# 온라인 쇼핑몰의 상품평 자동분류를 위한 감성분석 알고리즘

## A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in On-Line Shopping Mall

장재영(Jae-Young Chang)*

**초 록**

급속한 전자상거래의 발전으로 인하여 온라인상으로 상품을 구매하고 그에 대한 평가를 작성하는 것이 일반적인 구매 패턴이 되었다. 기존 구매자들의 상품평들은 다른 잠재적인 소비자들의 상품 구입을 이끌어내는데 큰 동기가 된다. 사용자가 작성한 상품평은 하나의 상품에 대해 실제 사용자의 좋고 나쁨에 대한 감정을 표현한 결과로, 개개인에 따라 긍정 또는 부정적인 의견으로 나눠진다. 상품평 중에서 소비자가 원하는 정보를 얻기 위해서는 이들을 일일이 수작업으로 확인해야하지만, 온라인 쇼핑몰에 상품평이 대용량으로 축적된 환경에서 이러한 작업은 비효율적일 수밖에 없다. 본 논문에서는 오피니언 마이닝 기술을 이용하여 제품 사용자의 주관적 의견을 자동으로 분류할 수 있는 감성분석 알고리즘을 제시한다. 본 논문에서 제시하는 알고리즘은 온라인 쇼핑몰에 등록된 개별 상품평을 대상으로 긍정 및 부정 의견으로 판단하여 요약된 결과를 제공하는 기능을 한다. 본 논문에서는 또한 제안된 알고리즘을 바탕으로 개발된 상품평 자동분석 시스템을 소개하고, 알고리즘의 효율성을 검증하기 위한 실험결과도 제시한다.

**ABSTRACT**

With the continuously increasing volume of e-commerce transactions, it is now popular to buy some products and to evaluate them on the World Wide Web. The product reviews are very useful to customers because they can make better decisions based on the indirect experiences obtainable through the reviews. Product Reviews are results expressing customer's sentiments and thus are divided into positive reviews and negative ones. However, as the number of reviews in on-line shopping increases, it is inefficient or sometimes impossible for users to read all the relevant review documents. In this paper, we present a sentiment analysis algorithm for automatically classifying subjective opinions of customer's reviews using opinion mining technology. The proposed algorithm is to focus on product reviews of on-line shopping, and provides summarized results from large product review data by determining whether they are positive or negative.

본 연구는 2009년도 한성대학교 교내연구비 지원과제임.

* 한성대학교 컴퓨터공학과 부교수

2009년 09월 23일 접수, 2009년 10월 06일 심사완료 후 2009년 11월 06일 게재확정.

Additionally, this paper introduces an automatic review analysis system implemented based on the proposed algorithm, and also present the experiment results for verifying the efficiency of the algorithm.

**키워드** : 전자상거래, 온라인 쇼핑몰, 상품평, 감성분석, 자동분류
E-commerce, On-line Shopping, Product Reviews, Sentiment Analysis, Automatic Classification

## 1. 서 론

인터넷에 산재한 정보는 두 가지 형태로 나눌 수 있는데 첫째는 “사실(fact)”정보이고 둘째는 “의견(opinion)”이다. 사실은 보편적인 현상에 대해서 서술한 객관적인 정보를 의미하며 의견이란 하나의 현상에 대해서 개개인의 주관적인 정보를 의미한다. 현재 인터넷 상의 대부분의 검색 엔진은 사실에 대한 탐색에 주안점이 맞춰져 있으며 의견에 대한 검색에는 상당히 미흡한 상태이다. 그러나 인터넷을 검색하다보면 사실 뿐만 아니라 개개인의 의견 정보가 필요한 경우도 많이 발생한다. 예를 들어, 많은 사람들은 어떠한 제품에 대한 주관적 의견(subjective opinions)을 인터넷에 올림으로써 그 제품을 구입하기 원하는 사람들의 선택에 막대한 영향을 주고 있다. 그 의견들에는 제품을 만드는 회사에서도 알지 못한 장점과 단점이 모두 포함되어있기 때문에 회사 입장에서는 그 의견들을 종합하여 활용하여 제품연구와 홍보 전략에 이용할 수 있다. 하지만 인터넷에 모여 있는 수많은 정보들을 모두 수작업으로 분석하려면 막대한 시간과 비용이 소요되기에 그 정보들을 자동으로 분석하고 계층화하기 위한 기법이 필요하게 되었다. 그리고 그런 목적들이 오피니언 마이닝(opinion mining)의 발전을 가져다주고 있다.

오피니언 마이닝은 일정한 틀이 정해져 있지 않은 문서에서 그 문서의 주제를 찾아내는 방법인 텍스트 마이닝(text mining)에서 발전되었다[1~8]. 텍스트 마이닝이 어떠한 문서의 주제를 찾아내고 계층화 한다면, 오피니언 마이닝은 그 문서를 작성한 사람의 감정(sentiment)을 추출해 내는 기술이다. 어떠한 문서의 주제가 무엇인지 보다는, 그 문서를 작성한 사람이 주제에 대하여 어떠한 감정을 가지고 있는가, 예를 들어 그 주제에 대하여 ‘좋은 감정을 가지고 있는가?’ 아니면 ‘안 좋은 감정을 가지고 있는가?’를 판단하여 분석한다.

현재까지 오피니언 마이닝이 가장 성공적으로 적용되는 분야는 온라인 쇼핑몰에서 사용자의 상품평(product reviews)에 대한 분석이다[4, 5]. 실제 사용자가 작성한 상품평은 하나의 상품에 대해 사용자의 좋고 나쁨에 대한 감정을 표현한 결과이다. 따라서 개개인에 따라 긍정(positive) 또는 부정적인(negative) 의견으로 나눠지며 오피니언 마이닝에서는 이러한 감성분석(sentiment analysis)[2, 3]을 통해 상품평을 자동적으로 분류 요약하여 유용한 상품 정보로 활용될 수 있다. 오피니언 마이닝 기술은 사용자들의 상품평을 효과적으로 추출하여 요약함으로써 잠재적 소비

자들로 하여금 모든 상품평을 살펴보지 않더라도 상품에 대한 다양한 의견들을 쉽게 열람 및 검색할 수 있는 기반 기술을 제공한다.

오피니언 마이닝 연구의 핵심은 주관적 상품평에 대해서 긍정 혹은 부정적 의견을 자동으로 판단하는 것이다. 상품평은 일반적인 문서보다 길이가 짧고 상품의 속성-혹은 특성-(feature)과 의견이 많이 포함되어 있다. 예를 들어 디지털 카메라는 크기, 가격, 무게, 디자인, 배터리, 렌즈, 액정, 셔터스피드, 플래시 등의 특성을 가지고 있다. 사용자는 이러한 상품의 특성에 대한 감정적인 의견을 가질 수 있다. 카메라의 디자인에 대해서 <좋다/나쁘다>라는 기본적인 의견부터 액정이 <밝다/어둡다/선명하다/흐리다/크다/작다> 등의 다양한 의견까지 표현할 수 있다. 따라서 소비자 개인이 작성한 전체 상품평에 대한 감정의 전반적 극성(polarity)도 중요하지만 각 속성에 대한 극성을 판단하는 것도 중요한 과제가 된다.

현재 국제적으로 오피니언 마이닝과 관련한 활발한 연구가 진행되고 있다[1~11]. 또한 이러한 연구결과를 바탕으로 [12, 13] 등에서 상품평에 대한 분석 결과를 사용자에게 제공하는 시스템을 실제 운용하고 있다. 하지만 국내에서는 최근 들어 오피니언 마이닝의 개념이 소개되었고 현재는 부분적인 기초연구가 이루어지고 있는 단계에 있다[14, 15].

이러한 배경을 바탕으로 본 논문에서는 오피니언 마이닝 기술을 이용하여 제품 사용자의 주관적 의견을 자동으로 분류할 수 있는 감성분석 알고리즘을 제시한다. 본 논문에서 제시하는 알고리즘은 온라인 쇼핑몰에 등록된 한글 상품평에 대해서 전체 혹은 각 속성별로 긍정 또는 부정 의견인지 판단하고 그 정도(strength)까지 계산하는 기능을 한다. 이를 위해 기본적으로 상품평의 각 단어들이 품사별로 분류되며, 상품에 대한 속성과 감성단어들에 대한 데이타베이스가 이미 구축된 환경을 가정하였다.

본 논문에서는 또한 제안된 알고리즘을 기반으로 개발된 상품평 자동분석 시스템을 소개한다. 이 시스템은 온라인 쇼핑몰에 등록된 상품평을 대상으로 제품 사용자의 의견을 자동을 분석 요약하여 그 결과를 제공한다. 개발된 시스템에서는 상품평의 단어들을 추출하고 품사별로 분류하기 위해서 한글 형태소 분석기[16]를 사용하였고, 개별 속성별 감성단어들의 극성을 판별하기 위해 감성사전을 구축하였다. 감성사전은 감성 단어들에 대한 극성과 그 정도에 대한 정보를 담고 있다. 마지막으로 제안된 알고리즘의 효율성을 검증하기위한 실험을 실시하였다. 실험은 네이트(Nate) 쇼핑몰을 대상으로 실시하였는데 그 이유는 이 사이트가 타 사이트에 비해서 비교적 풍부한 상품평 정보를 보유하고 있고, 인터넷을 통해서 상품평 정보를 비교적 수월하게 수집할 수 있도록 구성되었기 때문이다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구를 제시하고 제 3장에서는 감성분석 절차와 알고리즘을 제시한다. 제 4장에서는 구현 시스템을 소개하고 실험결과를 제시한다. 마지막으로 제 5장에서는 결론을 맺는다.

## 2. 관련연구

상품평의 정확한 요약을 위해 오피니언 마

이닝 기술을 이용한 다양한 연구가 진행되고 있다[1-11]. 상품평을 요약하는 방법으로는 크게 자연어 처리기법과 통계학적 접근법이 있다[14]. [1]에서는 기계 학습(machine learning) 및 자연어 처리 기술을 활용하여, 상품평 데이터에 대한 감성분석 및 분석결과 요약 기법을 제시하고 있으며, 결과물로서 연구목적의 Opinion Observer 라는 명칭의 시스템을 개발하였다. 그러나 실제 시스템 개발을 위해 필요한 인프라적 측면-예를 들어 텍스트 웨어하우스(text warehouse) 및 어휘 온톨로지(ontology)-을 소홀히 하고 있어 상용화 시스템 개발을 위한 방법론 측면으로서는 미흡한 면이 있다.

미국 카네기멜론 대학교에서는 RedOpal 시스템을 개발한 사례가 있으며[4], 이는 상품평 데이터와 사용자 평가점수를 활용하여 요약 보고서를 생성하는 기법을 제안하였다. 이 연구에서는 상품 속성과 평가 점수에 대하여 다차원 분석 결과를 보여주고 있지만, 주관적 긍정/부정 평가를 수행하지는 않고 있다.

[8]에서는 문장 구조와 문장 사이의 관계, 문장성분의 패턴 정보 등의 언어 규칙을 이용한 통계학적 방법으로 오피니언 마이닝에 접근하고 있으며, [10]에서는 워드넷(WordNet)을 활용하여 어휘의 긍정이나 부정적 의미를 판단하고, 이를 센티워드넷(SentiwordNet)으로 응용하여 감정의 폭을 정량화하는 방법을 제시하고 있다.

[7]에서는 사용자의 질의어에 대해 가장 관련 있는 상품평의 우선순위를 정하는 기법을 제안하였다. 이 방법에서는 주관적 혹은 객관적 상품평인가를 고려하지 않고 타 사용자들의 해당 상품평이 얼마나 도움이 되는가를 정량적으로 평가한 수치를 가지고 상품평의 가치를 계산하는 방법을 이용하였다.

한글 상품평에 관련해서 [15]에서는 한국어 문법구조와 의미 사전을 이용했다는 점에서 본 논문의 접근방법과 유사한 측면이 있다. 그러나 이 논문에서는 대체로 문법이 제대로 지켜지지 않고 문장 분리가 쉽지 않은 한글 상품평들의 특징을 고려하지 않고 있다. 반면에 본 논문에서는 자유도가 높은 한글의 특성을 고려하여 상품평 내에서 상품의 속성과 감성 어휘의 추출에 초점을 맞춘 상품평 모델의 정의하였으며, 이를 기반으로 상품평의 극성을 판단하는 알고리즘을 제안하였다.

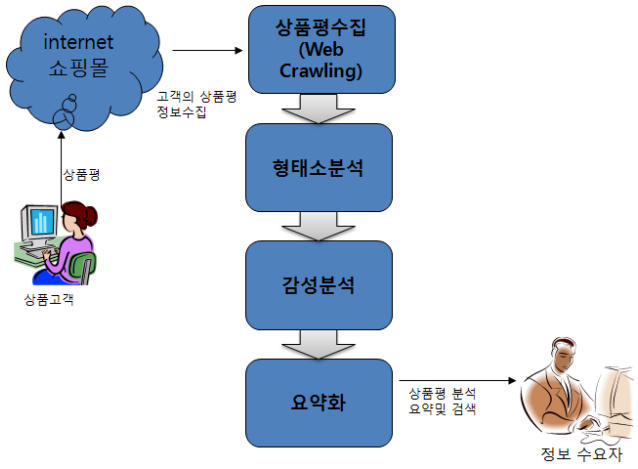
## 3. 상품평 자동분류

본 장에서는 오피니언 마이닝 기법을 활용하여 상품평을 자동분류하기 위한 과정을 설명하고, 이 과정의 핵심인 감성분석 알고리즘을 제시한다.

### 3.1 상품평 자동분류 과정

<그림 1>은 본 논문에서 가정하는 상품평 자동분류 과정을 보여준다. 우선 상품평 수집 단계는 상품평들을 수집하여 데이터베이스에 저장 관리하는 기능을 한다. 대부분의 상품평은 인터넷상의 온라인 쇼핑몰에 산재해 있으므로 웹 크롤링(web crawling) 기법을 이용하여 일괄적 혹은 실시간으로 수집하게 된다.

다음 단계로 형태소 분석 과정은 한글로 작성된 상품평을 각 문장별로 분리한 후, 각 문장에 대해서 형태소 분석을 통해서 단어들




〈그림 1〉 상품평 분석 과정

추출한다. 형태소 분석을 통해 추출된 단어들에는 품사가 부여되는데, 이들 중에서 감성분석에 필요한 품사들만을 별도로 모아서 관리하게 된다. 일반적으로 상품에 대한 속성들은 명사로 표현되며, 감성을 표현하는 대부분의 품사들은 형용사, 부사, 동사 등이다. 한글의 형태소 분석기의 결과는 이들 품사 이외에도 관형사, 감탄사, 조사, 수사 등의 품사들이 있지만 이들에 해당하는 단어들은 작성자의 감성을 표현과는 대부분 관련이 없으므로 고려대상에서 제외될 수 있다.

다음으로 감성분석 단계에서는 문장과 단어별로 분리된 각 한글 형태소에 대해서 상품의 속성별로 감성의 극성과 그 정도를 판단한다. 이를 위해서는 각 상품의 속성과 감성 단어에 대한 정보가 요구된다. 우선 상품 속성은 상품들의 성격에 따라 결정될 수 있다. 예를 들어 디지털 카메라와 레이저 프린터를 비교해 볼 때 크기, 가격, 무게, 디자인 등은 두 제품군에 공통적으로 적용할 수 있는 일반적인 속성이다. 반면에 배터리, 렌즈, 액정, 셔터스피드, 플래시 등은 디지털 카메라만의 고유의 속성이며, 인쇄속도, 토너 등은 레이저 프린터만의 속성이다. 이러한 상품의 속성 정보는 상품 카탈로그 등과 같이 외부로부터 직접적으로 구축할 수도 있고, 대용량의 상품평으로부터 기계학습을 통해 자동으로 구축될 수도 있다[1].

감성단어는 “예쁘다”, “좋다”, “빠르다” 등과 같이 대부분 부사, 동사, 형용사 등으로 구성된다. 감성 단어에 대한 데이터베이스 구축을 위해서는 크게 두 가지 측면을 고려해야한다. 첫째는 감성단어는 감성의 극성과 극성의 정도를 관리해야한다. 예를 들어 “예쁘다”, “엉망이다”는 각각 긍정과 부정의 의미를 담고 있다. 또한 단순히 “좋다”보다는 “환상적이다”란 단어가 더 강한 극성을 갖고 있다. 둘째로 일부 감성단어들은 상품의 속성과 연관되어 있다는 것이다. 예를 들어 “크다” 혹은 “작다”는 속성에 따라 긍정적 감정이 될 수도 있고 그 반대가 될 수도 있다. 따라서 경우에 따라 속성과 감성단어에 대한 연관정

보도 관리해야한다. 본 논문에서는 이와 같은 상품의 속성과 감성 단어에 대한 정보가 관리된다는 가정 하에 감성분석 알고리즘을 제시한다.

마지막으로 요약화 과정은 감성분석 단계에서 얻을 결과를 기반으로 통계 등의 기법을 이용하여 요약된 결과를 저장하게 된다. 이렇게 저장된 정보는 사용자의 검색이나 요청에 대응하여 시각화된 정보를 제공하게 된다.

## 3.2 상품평 모델

텍스트로 구성된 상품평은 언어학적 관점에서 문장의 집합으로 정의할 수 있다. 하지만 상품평의 모든 문장들이 상품에 대한 감성을 표현하는 것은 아니다. 따라서 감성분석의 대상이 되는 문장만을 선별하는 과정이 필요하다. 또한 한글은 자유도 높아 마침표(.)만으로 문장을 분리하는 것도 쉽지 않은 실정이다. 따라서 본 논문에서는 상품평의 각 문장들은 특정 속성에 관련되었다는 가정 하에 다음과 같은 논리적 상품평 모델을 정의한다.

일반적으로 상품의 속성에 대한 사용자들의 감정표현은 속성과 감성단어가 연속으로 나타나게 된다. 예를 들어, “디자인이 예쁘고 훌륭해요”라는 문장에서 “디자인”은 속성에 해당하고 “예쁘다”와 “훌륭하다”는 감성어휘가 된다. 이와 같이 속성과 감성어휘가 연속으로 나타나는 부분을 본 논문에서는 다음과 같이 속성문장(feature sentence)이라고 정의한다.

• 정의 : 속성 문장

$F$가 상품평에 대하나 속성 집합이고, $S$가 감성어휘 집합이고 가정하자. 이 때 속성 $f \in F$와 감성어휘 리스트 $s_1, \cdots, s_m (s_j \in S\ (1 \le j \le n))$가 $f, s_1, \cdots, s_m$와 같은 형태로 구성될 때, 이를 속성문장이라 한다.

예를 들어, “디자인이 예쁘고 훌륭해요”는 “디자인”에 대한 속성문장이 된다. 또한 실제 하나의 물리적인 문장 내에 두 개 이상의 속성들이 등장할 수 있는데 이 경우 해당 문장은 각 속성들에 대한 속성문장으로 중복되어 정의될 수 있다. 상품평에 대한 형태소 분석결과로 상품평 내의 특정 문장에 $f_1, \cdots, f_n, s_1, \cdots, s_m$와 같은 패턴이 나타난다고 가정하자.[1] 여기서 $f_i (1 \le i \le n)$는 상품에 대한 속성들이고 $s_j (1 \le j \le m)$는 감성어휘들이다. 그러면 이 부분은 다음과 같은 $n$개의 속성문장으로 분리될 수 있다.

$$\begin{matrix} f_1,\ s_1,\ \cdots,\ s_m \\ \vdots \\ f_n,\ s_1,\ \cdots,\ s_m \end{matrix}$$

예를 들어 “디자인과 화질이 모두 좋아요”라는 문장은 각각 “디자인이 좋아요”와 “화질이 좋아요”와 같이 두 개의 속성문장이 된다.

또 다른 경우로 하나의 물리적인 문장 내에 $f_1, s_1, \cdots, f_n, s_n$와 같은 패턴이 나타난다고 가정할 때, 이 문장은 다음과 같이 여러 개의 속성문장으로 분리된다.

$$\begin{matrix} f_1,\ s_1 \\ \vdots \\ f_n,\ s_n \end{matrix}$$

1) 감성분석과 관련 없는 형태소는 제외한다고 가정한다.

예를 들어 “디자인은 좋지만, 화질이 나빠요”라는 문장은 “디자인은 좋아요”라는 문장과 “화질은 나빠요”라는 두개의 논리적인 속성문장으로 나누어 정의할 수 있다. 이와 같은 가정 하에 상품평은 논리적으로 속성문장들의 집합이라고 정의할 수 있다. 따라서 상품평 내에 특정 속성과 관련 없거나 감성어휘를 포함하지 않는 문장이나 패턴들은 분석대상에서 제외된다.

다음으로 감성어휘에 대해서는 감성단어와 더불어 긍정/부정을 나타내는 극성 및 그 정도를 나타내는 점수가 부여된 사전이 구축되었다고 가정한다. 긍정의 경우에는 양수의 값을 갖게 되며, 부정은 반대로 음수를 갖는다. 또한 극성의 정도가 크면 절대값도 커지게 된다. 예를 들어 “좋다”와 “나쁘다”는 각각 긍정과 부정의 의미를 갖고 있으므로 +1, -1과 같은 값을 갖게 된다. “좋다”와 “환상적이다”는 극성의 정도가 다르므로 각각 +1, +2와 같이 절대값에 차이를 갖는다. 이와 같은 감성어휘에 대한 사전은 수작업으로 구축할 수도 있고, 기계학습을 통해 자동적으로 구축될 수도 있다[1]. 본 논문에서는 실제 구현을 위해 수작업과 기계학습 모두를 이용해서 구축하였으며 그 결과는 제 4장에서 기술한다.

### 3.3 감성분석 알고리즘

감성분석의 목표는 주어진 상품평을 문장별로 분석하여 상품의 속성별로 감성의 극성과 그 정도를 판단하는 것이다. 본 논문에서는 3.2절에서 가정한 상품평 모델을 기반으로 감성분석 알고리즘을 제시한다. 알고리즘은 한글로 작성된 상품평의 감성 표현 중에서 전형적인 패턴들을 고려하여 설계하였다. 한글과 같이 자유도가 높은 언어에서 모든 가능한 감성 표현을 완벽하게 추출하는 것은 거의 불가능하다. 특히 한글은 문법구조가 복잡하여 표준문법에 맞게 작성된 상품평은 거의 찾아볼 수 가 없는 실정이다. 또한 신조어, 줄임말 등 인터넷에서의 한글 파괴현상으로 한글에 대한 처리가 갈수록 까다로워지고 있다. 따라서 본 논문에서는 문법구조를 분석하기 보다는 상품에 대한 속성과 감성을 표현하는 단어들에 중점을 두고 알고리즘을 설계하였다.

제시된 알고리즘은 <그림 1>의 상품평 자동분석 과정 중에서 감성분석 단계의 핵심 알고리즘으로, 상품의 속성별로 감성분석을 실시하여 그 결과를 출력한다. 본 논문에서 제시하는 감성분석 알고리즘은 <그림 2>와 같다. 이 알고리즘에서 입력 데이터는 상품평이다. <그림 1>의 형태소 분석 단계에서는 텍스트 형태의 상품평을 형태소 별로 분류하는데, 감성분석 알고리즘에서는 형태소 단위로 구성된 형식의 상품평을 입력받게 된다.

이 알고리즘의 첫 단계에서는 3.2절에서 기술한 상품평 모델에 기반하여 물리적 문장들을 논리적 속성문장들로 분류한다. 속성문장으로 분류하게 되면 문장에 속한 각 감성어휘들이 어떠한 속성에 대한 표현인지 탐색할 필요성이 없어지므로 알고리즘의 복잡도가 낮아지는 장점이 있다.

다음 단계에서는 각 속성문장별로 감성의 극성과 정도를 계산한다. 감성의 극성은 형태소들의 품사에 따라 결정된다. 우선 감성어휘는 대부분 형용사와 동사로 구성되며 일부에서는 명사가 나탈 수 도 있다. 예를 들어 “좋

다", "나쁘다"와 같이 직설적인 감성어휘들은 대부분 형용사로 구성된다. 뿐만 아니라 "수준급이다"와 같이 경우에 따라 명사도 감성어휘가 될 수 있다. <그림 2>에서 score[$f_i$]은 속성 $f_i$에 대한 감성평가 결과를 저장하는 자료구조를 나타내며, 이 값은 감성어휘가 나타날 때마다 매번 누적된다. 또한 grade[$w$]는 감성단어 $w$의 긍정/부정을 나타내는 극성 및 그 정도를 저장한 값으로 감성어휘 사전에 저장된 값을 사용한다. 감성어휘들은 "매우", "정말로" 등과 같이 부사로 강조될 수 있는데, 이 경우에는 감성어휘에 가중치를 부과하게 된다. 가중치 역시 감성어휘 사전에서 관리되는데 <그림 2>에서 weight[강조부사]는 강조부사에 대한 가중치를 의미한다. 가중치는 정도에 따라 1보다 크거나 같은 값으로 정의되는데 1은 가중치가 없다는 것을 의미하며, 강조 정도에 따라 1보다 큰 값이 부여될 수 도 있다. 누적된 가중치는 감성어휘에 적용된 후에는 다음의 감성어휘를 위해 다시 초기화 된다.

마지막으로 부정어는 크게 부정을 나타내는 부사와 동사로 구분되는데 "안 좋다"와 같이 부사로 나타낼 수도 있으며, "좋지 않다"와 같이 동사로 표현 될 수 있다. 부정부사에 대해서는 가중치를 이용하여 적용할 수 있으

Sentiment Analysis Algorithm

입력 : 형태소별로 분류된 상품평
출력 : 속성별 감성분석 결과

입력된 상품평을 속성문장의 집합($fs_1$,...,$fs_n$)으로 분류
BEGIN
  for each $fs_i(1 \le i \le)$ {
    score[$f_i$] = 0;  // $f_i$는 $fs_i$의 속성
    가중치 = 1;
  }
  for each $fs_i(1 \le i \le)$ {
    for each word $w$ in $fs_i$ {
      if($POS(w)$ = 강조부사)  // $POS(w)$는 $w$의 품사
        가중치 = 가중치 × weight[강조부사]
      if($POS(w)$ = 감성어휘)
        score[$f_i$] = score[$f_i$] + grade[$w$] × 가중치
        가중치 = 1
      if($POS(w)$ = 부정부사)
        가중치 = -1
      if($POS(w)$ = 부정동사)
        score[속성] = -score[속성]
    }
  }
END

〈그림 2〉 감성분석 알고리즘

며, 부정동사의 경우에는 현재까지 계산된 결과를 역으로 재계산하게 된다.

예를 들어 다음과 상품평을 가정하자.

> 오늘 카메라 받았습니다. 우선 모양이 작고 예뻐요. 액정도 크고 화질도 깔끔하고 좋습니다. 다만 배송이 너무 느립니다. 사용법은 아직 잘 모르겠고요. 공부해야죠.

이 상품평에서 첫 문장과 마지막 문장은 카메라의 속성과 관련 없는 것들로 일단 제외된다. 나머지 문장들에 대해서 속성문장으로 분류하면 다음과 같은 4개로 분류된다.

| 속성 | 분류된 형태소 |
|---|---|
| 모양 | “작다” “예쁘다“ |
| 액정 | “크다” |
| 화질 | “깔끔하다” “좋다” |
| 배송 | “너무” “느리다” |
| 사용법 | “잘” “모르다” |

분류된 형태소 중에서 대부분의 형태소들은 형용사로서 감성어휘들이고, “너무”와 “잘”은 강조부사이다. 이 표의 감성어휘 중에서 “느리다”는 -1의 값을 갖고 나머지는 +1을 갖으며, 강조부사인 “너무”와 “잘”의 가중치는 2라고 가정하자. 그러면 <그림 2>의 알고리즘에 의해 계산된 감성분석 결과는 다음과 같다.

$$score[모양] = grade[작다] + grade[예쁘다] = 1 + 1 = 2$$

$$score[액정] = grade[크다] = 1$$

$$score[화질] = grade[깔끔하다] + grade[좋다] = 1 + 1 = 2$$

$$score[배송] = weight[너무] * grade[느리다] = 2 \times -1 = -2$$

$$score[사용법] = weight[잘] * grade[모르다] = 2 \times 0 = 0$$

이 계산에 따라 모양, 액정, 화질은 긍정의 극성을 갖게 되며, 반대로 배송은 부정의 극성을 갖게 된다. 사용법에 대해서는 “모르다”라는 단어가 감성어휘가 아니므로 중립적인 0의 값을 갖게 된다.

## 4. 구현 및 실험

### 4.1 구 현

본 논문에서는 제 3장에서 제시한 시스템 구성도와 감성분석 알고리즘을 기반으로 상품평 자동분석 시스템을 개발하였다. 프로그램 언어로는 Java와 JSP를 사용하였고, DBMS는 MySql 5.0을 사용하였다. 형태소 분석을 위해 국내의 대표적인 형태소분석기인 KLT [16]를 사용하였다.

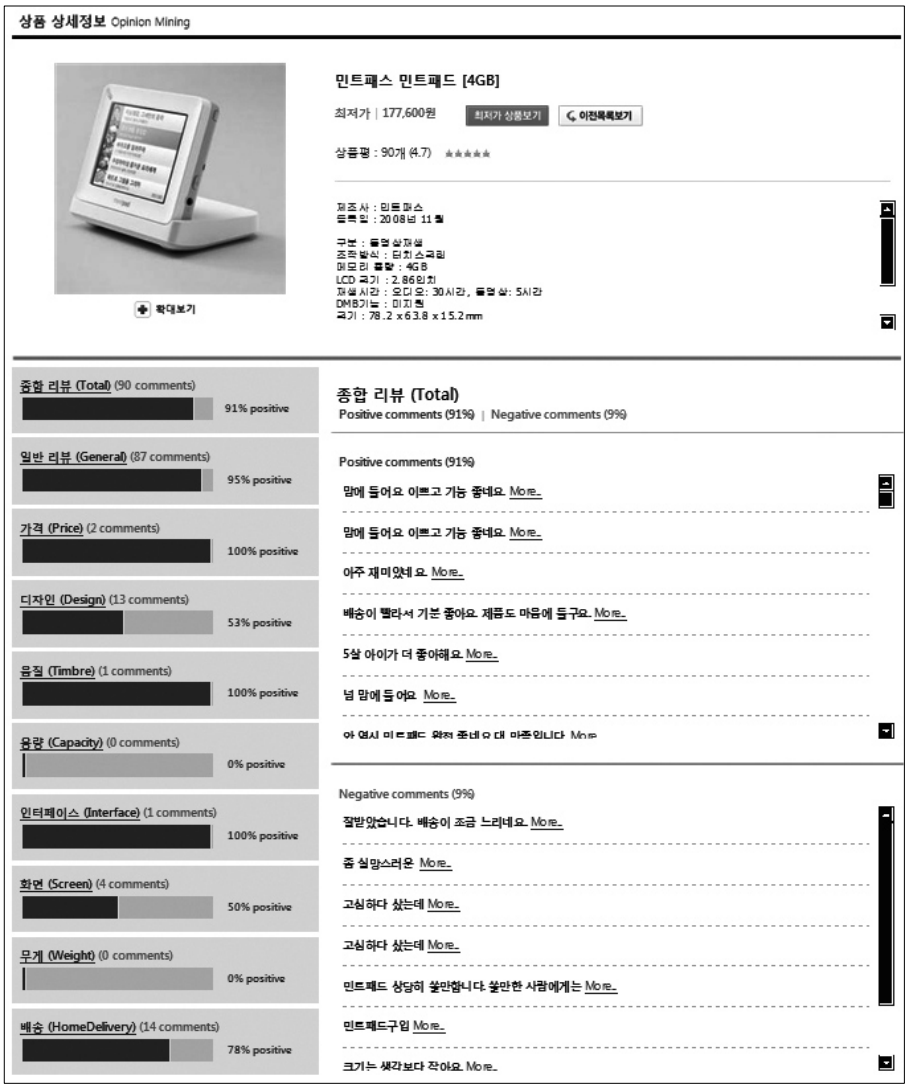
감성어휘 사전을 구축하기 위해서 두 가지 접근 방법을 활용하였다. 첫째는 기본적인 상품의 속성 사전을 구축하기 위해서 상품평 데이터 수집기에서 수집된 상품에 대한 기본 정보 중에서 상품의 속성을 자동 추출하였다. 예를 들어 PMP제품의 경우 상품에 대한 상세정보에는 동영상, 조작방식, 터치스크린, 메모리 용량, LCD, 재생시간, 동영상, 오디오, DMB, 크기 등에 대한 설명이 있으며, 이 정보들이 기본적인 속성들이 될 수 있다. 그러나 이러한 정보만으로는 제품에 대한 속성들

이 모두 얻어진다고 볼 수 없다. 따라서 두 번째 과정으로 상품평 텍스트 정보로부터 숨겨진 속성을 추출하였다. 본 논문에서 개발한 방식은 자주 출현하는 명사단어를 후보속성으로 구분하는 것이다. 이는 중요한 속성은 상품에 자주 출현하는 명사단어일 확률이 높기 때문이다.

<그림 3>은 네이트 쇼핑몰에서 PMP 상품군에 대해서 상품정보와 상품평 정보를 웹 크롤링 기법을 이용하여 수집한 후, 가상의 쇼핑몰을 구성한 화면이다. <그림 3>에서 각 개별 상품리스트들의 가장 오른쪽 부분은 감성분석 결과를 요약하여 보여주고 있다. 예를 들어 가장 상단에 있는 TG삼보 CP-100의 경우를 보면 전체의 상품평 중에 72%가 긍정적인 반응을 보인 반면 나머지 28%는 부정적이거나 특별한 의견이 없는 상품평들이다. 이 화면에서 나타난 감성분석 결과는 상품의 속성과 관련 없이 전반적인 상품평의 극성에 대한 평가 결과를 보여준다.

다음으로 <그림 4>는 특정 상품에 대한 구체적인 상품평 분석 결과를 보여주는 화면이다. 이 화면은 세 가지 부분으로 나뉘어져 있는데 상단은 상품에 대한 일반적인 내용을 보여주고 있고, 왼편은 상품평에 대한 감성분석 결과가 그래프로 표현되어 있다. 각 그래프는 속성별로 나뉘어져 있고 속성마다 상품평의 긍정/부정 비율이 표현되어 있다. 만약


〈그림 3〉 PMP 상품군 리스트




〈그림 4〉 상품평 분석결과

특정 속성에 해당하는 상품평이 10개가 등록되어있고 그중 7개의 상품평이 긍정을 나타내고 3개의 상품평이 부정을 나타내면 그래프의 바(bar)는 70%를 나타내게 된다. 속성은 총 10개로 구분 되어있고 종합, 일반, 가격, 디자인, 음질, 용량, 인터페이스, 화면, 무게, 배송으로 구성 되어있다. 여기서 종합은 속성과 관련 없이 전체적인 감정의 극성을 나타낸 것이며, 일반은 다른 8개의 속성이 포함되지 않은 감정의 극성을 나타낸다. 이 그래프 중에서 특정 속성을 선택하게 되면 해당 속성에 대한 상품평을 확인할 수 있다. 그 결과는 화면의 오른편에 나타나는데, 긍정과 부정으로 나뉘어져 있고 각각에 대해서 감정

의 극성이 큰 상품평부터 정렬되어 출력된다.

## 4.2 실 험

본 논문에서 개발한 감성분석 알고리즘의 정확도를 평가하기 위해 네이트 쇼핑몰에서 PMP군에 있는 상품들의 상품평에 대한 분류 정확도를 측정하였다. 상품평은 총 500개를 임의 선택하여 평가를 실시하였다. 우선 500개의 상품평에 대해서 수동으로 긍정 및 부정적 상품평을 분류하고, 본 과제에서 개발한 감성분석기에서 분류된 결과와 비교하였다. 또한 개별 속성에 국한된 경우에 대해서도 평가를 실시하였다. 평가 결과는 각각 <표 1>, <표 2>와 같다.

우선 <표 1>은 종합적 상품평에 대한 결과이다. <그림 4>에서 대상평가수는 전체 상품평을 수동적으로 긍정/부정을 결정한 결과이다. 이를 바탕으로 정분류수는 본 논문에서 개발한 감성분석 알고리즘을 통해서 수동평가와 일치하는 결과를 보인 상품평의 수이며 오분류수는 수동평가와 감성분석기의 결과가 다른 상품평의 수이다.

<표 1>에서 보는 바와 같이 긍정/부정 상품평에 대한 감성분석 결과의 정확도는 각각 85% 76%이다. 그리고 전체적인 정확도는 83% 수준이다. 비록 많은 양의 데이터에 대한 평가도 아니고 PMP라는 특정 상품군에 대해 제한적으로 실시된 면이 있지만 비교적 정확한 평가가 이루어진 것으로 판단할 수 있다.

<표 2>는 특정 속성(여기서는 배송)에 대한 분류 정확도의 평가결과이다. <표 2>에서 보는 바와 같이 배송에 관련된 긍정/부정 상품평에 대한 감성분석 결과의 정확도는 각각 82% 75%이다. 그리고 전체적인 정확도는 80% 수준이다. 따라서 개별적 속성에 대한 분류 정확도도 종합적 상품평에 대한 정

〈표 1〉 종합적 분류정확도 평가

| | 긍정 | 부정(중립) | 총계 |
|---|---|---|---|
| 대상 평가수 | 382 | 118 | 500 |
| 정분류수 | 325 | 90 | 415 |
| 오분류수 | 57 | 28 | 85 |
| 정확도 | 85% | 76% | 83% |

〈표 2〉 '배송'속성에 대한 분류정확도 평가

| | 긍정 | 부정(중립) | 총계 |
|---|---|---|---|
| 대상 평가수 | 326 | 174 | 500 |
| 정분류수 | 267 | 130 | 397 |
| 오분류수 | 59 | 44 | 103 |
| 정확도 | 82% | 75% | 80% |

확도와 유사한 결과를 보이고 있다.

이 평가에서 오분류된 상품평의 대부분은 문법이나 띄어쓰기가 많은 부분에서 오류가 있거나 사전에는 등장하지 않는 신조어 등을 사용한 것이 대부분이었다. 또한 부정 상품평에 대한 분류 정확도가 상대적으로 낮은 이유는 부정적인 상품평의 경우 많은 사용자들이 직접적인 부정어들을 사용하기 보다는 추상적으로 부정적인 감정을 표현한 것이 많아 감성분석기가 이를 올바르게 판단하지 못했기 때문이다.

## 5. 결 론

본 논문에서는 오피니언 마이닝 기술을 이용하여 제품 사용자의 주관적 의견을 자동으로 분류할 수 있는 감성분석 알고리즘을 제안하였다. 제안된 알고리즘은 한글의 특성을 고려하여 상품평을 분석하여 속성과 감성어휘 등을 추출하고, 감성의 극성정도를 판단하는 기능을 한다. 또한 제안된 알고리즘을 기반으로 개발된 상품평 자동분석 시스템 개발하였다. 이 시스템은 온라인 쇼핑몰에 등록된 상품평을 대상으로 제품 사용자의 의견을 자동을 분석 요약하여 그 결과를 제공하는 기능을 한다. 제안된 알고리즘의 효율성을 검증하기위해서 인터넷 유명 쇼핑몰을 대상으로 실험을 실시하였다. 실험결과 평균적으로 80%이상의 정확도를 보였으며 이 수치는 기존의 연구에서의 정확도와 유사한 것으로 평가된다.

최근 들어 인터넷 마케팅이 전체 광고시장에서 획기적인 성장을 이루고 있는 점을 감안할 때 상품평에 대한 자동분석 시스템은 인터넷 쇼핑몰에서의 마케팅이나 포털 사이트의 검색 광고 분야에 상당한 영향을 미칠 것으로 예상된다. 향후에는 감성분석 평가의 정확도를 개선하여 실제 상용화 가능한 수준의 상품평 분석 시스템을 개발할 계획이다.

## 참 고 문 헌

[1] Liu, B., Hu, M., and Cheng, J., “Opinion observer : analyzing and comparing opinions on the Web,” Proceedings of the 14th international conference on WWW, pp. 10-14, 2005.

[2] Narayanan, R., Liu, B., and Choudhary, A., “Sentiment Analysis of Conditional Sentences,” Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-09), August 6~7, 2009, Singapore.

[3] Liu, B., “Sentiment Anlaysis and Subjectivity,” Invited Chapter for the Handbook of Natural Language Processing, Second Edition, To appear in Oct/Nov, 2009.

[4] Sca±di, C., Bierho®, K., Chang, E., Felker, M., Ng, H., and Jin, C., “Red Opal : Product-Feature Scoring from Reviews,” Proceedings of the 8th ACM conference on Electronic commerce, 2007, pp. 11-15.

[5] Hu, M., and Liu, B., “Mining and sum-

marizing customer reviews," Proceedings of the tenth ACM SIGKDD 04, 2004, pp. 22-25.

[6] Smrz, P., "Using WordNet for Opinion Mining," Proceedings of the Third International WordNet Conference (GWC 2006), pp. 333-335.

[7] Miao, Q., Li, Q., and Dai, R., "A sentiment mining and retrieval system," Expert Systems with Applications, Vol.36, 2009, pp. 7192-7198.

[8] Xiaowen Ding, and Bing Lui, "The Utility of Linguistic Rules in Opinion Mining," SIGIR 2007, pp. 811-812.

[9] Esuli, A., and Sebastiani, F., "Page-Ranking WordNet Synsets : An Application to Opinion Mining", 2007 Association for Computational Linguistics, 2007, pp. 424-431.

[10] Courses, E., and Surveys, T., "Using SentiWordNet for multilingual sentiment analysis," Data Engineering Workshop ICDEW 2008.

[11] Cover, T. M., and Thomas, J. A., "Elements of information theory", Wiley, New York, 1991, pp. 12-14.

[12] http://www.amazon.com.

[13] http://live.com.

[14] 양정연, 명재석, 이상구, "상품특징별 점수화를 이용한 상품리뷰요약 시스템의 설계 및 구현", 지식정보산업연합학회 창립기념 학술대회, 2008, pp.339~347.

[15] 명재석, 이동주, 이상구, "반자동으로 구축된 의미사전을 이용한 한국어 상품평 분석 시스템", 정보과학회논문지 : 소프트웨어 및 응용 제35권, 제6호, 2009.

[16] 강승식, 한국어 형태소 분석과 정보 검색, 홍릉과학출판사, 2003.

## 저 자 소 개


| | |
|---|---|
| 장재영 | (E-mail: jychang@hansung.ac.kr) |
| 1992년 | 서울대학교 계산통계학과(학사) |
| 1994년 | 서울대학교 계산통계학과 전산과학전공 대학원(석사) |
| 1999년 | 서울대학교 계산통계학과 전산과학전공 대학원(박사) |
| 2000년~현재 | 한성대학교 컴퓨터공학과 부교수 |
| 관심분야 | 데이터베이스, 데이터마이닝 |