# 오피니언 마이닝을 이용한 상품평 분석

송준석*, 조경수, 김응모
*성균관대학교 정보통신공학부
e-mail : songjjunseok@naver.com

# Review Analysis by using the Opinion Mining Techniques

Jun Seok Song*, Kyung Soo Cho, Ung-mo Kim
School of Information and Communication Engineering,
Sungkyunkwan University

요 약

인터넷 시장이 빠르게 성장함에 따라 사용자들의 참여도가 매우 높아졌다. 인터넷 사용자들은 인터넷 쇼핑의 상품에 관한 의견을 웹 상에 표현하기 시작했고, 실제 소비자이 판단하는 데에 많은 영향을 미치고 있다. 하지만 현재에 들어 그 양이 엄청나게 방대해 졌기 때문에 사용자들이 원하는 정보만을 찾아내는 것은 어려운 일이다. 본 논문에서는 사용들이 작성한 인터넷 쇼핑에서 상품평에 관한 리뷰를 모아 방대한 양에서 오피니언 마이닝 기법을 이용해 유용한 정보를 효율적으로 도출해서 사용자가 원하는 정보를 요약하여 제공하는 방법을 제안한다. 이러한 방법을 통해서 사용자는 상품을 구매하기 전에 좀 더 객관적이고 효율적으로 판단을 내릴 수 있을 것이다.

## 1. 서론

최근 온라인쇼핑업계는 지속적인 성장을 거듭해 오는 2015년 44조원의 시장규모를 형성, 가장 큰 유통채널로 자리매김할 것이라는 예측이 나오고 있다.[1] 이와 더불어 사용자의 참여도가 매우 높아진 시대에 웹 상에 상품평이 크게 증가하고 있다. 그리고 사용자가 작성한 상품평은 다양한 활용성을 갖는 가치 있는 데이터이다. 특히 인터넷 쇼핑몰에서의 상품평은 사용자의 구매 결정에 직접적인 영향을 미치는 중요한 정보이고 구매자들의 상품평은 다른 잠재적인 소비자들의 상품 구입을 이끌어내는데 큰 동기가 된다. 하지만 온라인 쇼핑몰에서는 상품평의 성질에 부합하는 순위를 부여하지 않기 때문에, 사용자가 구입 결정을 위하여 수많은 상품평에 포함된 의견들을 효과적으로 검토하기는 쉽지 않다. 일반적으로 상품평은 감정적이며 주관적인 의견을 포함하고 있고 사용자들은 수많은 상품평들 가운데 그들에게 유용한 정보만을 가려내서 판단하는 것이 어려워졌다. 또 많은 상품평들이 필요 이상으로 길게 작성되어 있어 다 읽어 보기 힘들고 요점을 파악하기도 어려워졌다. 이러한 이유로 상품평을 분석하고 요약하는 기법이 필요하게 되었다.

본 논문에서는 오피니언 마이닝 기법을 이용해 인터넷 쇼핑몰에서 상품평들을 분석하여 사용자가 원하는 정보를 손쉽게 얻을 수 있고 상품을 구매하기 전에 객관적인 정보를 효율적으로 이용할 수 있도록 해서 좀 더 나은 구매를 할 수 있도록 하는 것이 목표이다.

본 논문은 다음과 같이 구성된다. 2장에서 오피니언 마이닝 기법과 기존의 대표적인 연구에 대해 설명한다. 3장에서는 본 논문에서 제안하는 상품평에 대한 한국어 오피니언 마이닝 기법에 대해서 설명하고, 4장에서 결론과 발전된 연구를 위한 향후 연구 과제를 제시하며 논문을 마친다.

## 2. 관련연구

### 2.1 오피니언 마이닝

오피니언 마이닝은 리뷰 데이터와 같은 대량의 정보 속에서 유용한 정보를 찾아내는 것이라는 특징을 가지고 있다. 이러한 오피니언 마이닝은 최근에 들어 활발히 연구되어 왔으며, 그 기반이 되는 기술은 자연어처리, 텍스트 마이닝, 통계 등의 분야로부터 기원하였다. 오피니언 마이닝은 대체로 크게 특징 추출, 의견분류, 요약 및 표현 등의 3가지 단계로 이루어진다. 첫째, 특징 추출단계는 유용한 정보라고 판단되는 여러 특징들을 정의하고 추출해 내는 단계이다. 이때 단순히 특징만을 추출하는 것이 아니라 해당 특징이 어떤 의미를 가지는가에 대한 의견을 나타내는 어휘정보도 함께 추출된다. 둘째, 의견 분류 단계에서는 추출된 특징과 의견을 나타내는 어휘가 해당 정보소스에서 어떤 의미로 사용되었는가에 대한 판단 및 분류를 하는 단계이다. 셋째, 요약 및 표현단계에서는 의견 성향이 밝혀진 의견정보들을 요약하여 전체 정보의 내용을 효율적으로 사용자에게 전달하는 단계이다. 이와 같은 오피니언 마이닝의 단계별 수행을 위한 다양한 방식들이 연구되고 있으며, 자연어처리기술 기반에서 통계적 기법에 이르기까지 여러 분야의 기술이 접목되고 있다.[2-5]

오피니언 마이닝은 다음과 같이 나눠볼 수 있다. 우선, Facts and opinions 이라고 하는 것은 현재 세상의 정보를 찾는 방법은 사실(facts)들의 나열 즉, 수많은 문서들을

키워드들의 연관도 순으로 정렬하는 방식이다. 의견은 이러한 키워드들의 관계만을 통해서 표현되기 어려우며, 요약되거나 통합되지 않는다면, 전달되기 힘든 개념이다.

오피니언 마이닝 추상화는 대상물을 추상화하는 과정이며, Basic components 이다. 그리고 Opinion holder는 Opinion을 가진 Person 혹은Organization을 뜻하고 Object는 의견이 표현된 대상이다. 또 Opinion은 외관, 태도 또는 의견이다.

Object와 entry는 Components, Sub-components 의 Hierachy 로 표현되고 각 Component는 attributes를 가지고 있다. 이러한 Component와 Attributes를 통칭하여 Features 라고 표현한다.

Model of review란 Opinion holder가 Positive, Negative 혹은 Neutral 한 의견을 Features 의 sub-set에 코멘팅 하는 과정을 통해서 opinion이 생성된다.

오피니언 마이닝 업무는 Document level은 리뷰의 감성적인 분류를 수행하고 Positive, Negative,
Neutral 로 분류된다. Sentence level은 주관적인 것과 객관적인 분류를 수행하고 Objective, Subjective
으로 분류된다. Feature level은 리뷰어의 코멘트 등을 통해 자질을 구분하고 Positive, Negative,
Neutral 으로 나뉜다.

사전에 기초한 Opinion words 에는 긍정(Positive)을 의미하는 것으로 beautiful, wonderful, good, amazing 등이 있으며, 부정(Negative)을 의미하는 것으로는 bad, poor, terrible 등이 있다. Context 에 독립적인 단어 (ex_ good) 와 그렇지 못한 단어 (small, long ..) 등을 구분하여 저장할 필요가 있다.
small, long 등의 단어는 때로는 positive 하게 때로는 negative 하게 쓰일 수 있다는 의미다.

Corpus-based approarches는 대량의 코퍼스 내에서 구문적인 규칙을 습득 및 공기관계(co-occurrence)와 패턴 추출을 통하여 도메인 독립적인 의견을 학습할 수가 있다. Conjunction의 특징은 유사한 의미를 지니는 특성을 가지므로 이러한 정보도 동시에 사용할 수 있다. 정답 셋을 가지고 있다면 기계학습을 통하여 해당 단어들의 가중치를 추출한다거나 정답에 미치는 영향 등을 파악하여 도메인에 독립적인지 그렇지 못한 지를 판별할 수 있다.

Approarch는 한 문장내에 Conjunction을 포함하는 경우, 한 문장내에 Conjunction이 포함되지 않더라도 유사한 패턴이 발견되는 경우 그리고 한 문장내가 아니더라도 연결된 문장의 경우 Conjunction Rule 을 적용한다.

Handling of many constructs으로는 "a good deal of"는 good이 의견에 사용되지 않은 경우이고 "not only ..."는 not 이 부정에 사용되지 않은 경우이다. 그리고 "... but also"는 but 이 반대의견에 사용되지 않은 경우이다.[6]

### 2.2 형태소분석기

형태소 분석이란 자연 언어 분석의 첫 단계로서 단어(한국어의 경우 어절)를 구성하는 각각의 형태소들을 인식하고 불규칙 활용이나 축약, 탈락 현상이 일어난 경우 원형을 복원하는 과정을 말한다.

문자열 유형의 파악하는 방법으로는 형태소 분석기는 우선 공백을 기준으로 텍스트를 나눠 배열에 담는다. 이를 쉽게 '어절'이라 부른다. 위에서 언급한 대로 어절의 문자열 유형을 파악한다. 그런 후 일정 기준에 따라 색인어로서의 어절을 분리한다. 한글을 포함한 어절일 경우 본격적인 형태소 분석 과정에 들어간다.

전자사전의 구성 및 탐색 기능이란 형태소 분석을 위해서 국내의 상용 제품들은 크게 세가지 방식을 사용한다. 하나는 모든 형태소 어휘와 일정 규칙을 사전에 담아 놓고 이를 탐색하여 결과를 반환하는 '사전 베이스', 별도의 전자사전 없이 무수한 오토마타 등 규칙의 연산을 통해 결과를 반환하는 '규칙 베이스', 그리고 이를 적절히 조화한 '절충형'이다.

규칙의 연산이란 특정 어절이 체언이라면 이는 명사, 수사, 또는 대명사와 기능어인 조사의 결합일 것이다. 만일 동사라면 어미와 동사 형태소의 결합으로 이루어져 있을 것이다. 이러한 판단을 하기 위한 특정 조건의 규칙이 필요하며, 최소한의 연산으로 속도 향상을 꾀해야 한다.

복합명사 처리하는 방법으로는 한글 명사, 특히 색인어로서 가치가 있는 명사는 주로 고유명사와 복합명사이다. 특히 고유명사 중에는 유난히 복합명사가 많다. 이를 단일명사로 하나 하나 분리해 내야만 탁월한 성능의 색인어 추출기가 될 수 있다.[7]

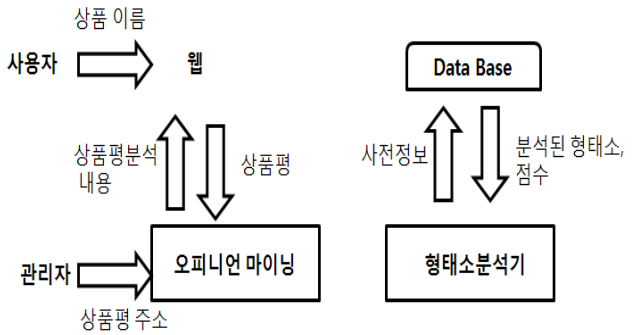
## 3. 오피니언마이닝을 이용한 상품평 분석과 요약

### 3.1 제안 배경

기존의 상품평은 상품평에 대해서 정확한 정보를 얻을 수 없었다. 그러나 오피니언 마이닝을 이용한 상품평 분석과 요약하는 방법은 고객이 직접 쓴 상품평에서 단어를 추출함으로 써 상품 평가에 대한 신뢰도를 높일 수 있다. 또한 편리한 시스템을 요구하는 사용자들의 요구를 충족시키고 기업들은 상품 판매 및 홍보 전략으로 이용할 수도 있다.

### 3.2 제안 내용

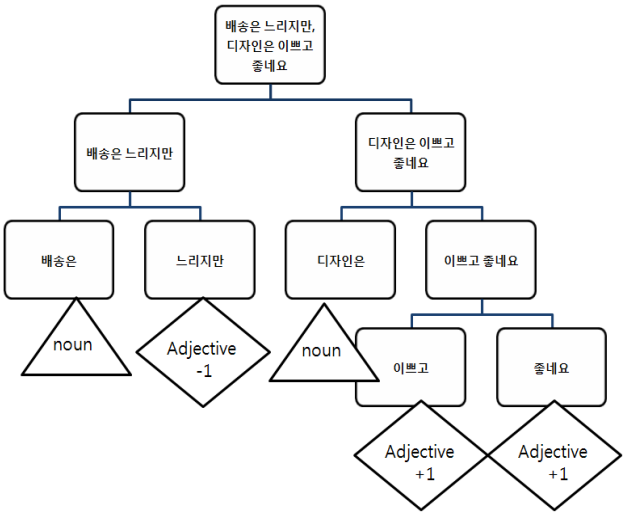
본 논문에서는 우선 시스템구성으로는 (그림1) 에서 사용자가 웹으로 상품 이름을 검색하면 관리자는 상품평 주소를 넘겨 인터넷 쇼핑의 상품 정보와 상품평을 가져와서 오피니언 마이닝을 통해 데이터 베이스의 정보를 참조[8]하여 형태소 분석기로 분석된 형태소와 점수를 데이터베이스에 넣고 분석된 모든 내용을 웹페이지로 보여준다.




(그림 1) 시스템 구성

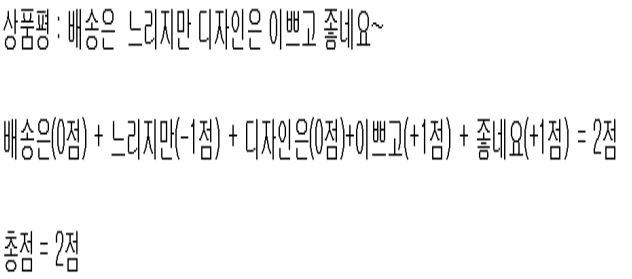
이때, 데이터에 흐름을 비춰보면 웹에서 상품평의 주소를 보내고 해당 주소에서 상품에 대한 상품평 정보를 가져와서 데이터베이스의 사전정보를 참조하여 형태소 분석기로 상품평을 분석한다. 그 다음 분석된 상품평을  평가하고 평가결과를 데이터베이스에 저장하고 웹에서 데이터 베이스의 평가된 상품평을 보여주게 된다.

그리고 형태소 분석기에서 상품평을 추출하고 추출한 상품평 문장에 대한 형태소 분석을 한 후에 감정을 표현할 수 있는 가중치 단어 추출한다. 평가방법으로는 (그림2)에서처럼 평가부분에서는 분석알고리즘을 통해 도출된 총점으로 0점이 기준이 되어 긍정, 중립, 부정으로 나누고 가중치 단어 추출시 가중 평가 처리를 한다. 예를 들어 '디자인이 대단히 좋아요'를 보면 '좋아요' 라는 표현에 '대단히' 라는 부사가 붙어 '좋아요' 라는 표현을 가중시키므로 평가점수도 가중처리를 한다.[9]




(그림 2) Evaluation

마지막으로 웹페이지에 보여줄 때는 먼저 평가부분이 긍정, 중립, 부정으로 되어있는 것을 보여주고 그림(3) 에서처럼 상품평 문장이 출력되고, 그 밑으로 이 상품이 왜 긍정 상품평으로 평가되었는지 계산한 결과가 화면에 출력된다.




(그림 3) 웹 페이지 표시

## 4. 결론 및 향후 연구

본 논문에서는 인터넷 쇼핑에서의 상품평 분석 및 요약하는데, 오피니언 마이닝 기법을 적용하여, 좀 더 효율적이고 빠르게 상품평을 확인 할 수 있는 방법을 제안하였고 상품평을 분석하여 요약제시 함으로써 먼저 경험이 있는 소비자의 평가를 통해 사용자가 객관적인 판단을 할 수 있도록 도와준다. 향후 과제로는 한글 문장에서 오피니언의 판단이 불가능한 문장을 분석 가능하도록 하는 연구와 함께 웹에서 많이 쓰이는 인터넷 용어와 신조어를 추출하여 그것의 사전적 의미뿐만 아니라 문맥적 의미를 분석할 수 있는 기술의 개발이 필요하다. 또한 상품평 점수에 상관이 없고, 유용하지 않은 정보를 걸러내는 기술의 연구가 필요할 것이다.

## 감사의 글

이 논문은 2009 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075771)

## 참고문헌

[1] '2010년 유통산업 전망', 대한상의보고서

[2]Namrata Godbole, Manjunath Srinivasaiah, StevenSkiena, "Large-Scale Sentiment Analysis for News and Blogs," Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2007), 2007.

[3] E. Boiy, P. Hens, K. Deschacht, M. Moens, "Automatic Sentiment Analysis in On-line Text," ELPUB2007 Conference on Electronic Publishing, June 2007.

[4] J. Yi, W. Niblack, "Sentiment Mining in Web-Fountain," International Conference on Data Engineering (ICDE'05), pp.1073-1083, 2005.

[5] T. Nasukawa, J. Yi, "Sentiment analysis: capturing favorability using natural language processing," Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture, pp.70-77, 2003.

[6] A Holistic Lexicon-Based Approach to Opinion Mining by Bing Liu

[7] 강승식 지음, “한국어 형태소 분석과 정보검색”, 홍릉과학출판사, 2002
[8] Lipika Dey. S K Mirajul Haque ″Opinion Mining from Noisy Text Data″, in Proceedings of the 2nd workshop on Analytics for noisy unstructured text data, pages 83-90, 2008.
[9] Minqing Hu, Bing Liu ″Mining and Summarizing Customer Reviews″, in Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168-177, 2004.