상품 리뷰 데이터와 감성 분석 처리 모델링

Product Review Data and Sentiment Analytical Processing Modeling

연종흠(Jongheum Yeon)*, 이동주(Dongjoo Lee)**,
심준호(Junho Shim)***, 이상구(Sang-goo Lee)****

초 록

전자 상거래 사이트의 상품 리뷰는 구매 예정자들에게 유용한 정보로 활용될 수 있지만, 방대한 양으로 인해 사용자가 모든 리뷰를 읽는 것은 불가능에 가깝다. 이를 보완하고자 전자 상거래 사이트들은 상품이나 그 특징에 대한 별점 통계, 유용한 리뷰 분류 등을 사용자의 참여나 수작업을 통해 제공하고 있다. 오피니언 마이닝(opinion mining) 혹은 감성 분석(sentiment analysis)은 이러한 일련의 과정을 자동화하는 연구로서, 상품 리뷰의 사용자 의견을 대상으로 그 의견이 긍정적인지, 부정적인지 판단한 후 요약하여 제공한다. 하지만 기존의 감성 분석은 구매예정자에게 유용한 정보, 즉 상품평의 극성을 판별하거나, 상품 특징별 평가 요약 등에만 초점을 맞추고 있어, 상대적으로 의견 정보의 활용도가 낮아지는 문제가 있다. 실제 상품 리뷰에는 상품의 평가 외에도 제품이 가지고 있는 문제점, 고객의 불만 등이 제시되어 있으며, 이를 관리자가 효과적으로 분석하여 의사 결정에 지원에 활용하고자 하는 요구가 늘어나고 있다. 이에 본 논문은 다양한 종류의 의견 정보를 파악하여 데이터 웨어하우스에 저장한 후, 의견 정보를 온라인에서 동적으로 분석하고 통합 처리하는 모델링 방안을 제시한다. 또한 이를 활용하여 실제 전자 상거래 사이트의 한 종류인 어플리케이션 판매 사이트의 리뷰에 대한 분석을 수행하였다.

ABSTRACT

Product reviews in online shopping sites can serve as a useful guideline to buying decisions of customers. However, due to the massive amount of such reviews, it is almost impossible for users to read all the product reviews. For this reason, e-commerce sites provide users with useful reviews or statistics of ratings on products that are manually chosen or calculated. Opinion mining or sentiment analysis is a study on automating above process that involves firstly analyzing users' reviews on a product to tell if a review contains positive or negative feedback, and secondly, providing a summarized report of users'

본 연구는 지식경제부의 산업원천기술개발사업(10038588, 앱스토어 환경을 지원하는 상황인지기반 고객경험관리 플랫폼)의 연구결과로 수행되었음.
이 논문은 한국전자거래학회 2011 춘계학술대회에서 발표된 "전자 상거래의 온라인 감성분석처리가 필요한 이유"의 제목으로 발표된 논문을 확장한 것임.

* 교신저자, 서울대학교 컴퓨터공학부 박사과정
** 삼성전자 DMC연구소 책임연구원
*** 숙명여자대학교 컴퓨터과학부 교수
**** 서울대학교 컴퓨터공학부 교수

2011년 07월 23일 접수, 2011년 08월 22일 심사완료 후 2011년 09월 29일 게재확정.

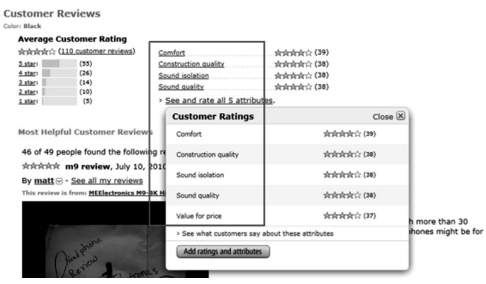
opinions. Previous researches focus on either providing polarity of a user's opinion or summarizing user's opinion on a feature of a product that result in relatively low usage of information that a user review contains. Actual user reviews contains not only mere assessment of a product, but also dissatisfaction and flaws of a product that a user experiences. There are increasing needs for effective analysis on such criteria to help users on their decision-making process. This paper proposes a model that stores various types of user reviews in a data warehouse, and analyzes integrated reviews dynamically. Also, we analyze reviews of an online application shopping site with the proposed model.

키워드 : 감성분석, 오피니언 마이닝, OLAP, 데이터 웨어하우스
Sentiment Analysis, Opinion Mining, OLAP, Data Warehouse

1. 서 론

웹을 통한 전자 상거래와 정보 공유가 활발해짐에 따라 상품에 대한 리뷰 문서가 기하급수적으로 증가하였다. 사용자들은 상품을 구매한 사이트뿐만 아니라 트위터, 블로그, 페이스북과 같은 소셜 미디어를 통해 자연스럽게 의견을 공유하고 상품 구매시 실제적인 도움을 받는다. 하지만 리뷰 문서의 방대한 양으로 인해 구매 예정자가 모든 리뷰 문서를 읽고 제품에 대한 전체적인 평가를 파악하는 것은 점점 더 어려워지고 있다.

상품에 대한 주요 의견을 효과적인 정보의 형태로 전달하는 것이 실제 구매에 영향을 주는 것이 밝혀져 있으며[19], 실제로 Amazon이나 Review Centre와 같은 사이트들은 상품 자체의 정보뿐만 아니라 상품에 대한 리뷰를 효과적으로 전달하는 것에 높은 비중을 두고 있다. 구체적으로 <그림 1> 상품 리뷰 요약 화면와 같이 상품 자체의 만족도 점수와 세부 특징별 점수를 구매자로부터 입력 받고 이들 점수를 종합하여 제공하거나, 구매 예정자들이 유용하다고 평가한 리뷰나 상품 특징을 강조하여 표시한다. 하지만 이 대부분의 과정은 사용자의 참여나 시스템 관리자에 의해 수작업으로 진행되기 때문에, 대량의 문서 리뷰를 대상으로 할 때는 효율성과 비용의 문제가 발생한다.



〈그림 1〉 상품 리뷰 요약 화면

이러한 문제를 해결하고자 최근 리뷰 문서에서 의견을 자동으로 추출하고 분석하는 오피니언 마이닝(opinion mining) 또는 감성 분석(sentiment analysis)과 관련된 연구들이 활발히 진행되어 왔다. 이들은 자연언어처리, 텍스트 마이닝, 통계 등의 분석 방법을 사용하여 사용자 들이 작성한 의견이 긍정적인지, 부정적인지 판단하고 그 정보를 요약하여 제공한다. 이러한 일련의 과정을 위해 감성 분석은 크게 의견 대상 추출, 극성 판단, 요약

및 시각화 등의 절차를 수행한다. 대표적으로 자연언어처리 기법이나 기계 학습에 기반하여 의견의 긍정, 부정을 판별하는 연구를 비롯하여[9, 13], 통계적 분석 기법에 기반하여 상품에 대한 사용자의 점수, 상품 특징 어휘 빈도수로부터 특징 단위의 평가를 도출하는 연구[16, 14, 8]가 진행되었다.

한편, 시스템에 모이는 의견 정보가 증가할수록 OLAP(On-Line Analytical Processing)처럼 의견 정보를 다양한 각도로 분석하고, 의사 결정 지원에 활용하고자 하는 요구가 증가하고 있다. 전자 상거래 사이트의 상품 후기를 작성할 수 있는 곳에는 사용자들이 상품에 대한 평가뿐만 아니라, 배송 중의 문제와 같은 고객 불만, 환불 요청과 같은 고객 지원, 제품 사용법에 대한 질문과 그 답변, 사용자들의 잡담이나 스팸 등 다양한 종류의 텍스트가 작성된다. 이러한 정보는 구매 예정자들에게 유용하기도 하지만, 만약 사이트의 관리자가 사용자들이 자주 언급하는 제품의 단점이나, 시간에 따른 사용자들의 성향 변화 등을 파악하고자 할 때 매우 유용하다고 볼 수 있다.

따라서, 본 논문에서는 의견 정보를 의사 결정 지원에 활용하기 위한 처리 기법으로 온라인 감성 분석 처리(On-Line Sentiment Analytical Processing, OLSAP) 방안을 제시하고자 한다. 기존의 감성 분석은 주로 상품 평가에 해당하는 텍스트를 특징 단위로 요약하거나 특정 키워드에 연관된 의견의 극성을 판단하는 것에 주로 초점을 맞춰 왔기 때문에 의견 데이터의 저장에 적합한 모델링, 유연한 활용을 위한 처리 방안에 대한 고려가 많은 부분 부족하였다. 특히, 의견 데이터는 자연언어에서 추출한 후 가공을 거치는 것으로, 전통적인 OLAP에 저장되는 구매이력과 같은 트랜잭션 데이터와 표현 형식이나 그 특성에 많은 차이가 있다. 따라서 OLSAP는 의견 데이터를 데이터 웨어하우스에 적합한 형태로 모델링하여 저장하는 것과 온라인에서 동적으로 의견 데이터를 통합적으로 분석하는 방안 등을 포함하게 된다.

본 논문은 다음과 같이 구성되어 있다. 제2장에서는 감성 분석과 오피니언 마이닝에 대한 이전 연구들에 대하여 설명한다. 제3장은 기본적인 의견 데이터의 추출과 모델링 방법을 제시한다. 제4장은 OLSAP에 대한 모델링 기법을 논의하며 제5장에서 실제 어플리케이션 판매 사이트의 상품평을 대상으로 분석을 수행한 결과를 설명한다. 제6장은 논의 및 향후 과제에 대해 기술한다.

2. 관련 연구

감성분석, 또는 오피니언 마이닝은 텍스트 내에서 나타나는 사용자들의 의견을 파악하는 연구이다. 크게 의견 추출, 극성 판단, 요약 및 시각화 등의 절차로 수행된다. 또한 이들을 종합한 감성분석 시스템들도 다수 개발되었다.

2.1 감성분석 프로세스

의견 추출은 문서나 문장 내에서 의견 표현과 그 표현의 대상을 찾아내는 단계이다. 이 경우 단순히 대상이 되는 어휘만을 추출하는 것이 아니라 해당 어휘와 관련된 의견

을 나타내는 어휘정보도 함께 추출한다. 자연언어처리 기반 추출 방법[14, 11, 8]은 이를 위해 문장 구조 정보를 활용한다. 전문가가 의견으로 활용될 수 있는 어휘들이 갖는 특징을 고려하여 일정 어휘 패턴을 정의 한 후, 패턴에 부합되는 어휘들이 의견 대상 어휘로 선택된다. 또한 해당 어휘를 수식하거나 서술하는 어휘가 의견을 나타내는 의견 표현 어휘로 추출된다.

사전에 정의된 문장 구조, 어휘 품사 등을 고려한 자연언어처리 기반 추출 방법 이외에도 문서 내 어휘 빈도수, TF-IDF 등의 수치에 기반하거나, 통계적 기법에 기반한 어휘 추출 방법도 시도되었다[19, 12].

극성 판단 단계는 추출된 의견 어휘가 문서나 문장 내에서 가지는 의미를 긍정적/부정적/중립적 등으로 판단하는 단계이다. 자연언어처리 기반의 방법은 워드넷(WordNet)[21] 또는 센티워드넷(SentiWordNet)[7] 등의 사전이 포함하는 의미 정보 등을 활용한다. 예를 들면, 형용사 및 동의어 집합을 활용하여 특정 어휘의 의미 극성 성향을 판단하거나, 어휘 사전을 생성, 확장하는데 워드넷을 활용하는 방법 등이 시도되었다[14, 11]. 기계학습에 기반하여 설계된 분류기(classifier)를 활용한 연구도 이루어졌다[9]. 이들은 의견이 표현된 문서와 그와 연관된 극성정보로 분류기를 학습시킨 후 새로운 문서에 대한 의미 극성을 판별한다.

이외에도, *PMI*(pointwise mutual information)로 어휘간의 관련 정도를 정의하여 의미 극성을 판별하는 방법도 시도되었다[21, 20]. 일반적으로 *PMI*는 다음과 같이 정의되며, 이는 두 어휘가 동시에 출현할 확률 값을 의미한다.

$$PMI(term_1,\ term_2) = log_2 \frac{p(term_1,\ term_2)}{p(term_1)p(term_2)}$$

이때 대상 어휘의 극성은 극성 정보가 모호하지 않은 긍정적 단어와 부정적 단어 사이의 관련도의 차이에 따라 결정할 수 있다.

$$SO(term) = PMI(term, \text{“}excellent\text{”}) - PMI(term, \text{“}poor\text{”})$$

위 식의 경우, *“excellent”*와 *“poor”*는 각각 긍정의 의미와 부정의 의미를 확실하게 지니고 있다. 따라서 어떤 어휘가 문장 또는 문서 내에서 *“excellent”*와 함께 많이 출현하였다면 *SO* 값은 양의 값을 가지게 되어 긍정적인 단어일 확률이 높은 것이며, 반대로 *“poor”*와 동시에 자주 나타났다면 *SO*는 음의 값이 되어 부정적인 단어로 판단될 확률이 높은 것이다. 만약 어떤 어휘가 *“excellent”*와 *“poor”*와 비슷한 출현 빈도를 지녔다면 *SO*값은 0에 가까워져 의견을 표현하지 않은 어휘가 되는 것이다.

감성 분석의 마지막 단계인 요약 및 시각화는 문서 내에서 추출된 의견 정보를 종합하여 긍정적, 부정적 표현들의 분포를 시각적으로 표현하여 사용자들에게 제공한다. 이들은 주로 요약된 정보와 구현된 시스템에 따라 다른 형태를 보인다.

2.2 감성분석 시스템

Opinion Observer[14]는 주로 상품을 다루

고 있는 리뷰 문서의 의견 정보를 분석하여 요약된 결과를 제공한다. 자연언어처리 기반의 방법에 기반을 두고 있으며, 사전에 정의된 패턴으로 의견 정보를 추출한 후, 워드넷을 활용하여 극성을 판단한다.

OPINE[18] 시스템은 사전에 정의된 문장 구조 패턴을 사용하여 의견 대상 어휘를 추출한다. 이때 문장 구조와 어휘 사이의 *PMI* 값을 계산하여 의견 대상 어휘를 선택하며, 구문 분석으로 이와 연관된 의견 어휘를 결정한다.

Red Opal[19] 시스템은 별점으로 표현된 상품의 점수를 활용하여 의견의 극성 및 강도를 계산한다. 대용량의 리뷰문서를 요약하기 위해 자연언어기법보다는 어휘의 빈도수와 상품에 대한 상품점수를 이용한 통계적 분석을 활용한다. 이를 통해 리뷰 요약 결과를 의견 표현 대상에 대한 점수의 형태로 제공한다.

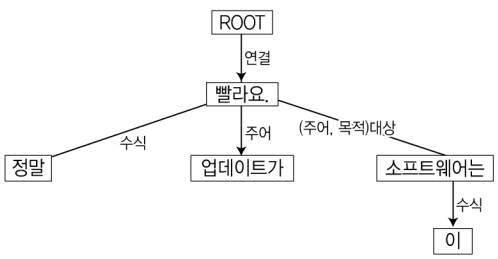
하지만 이들 시스템들은 감성 분석의 대상이 상품 리뷰 등의 문서로 제한되어 있으며, 그 결과도 사전에 정의한 형태로 고정되어 있는 등 매우 시스템 구현에 의존적이다. 만약 상품 리뷰가 아닌 다른 형태의 의견 텍스트를 분석하고자 할 때나, 분석 결과를 다양하게 활용하고자 할 경우에는 이러한 것이 제한이 된다는 단점이 있다.

3. 의견 데이터 모델링

상품 리뷰는 비구조적 텍스트 데이터이기 때문에 의견 문서나 문장이 포함하고 있는 의견 정보를 활용하기 위해서는 의견 표현의 추출 및 극성 판단의 과정이 반드시 선행되어야 한다. 이 장에서는 이 과정에서 활용할 수 있는 비교적 단순하고 명확한 방법을 기술하며, 이렇게 추출된 의견 데이터를 OLSAP에 적합한 형태로 모델링 하는 방안을 제시한다. 본 논문은 추출된 의견 데이터의 활용에 초점을 맞추고 있으므로, 의견 추출 및 극성 판단 과정에 활용할 수 있는 다양한 방법은 앞서 설명한 관련연구로 대신한다.

3.1 의견 데이터 추출 및 극성 판단

의견 데이터는 *"이 소프트웨어는 업데이트가 정말 빨라요."* 처럼 대상 그 자체나 대상의 일부 특징에 대해 사용자의 의견을 언급하는 텍스트 데이터로부터 추출한다. 이때 한글의 경우 사전 작업으로 형태소 분석과 구분분석을 거치게 된다. 형태소 분석은 텍스트로부터 각각의 형태소를 분리하여 원형을 복원하고 그 품사를 추가하는 과정이다. 앞의 예문은 형태소 분석으로 *"이/MAG 소프트웨어/NNG+는/JX 업데이트/NNG+가/JKS 정말/MAG 빠/VV+르라요/EFN+./SF"*와 같은 형태로 변환된다. 구문 분석의 경우에는 각각의 어절이 문장 내에서 가지는 문법적인 역할을 분석하는 과정이다. <그림 2> 구문 분석 결과는 앞의 예문으로 구문 분석을 수행한 결과이다.



〈그림 2〉 구문 분석 결과

여러 형태소 분석기와 구분 분석기가 개발되어 있으나, 본 논문에서는 오픈소스 형태로 공개된 형태소 분석기를 사용하였다[5, 2].

의견 텍스트는 형태소 분석 및 구문분석 등의 적절한 자연언어 처리를 거친 후에, 의견 표현과 그 대상, 수식어 등의 정보를 추출한다. 앞의 예제의 경우 <*"소프트웨어", "업데이트", "빠르다", "정말"*>와 같은 정보를 추출 할 수 있다. 문장 내에서 동사에 해당하는 형태소인 "빠르다"를 의견 표현으로 선택한 후, 구분 분석 트리에서 이 노드의 하위 노드들을 탐색하여 의견 표현의 대상으로 *"소프트웨어"*, 그 일부 특징으로 *"업데이트"*, 수식어로 *"정말"*을 추출하게 된다.

이후에 이렇게 추출된 정보에서 *"빠르다"*가 긍정/부정인지, *"정말"*의 경우 의미를 강화시키는지, 약화시키는지 등을 판단하여 최종적인 극성값을 구하게 된다. 이전에 살펴본 극성을 판단하는 다양한 방법들은 대부분 영어를 대상으로 하고 있으며, 한국어의 경우 형태소 분석, 구문분석을 거친 후 전문가에 의해 반자동으로 사전에 구축된 어휘 사전과 극성정보를 활용하여 의견 정보를 추출하는 방법이 연구되어 있다[1].

의견을 표현하는 문장은 앞의 예와 같이 *"소프트웨어"*, *"업데이트"*와 같이 그 대상 및 대상의 특징을 표현 내에서 직접적으로 언급하기도 하지만, *"이 게임은 재미없어요"*와 같이 대상 자체만 나와 있는 경우나 *"별로네요"*와 같이 대상과 그 특징이 모두 생략된 경우가 있을 수 있다. 의견을 대상의 특징별로 점수화하여 요약하는 시스템들의 경우 이 문제를 중요한 문제로 다루며 분석 대상을 한정하는 등의 방법으로 해결하기도 한다. 하지만 추출되는 정보들은 결과적으로 *"빠르다", "재미없다", "별로다"*와 같이 의견 표현이 반드시 포함된다고 볼 수 있다. 그러므로 이를 기준으로 한 표준화된 형태의 의견 데이터의 모델링 방법을 고려해보고자 한다.

3.2 의견 데이터 모델링

의견 데이터 모델링은 의견 정보의 추출 대상과 활용하는 응용에 따라서 상이할 수 있다. OLSAP는 가장 일반화된 형태와 추후 분석에 활용할 정보를 최대한 포함하는 것을 목적으로 의견 정보를 모델링하여 사용한다. 의견 정보의 요소는 앞서 살펴본 의견 표현과 의견 표현의 대상, 대상의 일부 특징, 극성 정보 외에도 다양한 정보가 존재한다. 우선적으로, 의견을 표현하는 주체로 사람이나 단체가 있다. 대부분의 경우 블로그나 상품 리뷰처럼 그 글을 작성한 사용자가 그 주체가 된다. 그러나, 신문의 사설 같은 경우에는 그 글을 작성한 기자가 아닌 신문사의 의견을 표현하고 있는 것이며, 보도 자료에서의 의견 표현은 기사를 작성한 기자가 아닌 취재를 당한 대상이라고 할 수 있다.

이러한 정보들을 종합하여 의견 데이터를 표현하면 다음과 같다.

$$< o_i, f_j, e_k, m_l, ve_{jk}, vm_{kl}, u_m, t_n, p_o >$$

o_i는 *"소프트웨어와"*같이 의견 표현의 대상이며, f_j는 *"업데이트"*와 같은 o_i의 세부 특징이다. 이들은 앞서 살펴본 바와 같이 그 값이 없을 수도 있는데, f_j 값이 없을 경우에는 o_i 자체를 언급하고 있는 것이라고 볼 수 있

고, o_i의 경우에도 문장 내에서 대상이 표현되어 있지 않다 하더라도 외부 정보를 사용하여 그 대상을 추측할 수 있다. 예를 들면, 상품 정보 페이지의 사용 후기가 분석의 대상이라면, o_i가 그 상품 자체가 되어 상품명이나 상품 ID 등으로 채워질 수 있을 것이다.

e_k는 "*좋다*", "*나쁘다*"와 같이 의견을 표현하고 있는 어휘가 되며, 일반적으로 형태소 분석된 어휘의 원형으로 표준화하여 사용한다. 예를 들면, "*좋네요*", "*좋았어요*" 와 같은 어휘들은 "*좋다*"로 변환한다. m_l는 "*꽤*", "*별로*" 의견의 강도에 대한 어휘로 와 같이 의견 극성에 영향을 준다. "*꽤*"와 같이 의견을 강조하는 경우도 있지만, 경우에 따라 "*안 좋네요*"의 "*안*"과 같이 의견 극성을 반대로 바뀌는 경우가 있어 이들에 대한 처리가 필요하다.

ve_{jk}와 vm_{kl}는 각각 의견과 표현과 의견강도에 대한 어휘로 실수 값을 가진다. 일반적으로 두 값을 곱하는 것으로 최종적인 의견 극성을 결정하게 되는데, 의견 표현의 극성값은 부정일 경우 음수, 긍정일 경우 양수가 된다. 의견 강도에 대한 값은 의견 극성 값을 강조하는 경우에는 1보다 큰 수를, 의견 성향을 반대로 바꾸는 경우에는 -1 이하의 수를 가지게 된다. 이들 수치값의 결정은 전문가에 의해 정제된 사전을 구축하여 하용하거나 앞서 관련연구에서 살펴본 다양한 방법들로 결정하게 된다.

마지막으로, u_m는 의견을 제시한 사용자, t_n는 의견이 작성된 시각, p_o는 의견이 작성된 위치를 나타낸다. 일반적인 구매 후기와 같은 경우에는 위치 정보가 없거나 IP 정보로 위치 정보를 추정하게 되지만, 트위터와 같은 소셜 미디어 들은 위치 정보를 중요한 요소로 GPS 정보를 직접 포함하기도 한다. 시각, 사용자, 위치 정보와 같은 요소들은 분석 대상에 따라서 그 값을 파악하기 어려울 수 있으나, 다양한 차원에서 의견 정보를 분석할 수 있게 하는 측면에서 중요한 역할을 한다.

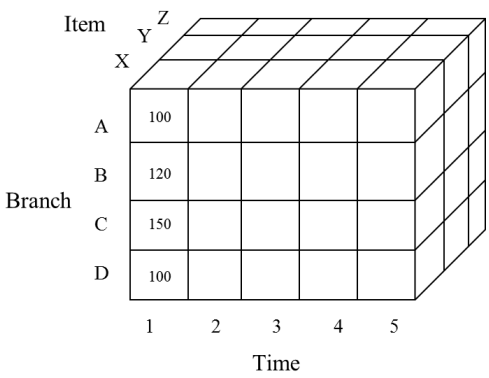
4. 온라인 감성 분석 처리

OLAP는 의사 결정 지원을 위해 데이터 웨어하우스에 기반하여 대규모의 다차원 데이터를 통합하고, 온라인에서 동적으로 분석하는 일체의 처리를 말한다[6, 10, 4]. 데이터 웨어하우스는 구매 이력과 같이 다수의 트랜잭션 데이터베이스에서 발생한 데이터를 정보 분석에 사용하고자 종합하여 저장 및 관리를 하는 데이터베이스이다. 일반적으로 변경과 삭제보다는 읽기 작업에 초점이 맞춰져 설계된다. OLAP는 다차원 모델을 물리적으로 저장하는 방식에 따라 다차원 OLAP (multi-dimensional OLAP), 관계형 OLAP(relational OLAP) 등으로 구분된다.

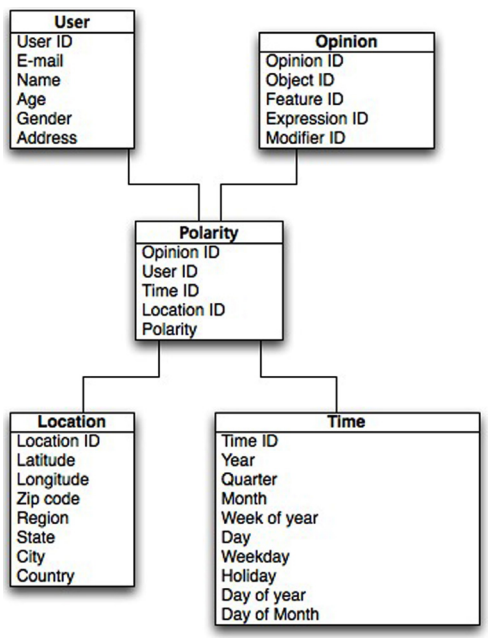
OLAP는 일반적으로 다차원으로 구성된 데이터를 다루며, 이를 위해 데이터 큐브(cube)라는 모델을 사용한다. 예를 들어, 매장의 상품 판매 이력 데이터는 특정 상품이 팔릴 때 그와 관련된 연관 데이터들이 발생한다. 연관 데이터는 판매 지점, 판매 기간, 판매량 등의 데이터 등이 있다. 데이터 큐브는 분석의 기준이 되는 차원으로 상품, 판매 지점, 판매 기간 등을 구성하게 되고 각각에 대한 측정치로 판매량 정보를 담게 된다. <그림 3> 판매량 데이터 큐브은 이러한 데

이터 큐브의 예시를 보여준다. 각각의 차원은 일반적으로 계층 구조를 이루게 되는데, 판매 기간(Time)의 경우 연도 - 분기 - 월 - 일 - 시와 같은 상세 계층으로 표현할 수 있다.

OLAP는 이렇게 각각의 차원과 그 계층구조에 기반하여 group by 연산이나 평균 또는



〈그림 3〉 판매량 데이터 큐브



〈그림 4〉 OLSAP의 스키마

합 등의 집계 연산을 통해 데이터를 분석하게 된다. 또한 각각의 차원에 대한 롤업(rollup), 드릴다운(drill-down) 등의 연산을 지원하여 요약된 데이터 수준에서 보다 구체적인 내용의 상세 데이터로, 또는 그 반대로 접근하면서 다양한 분석을 수행하게 된다.

OLSAP는 이러한 OLAP에 기반하여 다양한 의견 정보를 종합하여 다차원 데이터를 구성하게 되며, 이와 관련된 각종 질의 및 응용을 처리한다. 이를 위해서는 우선적으로 OLSAP에서는 어떤 방식으로 의견 데이터로 데이터 큐브를 구성할 수 있는지 고려하는 것이 필요하다.

의견 데이터는 앞서 살펴본 것처럼 $< o_i, f_j, e_k, m_i, ve_{jk}, vm_{ki}, u_m, t_n, p_o >$와 같이 다양한 요소들을 포함하여 모델링 된다. 의견 데이터는 상품 판매 이력 등과 같이 전통적으로 OLAP가 분석의 대상으로 삼아왔던 데이터와는 그 형태와 특성에 차이가 있다. 첫째로, 판매 이력은 판매량과 같이 차원 요소로 측정되는 측정치가 명확하게 나타나지만, 의견 데이터는 이와 같은 측정치가 불분명하다. 둘째로, 판매 이력은 그 차원 요소의 값이 없는 경우는 거의 발생하지 않지만[13], 의견 데이터는 앞서 설명한 것처럼 문장 내에 의견 표현의 대상이나 그 세부 특징의 값이 없는 경우가 빈번히 발생할 수 있다. 이와 연관하여, 판매 이력 같은 경우는 분석을 위한 차원 요소의 계층 구조를 사전에 명확히 정의할 수 있지만, 의견 데이터는 자연어를 포함하고 있어 이들 사이의 상관 관계를 명확하게 정의하는 것이 어렵다. 상품평 같은 경우에는 상품에 대한 전자 카탈로그 정보를 활용하여 의견 표현 대상에 대한 계층 구조를

구성할 수도 있으나, e_k와 같은 의견 표현 자체는 다양한 종류의 어휘로 인해 계층 구조를 만드는 것에 많은 노력이 들어간다. 이러한 두 번째 문제점을 해결하고자 이전 연구에서는 전문가가 어휘 추출 및 사전 구성을 효율적으로 할 수 있는 인터페이스를 구현하여 제공하였다[17, 1].

따라서 OLSAP는 이러한 사항들을 고려하여 구성해야 한다. 우선 감성 분석의 요약 과정은 각각의 의견 데이터의 극성 정보에 대해 집계 연산을 수행하는 것으로 볼 수 있으므로, 의견 데이터의 극성 정보를 판매량과 같은 측정치에 해당하는 요소로 볼 수 있다. 특히 극성 정보는 의견 데이터에서 ve_{jk}와 vm_{kl}와 같이 실수값으로 수치화 하여 표현할 수 있기 때문에 측정치로 사용하기에 적합하다. 측정치로 극성 정보를 사용하면 차원 요소들은 이외의 항목들로 결정되며, 이에 대한 스키마는 <그림 4> OLSAP의 스키마와 같다. 중심이 되는 테이블에는 측정치인 의견 극성 값과 이와 연관 되어 있는 차원 요소들이 포함되어 있다.

이렇게 구성된 데이터베이스를 이용하여 제품에 대한 사용자들의 전체적인 평가뿐만 아니라, 지역별 평가, 시간에 따른 의견의 변화 추세 등 복합적인 정보를 얻을 수 있게 된다.

5. 실험 및 분석

본 장에서는 스마트폰용 어플리케이션 판매 사이트의 평가 글을 대상으로 이들의 특성을 밝힌 후, 실제 의견 데이터의 추출을 수행하여 OLSAP에서의 활용 가능성에 대하여 논의한다.

데이터는 LG U+의 OZ 스토어에서 크롤하여 수집하였다. 데이터는 총 1,404개의 어플리케이션에 대해서 2010년 7월 29일부터 2011년 4월 26일까지 작성된 46,166개 평가글을 포함한다. 평가글을 작성한 총 사용자는 7306명이며, 이중 가장 많은 글을 작성한 사람은 480개의 상품평을 작성하였고, 1개의 상품평만을 작성한 사람은 657명이었다.

<표 1> 상품평 예시는 실제 데이터의 일부를 나타낸다. 이 데이터에서 나타나는 것처럼 실제로 후기를 적는 곳에는 어플리케이션의 평가뿐만 아니라 제작사 공지, 질문 및 답변, 환불 등의 요청, 잡담, 스팸과 같은 다양한 종류의 글이 나타난다. OLSAP에서 의견 데이터를 추출하고자 할 때는 의견에 해당하는 데이터만을 분류하는 사전 작업이 필요하다. 왜냐하면 이러한 데이터를 포함하여 분석 작업을 수행할 경우 전체적인 의견 성향에 오류를 가져올 수 있기 때문이다. 예를 들어 <표 1> 상품평 예시의 *"[컴투스] 010********님, 안녕하세요. 이곳 게시판을 통해서는 안내가 어렵습니다"*와 같은 텍스트는 이전 연구[1]의 방법으로 의견 데이터 추출을 수행하였을 때 *<안내, 어렵다>*와 같은 부정적인 표현을 추출할 수 있다. 하지만 텍스트 자체는 제작사 공지로서 사용자가 어플리케이션을 실제로 평가한 내용이라고 보기 어렵다. 그렇기 때문에 이러한 텍스트는 분석 대상으로 적절하지 않으며, 별도의 처리로 제거되는 편이 좋다고 볼 수 있다. 본 논문은 의견 텍스트와 그 활용에 초점을 맞추고 있으므로, 그 방법 등에 관한 논의는 생략한다.

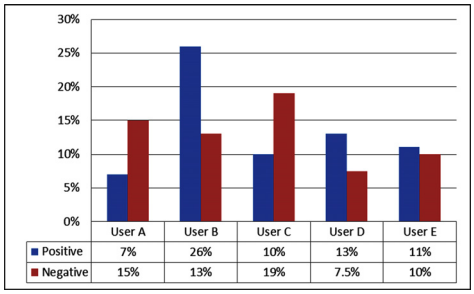
〈표 1〉 상품평 예시

No.	Text
1	*재미있어용 받을만함*
2	*[컴투스] 010********님, 안녕하세요. 이곳 게시판을 통해서는 안내가 어렵습니다. …*
3	*[마나스톤] 엔딩까지 오픈맵 버전 검증 중입니다. …*
4	*지우고시픈데어케해야하죠?*
5	*인터넷알바로돈버세요 만20이상 …*
6	*구문분석이 안된다네요 … 환불해주세요*

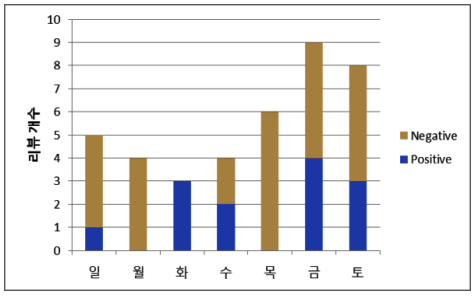
<그림 5>와 <그림 6>은 앞서 언급한 데이터베이스 스키마를 구축한 후 분석을 수행한 결과들의 예시이다.

<그림 5> 사용자별 평가 극성 비율은 사용자가 작성한 리뷰 글들 중 의견이 포함된 글들을 대상으로 각각 긍정, 부정을 판단 한 후, 이를 사용자별로 집계한 것이다. User A의 경우를 보면, User A가 작성한 전체 글들 중 의견이 포함된 글은 25%이며, 긍정 표현은 전체의 7%, 부정 표현은 전체의 15% 임을 나타낸다. 이를 통해, 사용자들의 기본적인 의견 평가 성향을 알 수 있는데, User B와 User C를 비교하면 User B의 경우 부정적인 표현보다 긍정적인 표현의 수가 2배이기 때문에 User B를 평가에 후한 사용자라 볼 수 있으며, 반대로 User C는 긍정적인 의견을 좀처럼 표현하지 않는 사용자라 판단할 수 있다. 이러한 평가 결과는 다양한 방법으로 활용될 수 있는데, 만약 User C와 User B가 같은 대상에 대해 각각 긍정적인 의견을 작성하였다고 하면, User B의 의견보다 User C의 의견이 좀 더 비중이 있다고 볼 수 있다.

<그림 6> User D의 요일별 극성 변화는



〈그림 5〉 사용자별 평가 극성 비율



〈그림 6〉 User D의 요일별 극성 변화

한 사용자가 작성한 긍정, 부정 글의 개수를 요일별로 집계한 결과이다. 이와 같은 시간 단위에 따른 의견 추이의 변화 또한 다양하게 활용될 수 있다. 예를 들면, 특정 어플리케이션의 업데이트에 따른 사용자들의 반응을 알아보고자 한다면, 업데이트 시점에 해당하는 특정 시점의 이전과 이후의 긍정, 부정 의견 개수나 비율의 추이를 살펴봄으로써 사용자들의 반응을 정량화된 값으로 얻는 등의 활용이 가능하다.

6. 결 론

구매 예정자나 관리자에게 의견 데이터는 매우 유용한 정보이지만, 상대적으로 이를 효

과적으로 활용하고자 하는 시도는 부족하였다. 이에 본 논문에서는 비구조적인 텍스트에서 추출한 의견 데이터를 다루기 위한 데이터 모델과, 축적된 의견 정보를 다각도로 분석하기 위한 처리기법인 OLSAP를 제안하였다. 이러한 방법으로 분석된 의견 정보는 제품의 평가뿐만 아니라, 향후 전자 상거래 시스템과 결합하여 마케팅 등 다양한 목적으로 활용 할 수 있을 것으로 기대된다.

참 고 문 헌

[1] 명재석, 이동주, 이상구, “반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템”, 정보과학회논문지 : 소프트웨어 및 응용, 제35권, 제6호, pp. 392-403, 2008.

[2] 이동주, 연종흠, 이상구, “한국어 문장의 띄어 쓰기 오류 교정과 최적 형태소 분석을 위한 통합 확률 모델”, 한국컴퓨터종합학술대회논문집, 제38권, 제1A호, pp. 237-240, 2011.

[3] 이현자, 심준호, “관계형 데이터베이스 상품 정보 질의 처리를 위한 인덱싱”, 한국전자거래학회지, 제13권, 제4호, pp. 209-222, 2008.

[4] 장재영, “OLAP 환경에서 다중 존 디스크를 활용한 실체뷰의 효율적 저장 기법”, 한국전자거래학회지, 제14권, 제1호, pp. 143-160, 2009.

[5] 꼬꼬마 한글 형태소 분석기, http://kkma.snu.ac.kr.

[6] Chaudhuri, S. and Dayal, U., “An overview of data warehousing and OLAP technology,” SIGMOD Record, Vol. 26, , No. 1, pp. 65-75. 1997.

[7] Denecke, K., “Using SentiWordNet for Multilingual Sentiment Analysis,” In Proceedings of the International Conference on Data Engineering : ICDE, Workshop on Data Engineering for Blogs, Social Media, and Web 2.0, 2008.

[8] Ding, X., Liu, B., and Yu, P. S., “A holistic lexicon-based approach to opinion mining,” In Proceedings of the international conference on Web search and web data mining, pp. 231-240, 2008.

[9] Esuli, A. and Sebastiani, F., “Determining Term Subjectivity and Term Orientation for Opinion Mining,” In Proceedings of 11th conference of the European chapter of the Association for Computational Linguistics : EACL, pp. 193-200, 2006.

[10] Gray, J., Bosworth, A., Layman, A., Reichart D., and Hamid Pirahesh, “Data cube : A relational aggregation operator generalizing group-by, cross-tab, and sub-totals,” Data Mining and Knowledge Discovery, Vol. 1, , No. 1, pp. 29-53, 1997.

[11] Hu, M. and Liu, B., “Mining and summarizing customer reviews,” In Proceedings of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data mining, pp. 168-177, 2004.

[12] Jin, W., Ho, H., and Srihari, R., “OpinionMiner : a novel machine learning system

for web opinion mining and extraction," In Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining, pp. 1195-1204, 2009.

[13] Jindal, N. and Liu, B., "Mining Comparative Sentences and Relations," In Proceedings of the 21st national conference on Artificial intelligence, pp. 1331-1336, 2006.

[14] Liu, B., Hu, M., and Cheng, J., "Opinion observer : analyzing and comparing opinions on the Web," In Proceedings of the 14th international conference on World Wide Web, pp. 342-351, 2005.

[15] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miler, K., "Introduction to WordNet : An on-line lexical database," International Journal of Lexicography, pp. 235-244, 1990.

[16] Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T., "Mining Product Reputations on the Web," In Proceedings of the 8th SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 341-349, 2002.

[17] Myung, J., Yang, J., and Lee, S., "PicAChoo : A Tool for Customizable Feature Extraction Utilizing Characteristics of Textual Data," In Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication : ICUIMC, pp. 650-655, 2009.

[18] Popescu, A. and Etzioni, O., "OPINE : Extracting product features and opinions from reviews," In Proceedings of the conference on Human Language Technology/Empirical Methods in Natural Language Processing : HLT/EMNLP, pp. 339-346, 2005.

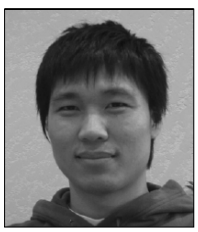
[19] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., and Jin, C., "Red Opal : Product-Feature Scoring from Reviews," In Proceedings of the 8th ACM conference on Electronic Commerce, pp. 182-191, 2007.

[20] Turney, P. and Littman, M., "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," Technical Report ERC-1094 (NRC 44929), National Research Council of Canada, 2002.

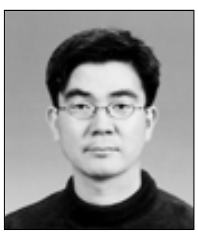
[21] Turney, P. and Littman, M., "Measuring praise and criticism : Inference of semantic orientation from association," ACM Transactions on Information Systems, Vol. 21, pp. 315-346, 2003.

저 자 소 개

연종흠 (E-mail : jonghm@europa.snu.ac.kr)
2008년 서울대학교 컴퓨터공학부 (학사)
2008년～현재 서울대학교 컴퓨터공학부 (박사과정)
관심분야 데이터베이스, 자연언어 처리, 데이터 마이닝



이동주 (E-mail : therocks@europa.snu.ac.kr)
2003년 서울대학교 응용생물화학부 (학사)
2011년 서울대학교 컴퓨터공학부 (박사)
2011년～현재 삼성전자 DMC연구소 책임연구원
관심분야 데이터베이스, 자연언어 처리, 상황인지 개인화



심준호 (E-mail : jshim@sookmyung.ac.kr)
1990년 서울대학교 계산통계학과 (학사)
1994년 서울대학교 계산통계학과 전산과학전공 (석사)
1998년 Northwestern University, Electrical and Computer Engineering (박사)
2001년～현재 숙명여자대학교 컴퓨터과학부 교수
관심분야 데이터베이스, 전자상거래, 상품정보, 온톨로지

이상구 (E-mail : sglee@europa.snu.ac.kr)
1985년 서울대학교 계산통계학과 (학사)
1987년 Northwestern University, Computer Science (석사)
1990년 Northwestern University, Computer Science (박사)
1992년～현재 서울대학교 컴퓨터공학부 교수
관심분야 e-비즈니스 기술, 시맨틱 웹, 온톨로지, 데이터베이스