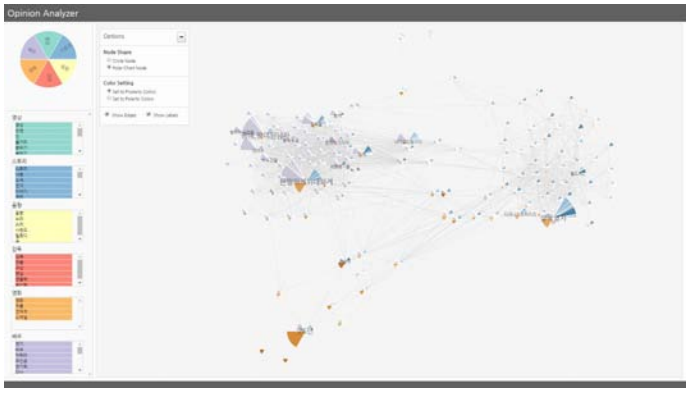
유사도 결과에 따라 네트워크를 형성하기 위한 알고리즘으로는 포스 다이렉티드(Force-directed) 알고리즘(Fruterman et al, 1991)을 사용하였다. 포스 다이렉티드 알고리즘을 사용함으로써 각 제품에 해당하는 노드들은 자연스럽게 비슷한 성질을 가지는 노드들과 군집을 형성하게 된다. 같은 성질을 가지는 노드를 직관적으로 쉽게 파악할 수 있도록 각각의 노드는 평가가 가장 많이 이루어진 평가요소의 범주에 해당하는 색상을 띄며, 그 색상은 그림 8에서 제시되었던 색상과 같다. 또한 그림 8과 같이 6개의 범주가 가지는 방향성을 활용하여 노드 군집이 모이는 방향이 일치하게 하여 분석이 용이하도록 하였다.



〈그림 8〉 6개 범주의 방향성과 노드 군집의 방향

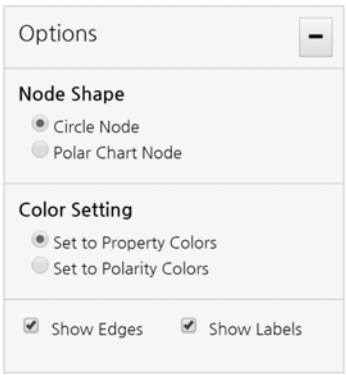
3-3-3. 인터페이스(Interface) 및 인터랙션(Interaction)

방대하고 복잡한 데이터를 효과적으로 분석하기 위해서는 시각화 디자인과 사용자 사이의 인터랙션이 무엇보다 중요하다(서진욱, 2011). 그래프를 확대하거나 색상을 변경, 또는 불필요한 정보를 필터링함으로써 정지된 상태의 그래프에 비해 좀 더 직관적이고 다양한 정보를 얻을 수 있다. 본 연구에서 제시하는 시각화도 인터페이스를 통해 여러 가지 인터랙션을 지원하여 분석의 용이성을 향상시키고자 하였다. 따라서 시각화는 웹 브라우저에서 동작이 가능하도록 HTML5 Canvas API를 활용하여 제작하였으며 ‘오피니언 분석기(Opinion Analyzer)’로 명명한 시각화의 메인화면은 그림 9와 같이 구성하였다.



〈그림 9〉 Opinion Analyzer의 메인화면

화면의 좌측에는 단일 대상에 대한 시각화를 볼 수 있다. 좌측의 상단에는 원형 폴라(Polar) 그래프로 선택한 대상의 평가요소의 빈도를 보여준다. 좌측 하단에는 막대그래프로 각 범주에 속하는 세부 평가요소들의 비율을 시각화 하였다. 화면의 우측에는 군집 시각화가 위치한다. 군집 시각화는 줌인(Zoom In), 줌아웃(Zoom Out) 및 드래그 앤 드롭(Drag and Drop)으로 원하는 크기와 위치로 노드의 군집을 이동시킬 수 있다. 군집 시각화의 왼쪽에는 옵션(Options) 메뉴로 군집 시각화에 대한 다양한 조작이 가능하도록 한다. 옵션 메뉴는 그림 10과 같이 구성되어있다.



〈그림 10〉 옵션 메뉴의 구성

노드 형태(Node Shape)옵션은 노드의 형태를 결정한다. 첫 화면에서 노드의 모양은 색상을 가진 원 모양을 하고 있으나, 폴라 그래프(Polar Chart Node)를 선택하면 그림 11과 같이 단일 대상에 대한 오피니언 결과를 부채꼴 모양의 원 그래프의 형태로 나타내는 시각화로 노드의 모양이 변경된다. 폴라 그래프에서는 전체 대상에 대한 오피니언 마이닝의 정보를 개략적으로 확인 할 수 있다는 장점

〈그림 11〉 'Polar Chart'인 노드 형태를 선택한 화면

을 가지고 있다.

컬러 세팅(Color Setting)옵션으로 노드의 색상을 변경할 수 있다. 그림 12(좌)는 이 기능을 이용하여 긍정과 부정을 나타내는 두 가지 색으로 표현되도록 변경한 결과로 부정적인 평가를 받은 노드가 붉은색으로 나타나는 것을 확인 할 수 있다. 노드형태가 폴라 그래프 일 경우 그림 12(우)와 같이 부정적인 평가를 받은 평가요소가 붉은색으로 변한 것을 확인할 수 있다.

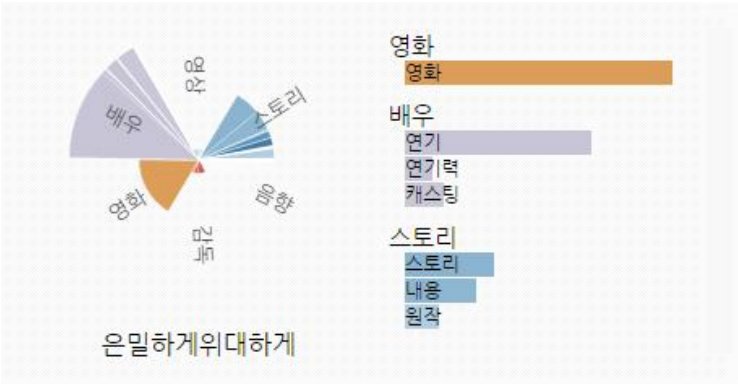
〈그림 12〉 긍정과 부정의 색으로 표현한 시각화

3-4. 시각화 분석

본 연구에서는 제품에 대한 전체 소비자들의 의견을 종합하여 보여주는 시각화 방법을 제안하고자 하였다. 이에 대한 사례 연구(Cast Study)로 2011년부터 2013년 까지 개봉된 영화 중 영화 리뷰가 8000개 이상인 영화 130작품을 선정하고 해당 영화에 대한 전체 관객의 의견을 오피니언 마이닝을 통해 분석하고 시각화 하였다. 개발된 시각화는 URL : 54.255.190.140/index/# 에서 확인할 수 있다.

3-4-1. 개별 영화에 대한 분석

본 연구에서 제안하는 시각화에서 개별 영화에 대한 오피니언 마이닝 결과는 하나의 원형 그래프와 막대그래프로 보여준다. 그림 13은 '은밀하게 위대하게'의 오피니언 마이닝 결과에 대한 시각화 그래프다.



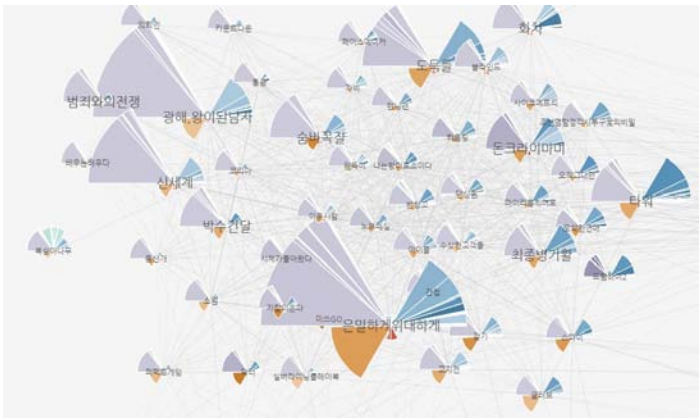
〈그림 13〉 영화 “은밀하게 위대하게”의 오피니언 마이닝 결과 시각화

원형 그래프는 12시, 2시, 4시, 6시, 8시, 10시의 총 6방향으로 각각 평가 요소에 대한 평가빈도와 극성을 나타낸다. 각각 영상에 대한 평가가 많을수록 12시 방향의 부채꼴의 크기가 증가하며, 스토리에 대한 평가가 많을수록 2시 방향의 부채꼴의 크기가 증가하며, 음향에 대한 평가가 많을수록 4

시 방향의 부채꼴의 크기가 증가하며, 감독에 대한 평가가 많을수록 6시 방향의 부채꼴의 크기가 증가하며, 영화자체에 대한 평가가 많을수록 8시 방향의 부채꼴의 크기가 증가하며, 배우에 대한 평가가 많을수록 10시 방향의 부채꼴의 크기가 증가한다. 예를 들어 영화 '은밀하게 위대하게'는 배우에 대한 평가가 많이 이루어짐을 그래프를 통해 알 수 있다. 세부적인 평가요소에 대한 결과는 막대그래프에서 확인할 수 있다.

3-4-2. 노드의 위치에 대한 분석

Opinion Analyzer에서 군집 시각화는 각 대상이 가진 평가요소의 비중에 따라 노드의 위치가 결정되기 때문에 반대로 노드의 위치는 평가요소의 비중에 대한 정보를 제공하게 된다. 그림 14는 배우에 대한 평가의 비중이 높은 영화들의 군집이다. 군집에서 "은밀하게 위대하게"를 중심으로 영화 "신세계"는 좌측에, 영화 "타워"는 우측에 위치하게 된다. 그림 15는 각각 영화 "신세계", "은밀하게 위대하게", "타워"의 평가요소의 비중이며, 배우에 대한 평가요소의 크기는 비슷하나 오른쪽에 위치할수록 스토리에 대한 평가의 비중이 높아지는 것을 확인할 수 있다.



〈그림 14〉 배우에 대한 평가의 비중이 높은 영화의 군집

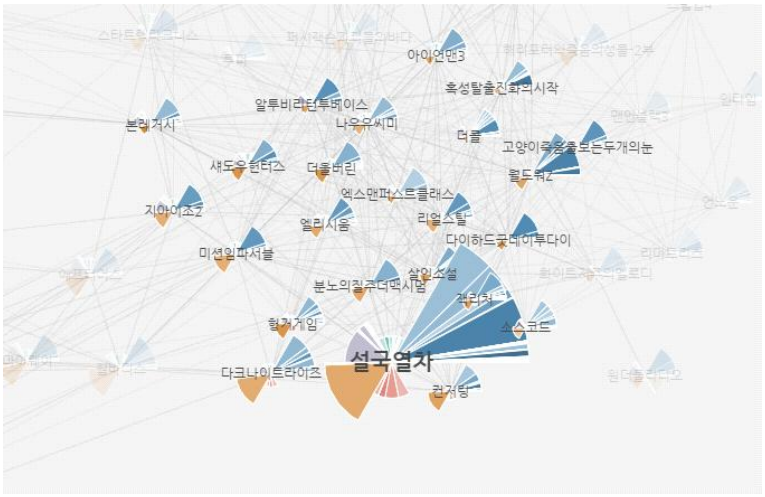
〈그림 15〉 영화 "신세계", "은밀하게 위대하게", "타워"의 평가요소의 비중

3-4-3. 군집된 영화 집단에 대한 분석

일반적으로 데이터의 양이 많은 경우 특성이 비슷한 데이터들을 집단끼리 군집화하여 집단 사이의 관계를 분석하는 것이 보편적이다. 본 케이스 스터디에서 최종적으로 형성된 집단은 크게 3가지 집단으로 나누어 볼 수 있었다.

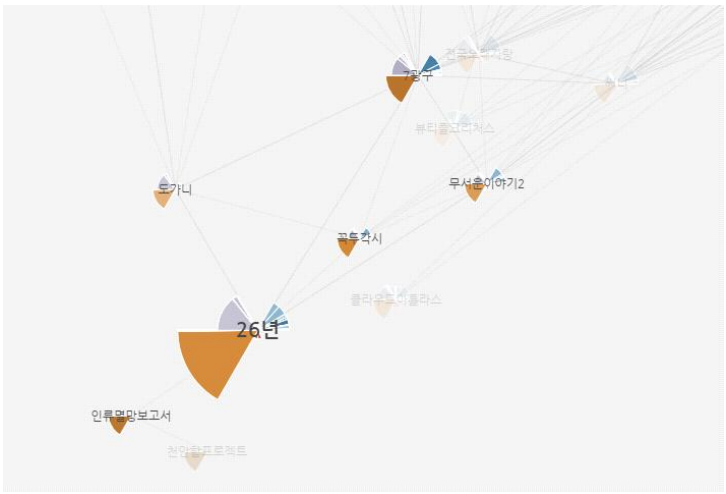
그림 16은 영화 '설국열차'을 중심으로 군집 된 첫 번째 집단이다. 시각화를 통해 집단의 특성을 확인한 결과 첫 번째 집단은 영화 리뷰에서 배우나 스토리에 대한 언급이 많은 집단이라는 것을 알 수 있다. 또한 군집의 중심으로 볼 수 있는 '설국열차'의 경우 대표 키워드 값이 스토리(0.438), 영화(0.213), 배우(0.149), 감독(0.114), 영상(0.083), 음향(0.002)의 순서로 분포하였고 이를 통해 첫

번째 집단에 속하는 영화들은 스토리나 영화에 대한 언급이 많은 집단이라는 것을 유추할 수 있다.



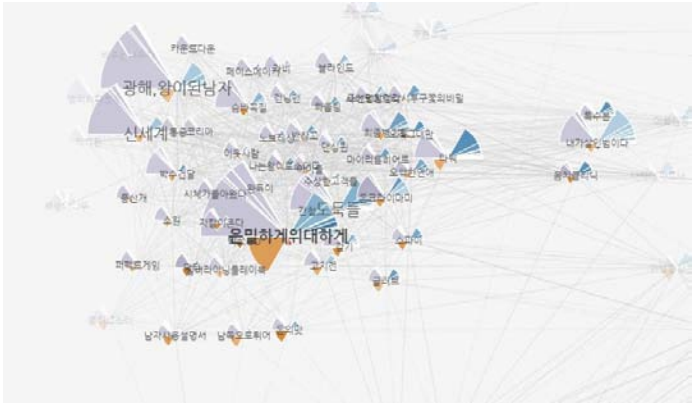
〈그림 16〉 영화 “설국열차”를 중심으로 군집된 첫 번째 집단

그림 17은 영화 ‘26년’을 중심으로 군집된 두 번째 집단이다. 시각화를 통해 집단의 특성을 확인한 결과 두 번째 집단은 영화 리뷰에서 영화, 배우, 스토리에 대한 언급이 많은 집단이라는 것을 알 수 있다. 또한 군집의 중심인 ‘26년’의 경우 대표 키워드 값이 영화(0.483), 배우(0.242), 스토리(0.192), 감독(0.054), 영상(0.021), 음향(0.007)의 순서로 분포하였고 이를 통해 두 번째 집단에 속하는 영화들은 영화, 배우, 스토리에 대한 언급이 많은 집단이라는 것을 유추할 수 있다.



〈그림 17〉 영화 “26년”을 중심으로 군집된 두 번째 집단

그림 18은 영화 ‘은밀하게 위대하게’를 중심으로 군집된 세 번째 집단으로 시각화를 통해 집단의 특성을 확인한 결과 세 번째 집단은 영화 리뷰에서 배우나 스토리, 영화에 대한 언급이 많은 집단이라는 것을 알 수 있다. 또한 군집의 중심인 ‘은밀하게 위대하게’의 경우 대표 키워드 값이 배우(0.577), 스토리(0.228), 영화(0.129), 감독(0.031), 영상(0.029), 음향(0.004)의 순서로 분포하였고 이를 통해 세 번째 집단에 속하는 영화들은 배우나 스토리에 대한 언급이 많은 집단이라는 것을 유추할 수 있다.

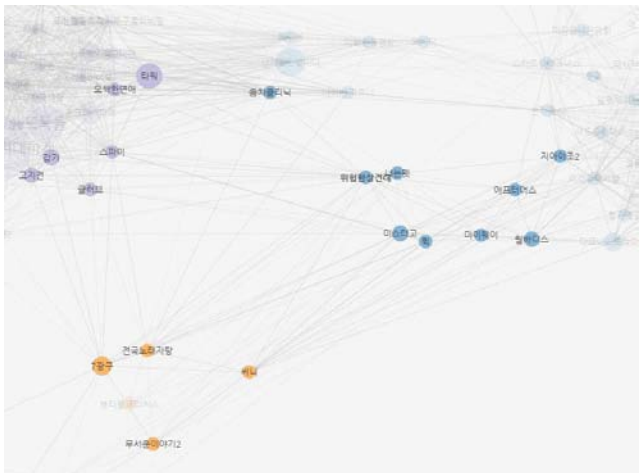


〈그림 18〉 영화 “은밀하게 위대하게”를 중심으로 군집된 세 번째 집단

3-4-4. 데이터간의 네트워크 분석

일반적으로 노드들 사이에 특성이 유사한 노드들을 표현하는 방법으로 네트워크를 활용한다. 네트워크를 사용하면 군집과 군집사이에 위치한 노드가 어떤 데이터들과 비슷한 특징을 가지고 있는지를 확인 할 수 있다.

그림 19의 경우 군집화 된 집단들의 중심점에 위치하는 영화 ‘위험한 상견례’와 이와 유사한 영화들을 링크로 연결하여 표현하고 있다. 영화 ‘위험한 상견례’와 유사한 영화는 ‘퀵’, ‘음치클리닉’, ‘너는 펫’, ‘스파이’, ‘글러브’, ‘에프터어스’등이다. 링크가 연결된 영화들은 모두 다른 집단에 포함되어 있다. 이를 통해 네트워크를 활용하면 군집화에서 발견하지 못한 새로운 군집을 확인 할 수 있다는 결과를 얻을 수 있다.



〈그림 19〉 영화 “위험한 상견례”와 연관된 영화들의
네트워크

3-5. 시각화 검증

본 연구에서는 시각화 툴에서 제공하는 여러 기능에 따라 분석용이성에 차이가 나는지를 알아보기 위해 사용성 평가를 시행하였으며 이를 통해 각각의 기능에 따른 장점과 단점을 도출하고자 하였다. 실험은 각각의 시각화 기능을 달리 하였을 때 사용자가 느끼는 분석용이성을 알아보기 위해 1대1 개별 실험을 통한 폐쇄적 실험 방법을 실시하였으며, 실험기간은 2015년 6월 22일부터 6월 26일까지 5일간이었다. 실험 대상은 시각화 분야에 대한 지식을 지니고 현재 데이터 시각화 분야를 공부중인 대학원생들을 표본으로 설정하였으며 총 30명을 대상으로 실험을 하였다.

분석을 위해 개발된 시각화는 옵션메뉴를 활용하여 총 16가지의 시각화의 형태를 보여줄 수 있다. 노드의 형태에 따라 원 모양의 노드(형태 1), 폴라 그래프 모양의 노드(형태 2)로 구분 될 수 있고, 노드의 색상에 따라 6개 평가요인의 범주 색상(색상 1)과 극성 정보에 따라 붉은색과 파란색으로 보여주는 방법(색상2)으로 구분 될 수 있다. 또한 유사한 노드사이의 연결선을 보이거나 안보이게 하는 방법(네트워크)과 노드의 이름(라벨)을 보이거나 안보이게 하는 방법이 있다. 다음 표 12 는 이러한 방법에 따라 나누어진 16가지 집단에 대한 설명이다.

〈표 12〉 시각화 기능에 따른 집단

집단	형태1	형태2	색상1	색상2	네트워크	라벨
1	O	X	O	X	O	O
2	O	X	O	X	O	X
3	O	X	O	X	X	O
4	O	X	O	X	X	X
5	O	X	X	O	O	O
6	O	X	X	O	O	X
7	O	X	X	O	X	O
8	O	X	X	O	X	X
9	X	O	O	X	O	O
10	X	O	O	X	O	X
11	X	O	O	X	X	O
12	X	O	O	X	X	X
13	X	O	X	O	O	O
14	X	O	X	O	O	X
15	X	O	X	O	X	O
16	X	O	X	O	X	X

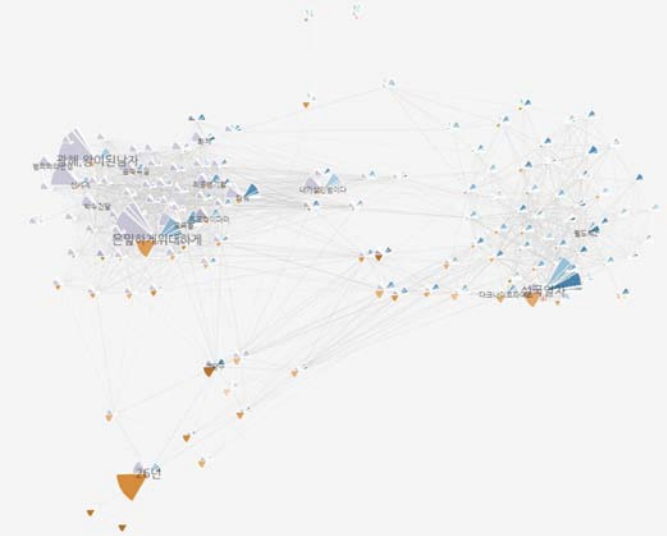
실험 방법은 피험자 내 설계(within subject design)로써 각 실험자에게 16가지 다른 형태의 시각화를 보여주고 16가지의 시각화에 대하여 분석 용이성 값을 최소 1에서 최대 10으로 주는 방법을 취하였다. 피험자 내 설계란 독립변인의 모든 수준에 피험자들을 할당하는 실험 설계 방법으로 본 연구에서는 피험자를 16가지 각기 다른 시각화를 사용하고 분석 용이성 점수를 주게 함으로써 실험 횟수는 16 * 30 으로 총 480번 진행되었다. 측정된 16가지 시각화에 따른 분석 용이성에 대한 기술통계 결과 값은 표 13 과 같다.

〈표 13〉 집단에 따른 분석 용이성

집단	평균	표준편차	순위
1	8.1	1.100505	2
2	5.9	1.370320	12
3	7.2	1.032796	7
4	4.9	1.523884	15
5	7.8	1.2292373	5
6	6.1	1.449138	11
7	6.7	1.059350	8
8	4.4	1.577621	16
9	8.6	1.074968	1
10	6.7	1.159502	8
11	8.0	1.054093	4
12	5.5	1.900292	13
13	8.2	1.032796	3

14	6.4	1.264911	10
15	7.4	1.349897	6
16	5.4	1.955050	14

16가지 시각화에 따른 분석 용이성을 확인한 결과 단일 대상에 대한 오피니언 결과를 부채꼴 모양의 원 그래프의 형태로 나타내는 폴라 그래프 형태의 노드, 6개 범주의 색상, 네트워크 보이기, 라벨 보이기의 시각화 형태가 8.6으로 가장 용이성이 높았고, 그 다음으로 원 형태의 노드, 6개 범주의 색상, 네트워크 보이기, 라벨 보이기의 시각화 형태가 8.1 의 용이성을 보였다. 16가지 시각화 모습에 따른 집단 중 제일 분석용이성이 높게 나온 9번 집단은 그림 20 과 같다.



〈그림 20〉 가장 높은 용이성을 보인 시각화의 형태

4. 결론

본 연구는 제품에 대한 전체 소비자들의 의견을 종합하여 보여주는 시각화 방법을 제안하는데 연구 목적을 두고 진행하였다. 이를 위해 소비자들의 의견을 대변하는 사례 연구로 영화 리뷰를 분석하기 위한 온톨로지를 구축하였다. 시각화 단계에서는 기존의 시각화 방법을 참고하여 단일 대상에 대한 오피니언 마이닝 분석뿐만 아니라 유사한 평가를 받은 집단의 군집 시각화를 통해 다른 대상과 비교분석 할 수 있는 새로운 방법의 시각화 방안을 제안하고 사용자 분석을 통해 시각화의 분석용이성을 측정하였다. 본 연구의 종합적인 결과와 의의는 다음과 같다.

첫째, 개별 영화에 대한 시각화 분석을 시행하여 관객의 종합적인 의견이 서로 상이하다는 점과 관객의 의견에 따라 분석에 사용된 28개의 영화는 크게 3개의 집단으로 군집화 된다는 사실을 확인하였다. 군집된 첫 번째 집단의 경우 분류된 리뷰 온톨로지가 스토리(0.438), 영화(0.213), 배우(0.149), 감독(0.114), 영상(0.083), 음향(0.002)의 순서인 집단이었고, 두 번째 집단의 경우 분류된 리뷰 온톨로지가 영화(0.483), 배우(0.242), 스토리(0.192), 감독(0.054), 영상(0.021), 음향(0.007)의 순서인 집단이었으며 세 번째 집단의 경우 분류된 리뷰 온톨로지가 배우(0.577), 스토리(0.228), 영화(0.129), 감독(0.031), 영상(0.029), 음향(0.004)의 순서인 집단이었다. 또한 군집 시각화를 활용하면 개별 영화와 유사한 영화 집단을 확인 할 수 있다는 결과를 도출하였다.

둘째, 시각화 기능에 따른 분석용이성 분석을 위해 사용자 분석을 시행한 결과, 16개의 집단 중 분석용이성이 가장 높게 나온 집단은 9번 집단으로 폴라그래프 형태의 노드, 6개 범주의 색상, 네트워크 보이기, 라벨 보이기의 시각화 형태가 분석이 가장 용이하다는 것을 도출하였다.

셋째, 기존의 연구에서는 일반적으로 사용되는 온톨로지를 활용하여 데이터를 분석하는 연구가 대부

분이었다. 또한 오피니언 마이닝의 경우 명사 분류만을 사용한 연구들이 대부분이었다. 하지만 본 연구에서는 기존에 구축된 온톨로지를 활용하지 않고 키워드 추출과 토픽모델링을 활용하여 영화 도메인에 대한 온톨로지를 구축하였다. 또한 명사 분류와 더불어 본 연구만의 논항구조 파악 방법을 오피니언 마이닝에 적용하였다는데 의의가 있다.

다만 본 연구에서 구축한 온톨로지가 기 구축된 온톨로지에 비해 오피니언 마이닝 분석의 정확도 향상에 도움이 되는지 측정하지 못했다는 점과 실험적인 규모의 서술어 사전과 구문분석기의 부재로 인한 낮은 재현율, 마지막으로 대규모의 모집 군을 대상으로 분석용이성 실험을 진행하지 못한 점은 본 연구의 한계점이라 할 수 있다.

이러한 한계점에도 불구하고 본 연구 결과와 시각화를 통해 온톨로지 시각화를 활용하여 새로운 관점으로 오피니언 마이닝을 수행할 수 있다는 가능성을 제시하였으며 앞으로 국내 오피니언 마이닝 방법론과 오피니언 마이닝 시각화에 대한 연구에 도움이 될 수 있기를 바란다.

참고문헌

논문

- Anaïs Cadilhac, Farah Benamara, Nathalie Aussenac-Gilles, “Ontolexical resources for feature based opinion mining : a case-study”, Proceedings of the 6th Workshop on Ontologies and Lexical Resources(Ontolex 2010), pp.77-86, 2010.
- Bing Liu, Minqing Hu, Junsheng Cheng, “Opinion Observer: Analyzing and Comparing Opinions on the Web”, In Proceedings of the 14th international conference on World Wide Web, ACM, pp.342-351, 2005.
- Eivind Bjørkelund, E., Burnett, T. H., & Nørvåg, K., "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, ACM, pp. 229-238, 2012.
- Borth, D., Chen, T., Ji, R., & Chang, S. F, "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content." Proceedings of the 21st ACM international conference on Multimedia. ACM, pp.451-460, 2013.
- Fruchterman, T. M., & Reingold, E. M., “Graph drawing by force-directed placement”. Softw., Pract. Exper., 21(11), pp.1129-1164. 1991.
- Zhuang, L., Jing, F., Zhu, X. Y. (2006, November). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, pp. 43-50, 2006.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent dirichlet allocation”, the Journal of machine Learning research, pp. 993-1022, 2003.
- Gruber T. R., "A translation approach to portable ontology specifications.", Knowledge acquisition, 5(2), pp.199-220, 1993.
- Kim, G. N., Ha, H., On, B. W., Lee, K., & Lee, M. "Bubble heap graphs." Proceedings of IEEE Information Visualization Conference (InfoVis' 13), Atlanta, USA. 2013.
- Larissa A. de Freitas, Renata Vieira, "Ontology-based Feature Level Opinion Mining for Portuguese Reviews", WWW 2013 Companion, ACM, 2013.
- 김석환, 김인규, “온톨로지 기반 전문지식 시각화 시스템 제안 및 구현”, 한국지능정보시스템학회 2011년 춘계학술대회, pp.323-330, 2011.
- 명재석, 이동주, 이상구, “반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템”, 정보과학회논문지: 소프트웨어 및 응용, 35(6), pp.392-402, 2008.
- 양정연, “상품 리뷰 요약에서의 문맥 정보를 이용한 의견 분류 방법”, 정보과학회논문지:데이타베이스, 36(4) pp.254-262, 2009.
- 윤영선, “온라인 리뷰가 온라인 쇼핑행동에 미치는 영향”, 국제회계연구, 52, pp.139-159, 2013.
- 이윤정, 지정훈, 우균, 조환규, “인터넷 게시물의 댓글 분석 및 시각화”, 한국콘텐츠학회논문지,

9(7), pp.45-56, 2009.
• 이은영, “온라인 구전이 소비자의 신제품 수용에 미치는 영향에 관한 연구”, 광고홍보연구, 16, pp.1-21, 2008.
• 서진욱, “인간-컴퓨터 상호작용과 정보가시화”, 정보과학회지, 29(8), pp.63-68, 2011
• 조하나, 정연오, 이재동, 이지형, “인터넷 뉴스 댓글의 감성 분석을 통한 오피니언 마이닝”, 한국지능시스템학회 2013년도 춘계학술대회 학술발표논문집, 23(1), pp.149-150, 2013.

도서

• Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
• Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.), Readings in information visualization: using vision to think, Morgan Kaufmann, 1999.
• Nahum Gershon, Stephen G. Eick, Stuart Card, Information Visualization, ACM, 1998.

인터넷 사이트

• NAVER 영화, http://movie.naver.com
• Social Metrics, http://insight.some.co.kr
• Opinion Analyzer,, 54.255.190.140/index/#