



## 텍스트 마이닝 기반의 온라인 상품 리뷰 추출을 통한 목적별 맞춤형 정보 도출 방법론 연구

A Study on the Method for Extracting the Purpose-Specific Customized Information from Online Product Reviews based on Text Mining

---

저자 (Authors)	김주영, 김동수 Joo Young Kim, Dong soo Kim
출처 (Source)	<a href="#">한국전자거래학회지 21(2)</a> , 2016.5, 151-161 (11 pages) <a href="#">The Journal of Society for e-Business Studies 21(2)</a> , 2016.5, 151-161 (11 pages)
발행처 (Publisher)	<a href="#">한국전자거래학회</a> Society for e-Business Studies
URL	<a href="http://www.dbpia.co.kr/Article/NODE06687695">http://www.dbpia.co.kr/Article/NODE06687695</a>
APA Style	김주영, 김동수 (2016). 텍스트 마이닝 기반의 온라인 상품 리뷰 추출을 통한 목적별 맞춤형 정보 도출 방법론 연구. 한국전자거래학회지, 21(2), 151-161.
이용정보 (Accessed)	성균관대학교 자연과학캠퍼스 115.***.238.89 2018/09/17 16:41 (KST)

---

### 저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

### Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s) and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s) for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

# 텍스트 마이닝 기반의 온라인 상품 리뷰 추출을 통한 목적별 맞춤화 정보 도출 방법론 연구

## A Study on the Method for Extracting the Purpose-Specific Customized Information from Online Product Reviews based on Text Mining

김주영(Joo Young Kim)\*, 김동수(Dong soo Kim)\*\*

### 초 록

개방, 공유, 참여를 특징으로 하는 웹 2.0 시대로 들어서면서 인터넷 사용자들의 데이터 생산 및 공유가 쉬워졌다. 이에 따른 데이터의 기하급수적인 증가와 함께 디지털 정보의 대부분인 비정형적 데이터(Unstructured Data)의 양도 증가하고 있다. 인터넷에서 정해진 형식 없이 자연어 형태로 만들어진 비정형 데이터 중, 특정 상품들에 대해 개인이 평가한 리뷰들은 해당 기업이나 해당 상품에 관심이 있는 잠재적 고객에게 필요한 데이터이다. 많은 양의 리뷰 데이터에서 상품에 대한 유용한 정보를 얻기 위해서는 데이터 수집, 저장, 전처리, 분석, 및 결론 도출의 과정이 필요하다. 따라서 본 연구는 R을 이용한 텍스트 마이닝(Text Mining) 기법을 사용하여 텍스트 형식의 비정형 데이터에서 자연어 처리 기술 및 문서 처리 기술을 적용하여 정형화된 데이터 값을 도출하는 방법에 대해 소개한다. 또한, 도출된 정형화된 리뷰 정보를 데이터 마이닝 기법에 적용하여 목적에 맞게 맞춤화된 리뷰 정보를 도출시키는 방안을 제시하고자 한다.

### ABSTRACT

In the era of the Web 2.0, characterized by the openness, sharing and participation, it is easy for internet users to produce and share the data. The amount of the unstructured data which occupies most of the digital world's data has increased exponentially. One of the kinds of the unstructured data called personal online product reviews is necessary for both the company that produces those products and the potential customers who are interested in those products. In order to extract useful information from lots of scattered review data, the process of collecting data, storing, preprocessing, analyzing, and drawing a conclusion is needed. Therefore we introduce the text-mining methodology for applying the natural language process technology to the text format data like product review in order to carry out extracting structured data by using R programming. Also, we introduce the data-mining to derive the purpose-specific customized information from the structured review information drawn by the text-mining.

**키워드 :** 텍스트마이닝, 데이터마이닝, 상품 리뷰, R 프로그래밍  
Text-mining, Data-mining, Product Review, R Programming

\* Dept of Industrial and Information Systems Engineering, Soongsil University(whytimesgone@ssu.ac.kr)

\*\* Corresponding Author, Dept of Industrial and Information Systems Engineering, Soongsil University (dskim@ssu.ac.kr)

Received: 2016-04-18, Review completed: 2016-05-16, Accepted: 2016-05-23

## 1. 서 론

닷컴 붐과 이후, 웹 2.0 시대가 도래됨에 따라 사용자가 적극적으로 참여하여 정보와 지식을 만들고 타인과 산출물을 공유하는 열린 인터넷의 시대가 시작되었다. 사용자들은 개인이 가지고 있는 정보나 콘텐츠를 인터넷에 제공할 수 있게 되었다. 이로 인해 해당 내용에 관심이 있는 개인에게는 새로운 정보를 얻을 수 있는 기회를, 기업에게는 자사 상품들에 대한 고객들의 견해와 관련된 텍스트 데이터를 데이터베이스에 구축할 수 있는 방법 중 하나인 집단 지성(Collective Intelligence)이 탄생하게 되었다. 집단 지성이란 많은 사람들로 부터 산출된 데이터를 협력 혹은 경쟁을 적용하여 나온 결과물을 의미한다[20]. 이는 전문가 개인이 제공하는 정보보다 다수의 데이터로 산출된 정보가 올바른 결론에 가깝다고 할 수 있다.

과거에 기업들은 자사 상품들에 대한 고객들의 의견, 평가와 같은 피드백을 받기 위해서 온라인이나 오프라인에서 설문 조사를 수행하였다. 하지만 이와 같은 방식은 시간과 비용이 많이 들고 고객의 응답 의지에 따라 데이터의 결과의 질이 달라질 수밖에 없었다[7].

인터넷에서 발생하는 데이터가 기하급수적으로 늘어남에 따라, 디지털 정보의 대부분을 차지하고 있는 비정형적 데이터(Unstructured Data)의 양도 증가하고 있다. 많은 기업들은 방대한 양의 데이터를 통해 비즈니스 현황을 파악하고 고객에 대한 이해를 높이기 위해 데이터 마이닝(Data Mining)을 이용하고 있다[2, 14]. 기업들은 이런 많은 양의 데이터를 정형화시키고 모델링함으로써 연관 관계(Association Relationship) 파악, 이상 현상(Anomaly) 탐지, 전

반적 평가와 같은 유용한 정보를 도출시킨 후, 이를 바탕으로 의사결정 활동을 수행한다[15].

인터넷에 산재한 다양한 비정형적인 데이터들 중, 특정 상품들에 대해 개인이 평가한 온라인 리뷰들은 해당 기업에 의미 있는 정보를 제공할 뿐만 아니라 상품에 관심이 있는 잠재적 고객에게도 필요한 정보를 제공해준다[1]. 잠재적 고객들에겐 온라인 리뷰가 구매를 결정하는 요소 중 하나로 작용한다[12]. 즉, 리뷰를 통한 구전 마케팅(Word-of-Mouth Marketing)도 타인에게는 상당한 영향을 미친다[11].

하지만 인터넷에 존재하는 대부분의 리뷰들은 자연어 형태(Natural Language)로 작성된 비정형적, 반정형적 텍스트 형태이다[3, 4]. 이는 텍스트 마이닝(Text Mining) 기법 사용 없이는 정해진 스키마가 없는 비정형 데이터에서 의미 있는 정보를 추출하기 어려움을 의미한다. 또한, 사용자들의 수동적, 능동적인 참여에 따라 방대한 양의 리뷰들이 존재하기 때문에 기업이나 불특정 개인이 필요로 하는 가치 있는 정보를 찾기 위해서는 온라인 상품 리뷰들을 취합하고 가공하는 분석 기술을 요한다.

많은 양의 정형화된 데이터에서 정보를 추출하여 가공하는 데이터 마이닝은 이미 많은 분야에서 사용되고 있다[19]. 해당 연구에서는 비정형적인 리뷰 데이터를 유용한 정보로 가공하는 방안을 제시하는 것에 초점을 두고 연구하였다. 따라서 본 연구는 R 프로그램을 바탕으로 한 텍스트 마이닝을 이용하여 비정형적인 상품 리뷰 데이터를 자연어 처리 기술 및 문서 처리 기술을 적용하여 이를 정형화된 데이터로 정제하는 방법에 대해 설명한다. 그 후, 텍스트 마이닝을 통해 추출한 데이터로 이루어진 리뷰 데이터셋(Review Dataset)의 속성

값(Attribute)들과 속성 값을 구성하는 요소(Element)들에 데이터 마이닝을 이용하여 목적에 맞춤화된 리뷰 정보를 도출시키는 방법에 대해 소개하고자 한다.

## 2. 관련 연구

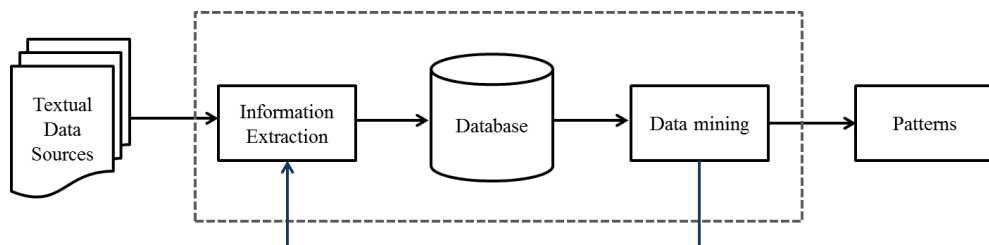
웹 2.0 시대에서는 개인이 쉽게 정보를 생산할 수 있다는 장점 때문에, 개인이 특정 상품을 평가한 온라인 리뷰의 수는 데이터의 증가만큼이나 빠르게 증가하였다[17]. 이로 인해 구전 마케팅으로 자리매김한 리뷰 데이터는 특정 상품에 관심이 있는 잠재적 고객의 구매 의사결정에 큰 요소가 되었고, 해당 상품을 생산하는 기업에게는 고객의 의견을 습득할 수 있게 되었다[5].

현재는 의미론적인 탐색(Semantic Search)라는 특성이 포함된 웹 시대로 도입하는 과도기이다[13]. 이는 단순한 집단 지성으로 산출된 데이터의 나열을 넘어 데이터를 이용하여 의미와 규칙 정보 발견을 목표로 한다는 특징을 가지고 있다. 하지만 많은 곳에 산재해 있는 온라인 리뷰 데이터는 그 자체만으로는 의미 있는 정보를 얻기 힘들다. 이로 인해 잠재적 고객은 단순히 가장 인기 있는 리뷰나 최근의 리뷰를 읽고 구매 여부를 결정한다[10]. 따라서 단순한

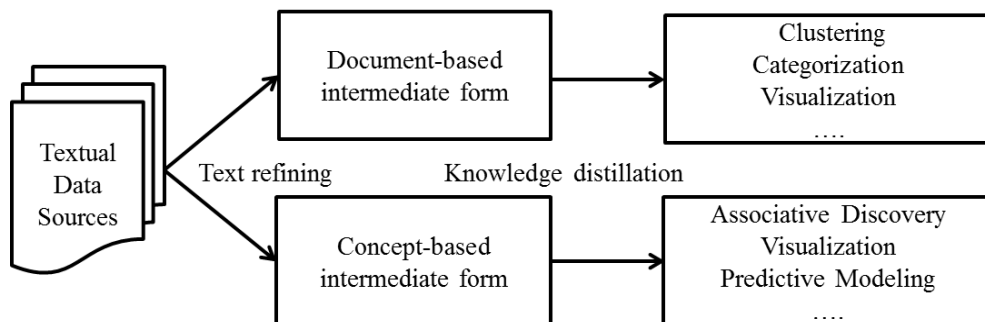
집단 지성으로 산출된 데이터의 나열이 아닌 리뷰 데이터를 가치가 있는 정보로 가공하여 의미와 규칙 정보를 발견하는 것이 필요하다[6]. 리뷰 데이터를 가공하기 위해서 비정형 데이터에 마이닝 기법을 적용한 분석이 필요하다. 기업의 관점에서 분석된 리뷰 데이터는 해당 상품의 전반적인 평가, 장·단점, 및 품질 불량 등을 파악할 수 있다. 또한, 특정 개인의 관점에서는 해당 상품에서 자신이 원하는 정보를 쉽게 얻을 수 있다.

사람이 사용하는 자연어는 문법적, 어휘적인 독특함이 있을 뿐만 아니라 다양한 표현 범위 때문에 일관된 규칙을 찾기 힘들다. 텍스트 마이닝은 이런 자연어 처리를 기반으로 한 기술이다. 자연어 처리 기술이란 사람이 사용하는 언어를 분석하고 그 구조와 의미를 파악하는 것을 뜻한다. 즉, 텍스트 마이닝이란 비정형 텍스트 문서들에서 관심이 있거나 가치가 있는 패턴 혹은 지식을 추출하는 과정을 뜻한다[8]. 대부분의 비정형 데이터가 텍스트로 이루어져 있기 때문에, 텍스트 마이닝은 다양한 목적으로 사용 된다[9].

<Figure 1>은 텍스트 마이닝을 통해 텍스트로부터 의미 있는 패턴을 도출시키는 과정을 보여준다. 사용자는 텍스트 마이닝 기술을 통해 방대한 양의 데이터에서 의미 있는 정보



<Figure 1> The Components of Text Mining Framework



〈Figure 2〉 Text Mining Framework on the Basis of Intermediate Form

를 추출해 내고, 다른 정보와의 연계성을 파악한다. 텍스트 마이닝을 통해 컴퓨터가 자연어 처리를 하기 위해서는 대용량 언어 자원과 복잡한 통계적, 규칙적 알고리즘이 적용되어야만 한다.

〈Figure 2〉는 텍스트 마이닝에서 두 개의 중간 형식(Intermediate Form)을 보여준다. 먼저, 문서를 기반으로 한 중간 형식(Document-based IF)의 각 개체(Entity)는 문서를 낸다. 이 중간 형식으로 정보를 정제하여 나올 수 있는 결과물로는 군집화, 시각화, 그리고 범주화가 있다. 두 번째로, 개념을 기반으로 한 중간 형식(Concept-based IF)의 각 개체는 특정 분야에 해당되는 단어들을 뜻한다. 이 중간 형식으로 정보를 정제 후, 나올 수 있는 결과물로는 연관관계 분석, 예측 모델링, 시각화 등이 있다. 본 연구는 핵심어로 추출된 리뷰 단어들 사이에 연관 관계를 분석하는 기법을 방법론에 포함하였다.

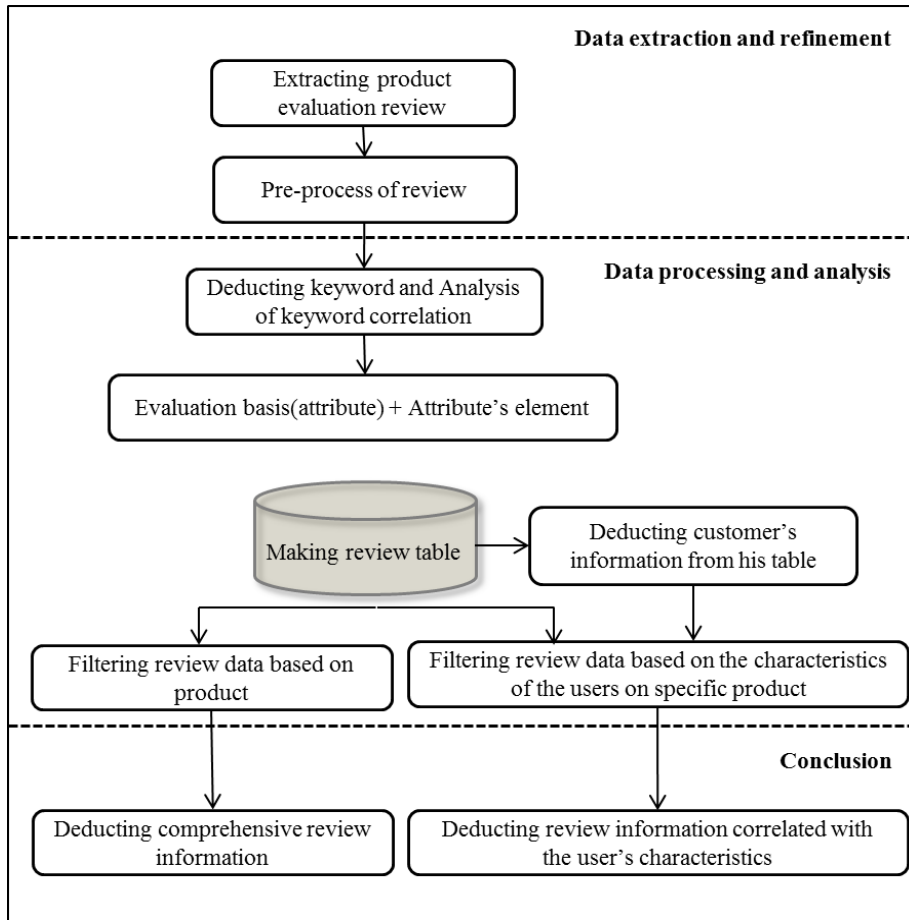
### 3. 리뷰 텍스트 마이닝 프레임워크

본 연구에서는 온라인 쇼핑몰에서 판매되는

특정 상품에 대한 리뷰를 추출하고 정리하여 정형 데이터로 만드는 것을 목표로 한다. 연구의 흐름은 리뷰 데이터 수집 및 저장, 자연어 처리, 키워드 추출, 그리고 자료 분류 및 요약 순으로 진행된다.

〈Figure 3〉은 리뷰 데이터 마이닝 방법론에 대한 아키텍처이다. 먼저 데이터 추출 및 정제 단계에서 상품 평가 리뷰를 추출하고 자연어 처리를 진행한다. 웹 크롤링(Web Crawling)을 이용하여 상품에 대한 온라인 리뷰와 공개된 고객의 정보를 수집(Data Aggregation)한다. 단, 파일의 단위는 하나의 제품에 있는 모든 리뷰로 설정한다. 각 단어의 연관 관계를 분석하기 위한 데이터 파일과 데이터베이스에 저장하기 위한 테이블 형태의 데이터 파일을 구분 지어 리뷰 텍스트를 저장한다. 저장한 텍스트를 통계분석용 소프트웨어인 R 프로그램을 이용하여 비정형적인 리뷰 데이터를 단어 단위로 추출하고 불용어(Stopword)나 의미가 없는 문자 및 기호를 제거하는 전처리(Preprocessing)를 수행한다.

그 후, 데이터 가공 및 분석 단계에서는 먼저, 산업 군에서 평가 요소(Attribute)로 사용되는 값들과 평가 요소를 표현할 수 있는 서술

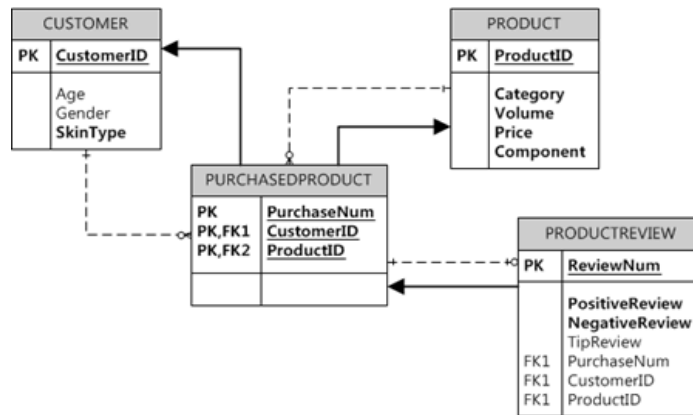


〈Figure 3〉 Framework for the Review Text Mining Methodology

어로 이루어진 형태의 데이터 사전(Data Dictionary)을 정의한다. 데이터 사전은 앞서 전처리를 통해 도출된 단어들을 범주화하여 평가 요소로 설정하는 방법과 산업 군에서 기존에 제품을 평가할 때 고려되는 요소를 평가 요소로 설정하는 방법이 있다.

평가 요소 중, 예상되는 범위에서 서술어가 나타나는 항목의 경우, 상품의 사전 정보를 입력하여 관련된 단어 정보를 추출한다. 그 후, 단어 단위로 추출된 요소들을 군집화를 통해 분류한다. 데이터 사전을 통해 1차적인 분류

이후, 정보 추출을 진행한다. 이는 특정 상품 리뷰들 중, 각 고객 식별자(Customer ID)에서 추출된 키워드에 연관 규칙 학습(Association Rule Learning)을 수행하여 각 단어 간 관계를 파악하여 연관성이 있는 단어들끼리 결합한다 [16]. 그 후, 다시 데이터 사전에 값을 수정하고 다시 해당 알고리즘을 통해 단어들을 결합시킨다. 반복 수행을 통해 보다 분류의 정확도를 높인 후, 군집화된 단어들을 대표하는 용어를 리뷰 데이터 테이블의 열 값에 각 단어들을 분류된 열 아래의 요소 값으로 지정한다.



〈Figure 4〉 Core Tables before Applying Text Mining

ReviewNum	PurchaseNum	ProductID	CustomerID	Ingredient	Effectiveness	SideEffect

〈Figure 5〉 Review Table after Applying Text Mining

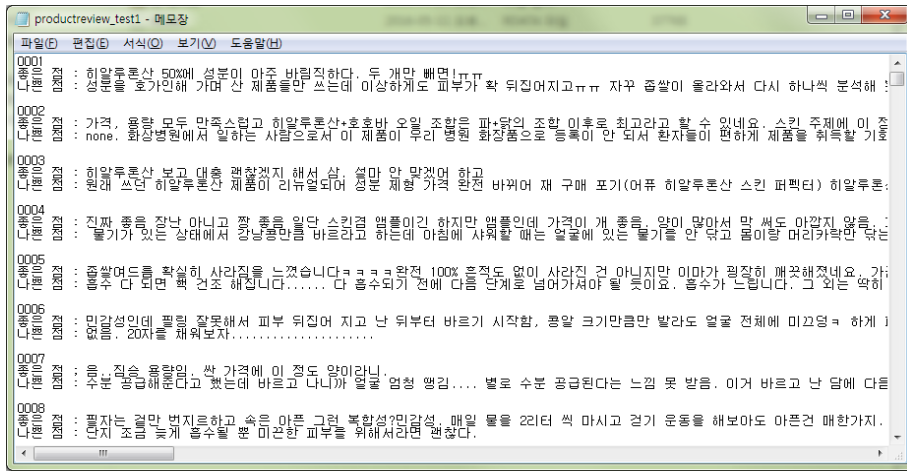
〈Figure 4〉는 텍스트 마이닝을 적용하기 이전 관계형 데이터베이스(Relational Database) 중, 핵심 테이블 간의 관계를 보여준다. 마이닝 이전 테이블에 저장된 상품 평가의 경우, 각각의 값이 단어 단위로 구성되어 있는 것이 아니라 문자열의 나열로 되어 있다. 하지만 데이터 사전 정의를 통한 분류와 단어 간 연관 관계를 파악한 텍스트 데이터는 정형화된 후, 〈Figure 5〉와 같은 형태로 저장된다.

각 열에 들어가는 값들은 군집화한 단어들을 대표하는 용어를 기재하고 테이블의 요소 값에는 해당 집합에 속해 있는 단어들 값을 넣는다. 이 방법론을 통해 정형화된 데이터로 변환 후, 데이터 마이닝 방법론을 적용하면 두 가지 형태의 리뷰 정보를 도출할 수 있다. 첫 번째는 전반적인 상품에 대한 평가로 고객의 특성에 상관없이 단어의 조합을 나타내는 것이

다. 이를 통해 해당 상품을 생산하는 기업은 기업이 제품을 만들 때 구상한 성능과 고객이 체감하는 성능을 비교 분석할 수 있다. 두 번째 리뷰 정보 형태는 고객의 특성에 따라 리뷰 테이블 값을 다시 정제하여 단어의 조합을 나타내는 것이다. 이를 통해 고객이 상품의 리뷰를 관찰할 때 특성이 비슷한 다른 고객들의 리뷰를 정제 및 정리하여 의미 있는 리뷰 정보를 제시할 수 있다.

#### 4. 리뷰 텍스트 마이닝 적용 사례

데이터 마이닝을 적용하여 리뷰에서 통계적 규칙이나 패턴을 찾기 위해서는 데이터 수집이 우선시 되어야한다. 많은 양의 리뷰 데이터를 수집하기 위해서는 웹 크롤러(Web Crawler)



〈Figure 6〉 Result of Crawling the Product Reviews(part)

Effect Data Dictionary	Price Data Dictionary	Ingredient Data Dictionary
Extract nouns from review data	Expensive Inexpensive Cheap Cost-Effectiveness ...	Hyaluronic acid Purified water Aloe Extract ....

〈Figure 7〉 An Example of Cosmetic Data Dictionary(part)

reviewtest2	습수	수분	제품	사용	스킨	성분	얼굴	이거	가격	건조	
피부	140	113	87	82	74	69	64	55	47	45	38
습수		상대	트러블	정도	각질	진짜	꽃물	때문	여드름	화장품	
습수	37	34	31	30	27	27	25	25	24	24	23
화장		시간	토너	세안	보습	세수	이슬	지성	크림	건성	
화장	22	20	20	19	18	18	18	18	18	17	17
습수		처음	부족	무난	민감성	구매	다음	효과	부스터	저렴	해서
습수	17	17	16	15	15	14	14	14	13	13	13
습수		사람	생각	아침	이상	공급					
습수	13	12	12	12	12	11					

〈Figure 8〉 Array of Keywords in Effect Data Dictionary(part)

를 사용하여 한글 텍스트를 뽑아내야한다. 비정형데이터의 경우 정해진 스키마가 없기 때문에 크롤러가 직접 파싱(parsing)해야 된다. 〈Figure 6〉은 R을 통해 상품 리뷰를 추출한 결과를 보여준다. 단어 추출 및 연관 관계 파악을 위하여 아래와 같은 형태로 저장한다. 공개된 고객 정보의 경우, 크롤링 후 정형적 데이터인 테이블 형태로 저장한다.

〈Figure 7〉은 데이터 가공 및 분석 단계에서 정의되어야 하는 데이터 사전의 예제를 보

여준다. 특정 화장품 상품에 대한 데이터 사전을 명시화하였다. 특정 화장품의 성분은 기존에 명시되어 있기 때문에 데이터 사전에 미리 값을 입력한다. 또한, 가격을 나타내는 서술어도 정해져 있기 때문에 데이터 사전에 미리 값을 입력한다. 위의 두 데이터 사전과 관련한 단어를 제한 키워드에서 연관 규칙 학습을 통해 단어 간 연결 관계를 파악한다. 그 후, 다시 데이터 사전을 정의하고 연관 규칙 학습을 반복한다.



이 때, 키워드에서 지지도(support), 신뢰도(confidence) 등을 구하여 단어 간 연관성을 파악한 후, 이를 군집화 시킨다. 군집화로 묶여진 단어 묶음을 대표할 수 있는 용어를 정형화된 리뷰 데이터 테이블의 속성 값에 입력을 하고 각각의 단어들은 해당 속성 값의 요소 칸에 값을 넣는다. 비정형적 형태로 이루어진 리뷰를 데이터 가공과 분석을 통해 정형화된 테이블 값에 넣은 후, 목적에 맞게 필터링하여 리뷰 정보를 도출한다.

이와 같이 도출된 정형화된 데이터를 이용해서 각 단어 간에 연관 관계를 도출하여 하나의 주어에 표현되는 서술어의 빈도를 파악할 수 있다. 예를 들어 ‘효과’라는 주어와 ‘좋다’, ‘나쁘다’라는 서술어의 빈도를 통해 상품에 대한 고객들의 평가를 정량적으로 나타낼 수 있다. 이를 통해 상품의 전체적인 리뷰를 읽지 않아도 타인이 평가한 대체적인 의견을 쉽게 알 수 있다는 장점이 있다. 또한, 각 단어 간의 선·후행 연결 관계를 파악할 수 있어 분석된 결과를 통해 단어의 재조합을 통해 전반적으로 평가된 키워드로 평가를 정리할 수 있다는 장점이 있다.

## 5. 결 론

정보의 홍수라는 단어처럼 웹 2.0 시대의 사용자는 인터넷에서 많은 양의 정보를 쉽게 찾아 볼 수 있다. 여러 사용자들의 인터넷 정보 생산이 쉬워지면서 상품의 리뷰와 관련된 정보도 방대한 양으로 생성되었기 때문이다. 하지만 많은 양의 데이터에서 종합적인 의미를 찾기 위해서는 비정형적인 데이터들을 정제하

는 텍스트 마이닝 기법이 사용되어야 한다. 또한, 전체 데이터를 통한 일반적인 정보 외에도 기업이나 소비자가 목적에 해당하는 맞춤형 정보만 볼 수 있게끔 데이터의 분류 및 군집화가 필요하다[18]. 따라서 본 연구는 목적에 맞춤형 유용한 리뷰 정보를 도출하는 방안 및 특정 제약조건과 연관된 리뷰 정보를 제공하는 방안에 대해 제시하였다. 기업은 정제된 리뷰 정보를 통해 상품의 특성별 고객의 기호도를 파악할 수 있고, 고객의 경우, 자신의 정보와 비슷한 고객들이 남긴 리뷰를 정리하여 보여줌으로써 추천 리뷰 서비스로 확장이 가능할 것이다.

하지만 본 연구에서는 비정형화된 데이터를 정형화된 데이터로 변환할 때의 성능 문제와 각각의 고객 리뷰의 진실성 검증이 어려운 한계점이 있다. 한글 단어의 특징 때문에 R 프로그래밍의 텍스트 마이닝 패키지(tm Package)를 그대로 사용할 수 없다는 점과 기존의 단어추출 패키지의 정확도가 부족하기 때문이다. 특히, 오타나 띄어쓰기가 부족한 리뷰의 경우, 명사 단위로 추출할 때에 정확한 분류가 불가능하다. 또한, 직접 사용자가 제품을 사용한 후, 작성한 후기인지 수동적으로 긍정적, 부정적 단어를 언급한 것인지에 대한 진위 여부를 판단하기 어렵다는 점이 있기 때문이다.

추후 연구에서는 방법론에 제시된 형태로 알고리즘을 구현하고 특정 산업의 리뷰 데이터를 추출하여 방법론에 대한 검증을 수행할 예정이다. 또한, 기존 데이터를 이용하여 파악한 키워드의 연관 관계 및 고객 데이터 속성을 이용하여 고객의 평가를 예측할 수 있는 모델링(Predictive Modeling)도 설립할 것이다.

## References

- [1] Archak, N., Ghose, A., and Ipeirotis, P. G., "Deriving the pricing power of product features by mining consumer reviews," *Management Science*, Vol. 57, No. 8, pp. 1485-1509, 2011.
- [2] Baars, H. and Kemper, H.-G., "Management support with structured and unstructured data-an integrated business intelligence framework," *Information Systems Management*, Vol. 25, No. 2, pp. 132-148, 2008.
- [3] Blumberg, R. and Atre, S., "The problem with unstructured data," *DM Review Magazine*, 2003.
- [4] Buneman, P., "Semistructured data," *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, ACM, 1997.
- [5] Chevalier, J. A. and Mayzlin, D., "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, Vol. 43, No. 3, pp. 345-354, 2006.
- [6] Collins, M., *Head-driven statistical models for natural language parsing*, *Computational linguistics*, Vol. 29, No. 4, pp. 589-637, 2003.
- [7] Decker, R. and Trusov, M., "Estimating aggregate consumer preferences from online product reviews," *International Journal of Research in Marketing*, Vol. 27, No. 4, pp. 293-307, 2010.
- [8] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., "From data mining to knowledge discovery in databases," *AI magazine*, Vol. 17, No. 3, pp. 37-54, 1996.
- [9] Holton, C., "Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem," *Decision Support Systems*, Vol. 46, No. 4, pp. 853-864, 2009.
- [10] Kangale, A., Kumara, S. K., Naeema, M. A., Williamsb, M., and Tiwaria, M. K., "Mining consumer reviews to generate ratings of different product attributes while producing feature-based review-summary," *International Journal of Systems Science*, Vol. 47, No. 13, pp. 1-15, 2016.
- [11] Kozinets, R. V., de Valck, K., Wojnicki, A. C., and Wilner, S. J. S., "Networked narratives: Understanding word-of-mouth marketing in online communities," *Journal of marketing*, Vol. 74, No. 2, pp. 71-89, 2010.
- [12] Lee, J., "How eWOM Reduces Uncertainties in Decision-making Process: Using the Concept of Entropy in Information Theory," *The Journal of Society for e-Business*, Vol. 16, No. 4, pp. 241-256, 2011.
- [13] Mangold, C., "A survey and classification of semantic search approaches," *International Journal of Metadata, Semantics and Ontologies*, Vol. 2, No. 1, pp. 23-34, 2007.
- [14] Mayer-Schönberger, V. and Cukier, K.,

- Big data: A revolution that will transform how we live, work, and think., Houghton Mifflin Harcourt, 2013.
- [15] McAfee, A. and Brynjolfsson, E., "Big data," *The management revolution*, Harvard Bus Rev, Vol. 90, No. 10, pp. 61-67, 2012.
- [16] Mei, Q. and Zhai, C. X., "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, 2005.
- [17] O'reilly, T., "What is Web 2.0: Design patterns and business models for the next generation of software," *Communications and strategies*, No. 1, p. 17, 2007.
- [18] Tan, A.-H., "Text mining: The state of the art and the challenges," *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, pp. 65-70, 1999.
- [19] Washio, T. and H. Motoda., "State of the art of graph-based data mining," *Acm Sigkdd Explorations Newsletter*, Vol. 5, No. 1, pp. 59-68, 2003.
- [20] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W., "Evidence for a collective intelligence factor in the performance of human groups," *science*, Vol. 330, No. 6004, pp. 686-688, 2010.

## 저 자 소 개



김주영

2011년

2015년~현재

관심분야

(E-mail: whytimesgone@ssu.ac.kr)

승실대학교 산업·정보 시스템 공학과 (학사)

승실대학교 산업·정보 시스템 공학과 (석사 과정)

Text Mining, Data Mining



김동수

1994년

1996년

2001년

2001년~2003년

2003년~2006년

2006년~현재

관심분야

(E-mail: dskim@ssu.ac.kr)

서울대학교 산업공학과 (학사)

서울대학교 산업공학과 (석사)

서울대학교 산업공학과 (박사)

한국전산원 전자거래연구부 e-Biz 표준팀장

가톨릭대학교 의료경영대학원 전임강사, 조교수

승실대학교 산업·정보시스템공학과 교수

BPM, 프로세스 마이닝, O2O, e-Health, 정보보호