

# 온라인 리뷰 데이터의 오피니언마이닝을 통한 콘텐츠 만족도 분석 시스템 설계<sup>☆</sup>

## A Design of Satisfaction Analysis System For Content Using Opinion Mining of Online Review Data

김문지<sup>1</sup>      송은정<sup>1</sup>      김윤희<sup>1\*</sup>  
MoonJi Kim    EunJeong Song    YoonHee Kim

### 요약

소셜 네트워크 서비스(SNS)의 활성화로 웹상에는 방대한 양의 온라인 리뷰들이 생산되고 있으며, 이러한 온라인 리뷰들은 다양한 콘텐츠들에 대한 의견 데이터로서 콘텐츠 이용자와 제공자들에게 가치 있는 정보로 활용되고 있다. 한편, 온라인 리뷰에 대한 중요도가 높아짐에 따라 온라인 리뷰를 분석하여 글쓴이의 의견이나 평가, 태도, 감정 등을 추출해 내는 오피니언마이닝에 대한 연구가 활발하게 진행되고 있다. 그러나 기존의 오피니언마이닝 연구들에서는 리뷰의 의견 분류에만 초점을 맞추어 감성 분석 기법을 설계하였기 때문에 리뷰 속에 내포되어있는 작성자의 자세한 만족도까지는 알 수 없었으며, 감성 분석 기법이 특정 콘텐츠에 한정되어있어 도메인이 같지 않은 다른 콘텐츠들에는 적용될 수 없다는 문제점이 있었다. 이에 본 연구에서는 기존 의견 분류 방법에 강도를 주어 좀 더 세밀한 감성 분석을 수행하고, 이 결과를 통계적 척도에 적용하여 리뷰에 내포되어 있는 작성자의 자세한 만족도를 도출할 수 있는 감성 분석 기법을 제안한다. 그리고 제안한 기법을 바탕으로 도메인에 상관없이 다양한 콘텐츠에 적용되어 콘텐츠의 만족도를 분석할 수 있는 시스템을 설계하였다. 또한 방대한 양의 리뷰 데이터들을 빠르고 효율적으로 처리하기 위해 빅 데이터 처리도구인 하둡을 기반으로 시스템을 구축하였다. 본 시스템을 통해 콘텐츠 이용자는 보다 효율적인 의사결정을, 제공자들은 빠른 반응분석을 할 수 있어 본 시스템은 사용자의 의견을 필요로 하는 다양한 분야에 매우 실용적으로 활용될 것으로 기대한다.

☞ 주제어 : 오피니언마이닝, 만족도 분석, 하둡, 온라인 리뷰

### ABSTRACT

Following the recent advancement in the use of social networks, a vast amount of different online reviews is created. These variable online reviews which provide feedback data of contents' are being used as sources of valuable information to both contents' users and providers. With the increasing importance of online reviews, studies on opinion mining which analyzes online reviews to extract opinions or evaluations, attitudes and emotions of the writer have been on the increase. However, previous sentiment analysis techniques of opinion-mining focus only on the classification of reviews into positive or negative classes but does not include detailed information analysis of the user's satisfaction or sentiment grounds. Also, previous designs of the sentiment analysis technique only applied to one content domain that is, either product or movie, and could not be applied to other contents from a different domain. This paper suggests a sentiment analysis technique that can analyze detailed satisfaction of online reviews and extract detailed information of the satisfaction level. The proposed technique can analyze not only one domain of contents but also a variety of contents that are not from the same domain. In addition, we design a system based on Hadoop to process vast amounts of data quickly and efficiently. Through our proposed system, both users and contents' providers will be able to receive feedback information more clearly and in detail. Consequently, potential users who will use the content can make effective decisions and contents' providers can quickly apply the users' responses when developing marketing strategy as opposed to the old methods of using surveys. Moreover, the system is expected to be used practically in various fields that require user comments.

☞ keyword : Opinion mining, Satisfaction Analysis, Hadoop, Online Review

## 1. 서론

인터넷과 소셜 네트워크 서비스(Social Network Service : SNS)의 발달로 웹상은 개인의 의견을 표현하고 공유하는 대표적인 공간이 되었다. 사람들은 SNS를 통해 자신이 경험한 공간, 제품, 서비스와 같은 콘텐츠들에 대한 의견을 작성하고 공유하며, 이러한 방대한 양의 의견데이터

<sup>1</sup> Department of Computer Science, Sookmyung Women's University, Seoul, 04310, Korea

\* Corresponding author (yulan@sookmyung.ac.kr)

[Received 8 March 2016, Reviewed 10 March 2016, Accepted 27 April 2016]

☆ 본 논문은 2015년도 한국인터넷정보학회 추계학술발표대회

우수 논문 추천에 따라 확장 및 수정된 논문임.

이터들은 온라인 리뷰로써 콘텐츠 이용 결정여부를 위한 참고자료가 될 뿐만 아니라 콘텐츠 제공자들에게는 이용자의 의견을 파악할 수 있는 매우 가치 있는 정보로써 활용되고 있다.

이렇게 온라인 리뷰에 대한 활용도와 중요도가 높아짐에 따라, 웹상에 존재하는 방대한 양의 온라인 리뷰들 속에 내포되어 있는 작성자의 의견을 분석하고 요약해주는 오피니언마이닝에 대한 다양한 연구[3-9]들이 활발하게 진행되고 있다. 하지만 기존 연구들에서의 오피니언마이닝은 특정 콘텐츠에 대한 작성자의 의견 분류, 즉 작성자가 해당 콘텐츠에 대해 단순히 긍정/부정의 의견을 갖는지를 분류 하는데 그쳐 리뷰 속에 내포되어있는 작성자의 자세한 만족도까지는 알 수 없었다. 또한 기존에 연구되었던 감성 분석 기법들은 상품, 영화등과 같이 특정한 콘텐츠에 한정되어 도메인이 다른 콘텐츠들에는 적용할 수 없다는 단점이 있다. 이에 본 연구에서는 기존의 의견 분류 방법에 강도를 주어 좀 더 세밀한 의견분류를 수행하는 감성 분석 기법을 적용하고, 이를 통해 도출된 작성자의 의견 극성을 통계적 척도에 적용해 콘텐츠에 대한 작성자의 자세한 만족도를 도출해 내는 시스템을 구축한다. 또한 특정 콘텐츠에 한정된 감성 분석이 아닌, 다양한 콘텐츠들을 대상으로 도메인에 따라 자동적으로 감성 분석을 할 수 있도록 기존연구[9]를 확장한 감성 분석 기법을 제안하여 이를 시스템에 적용하였다. 본 연구에서 제안한 시스템의 구현에는 방대한 양의 리뷰 데이터를 시스템에 안정적으로 수집하고 저장하기 위한 하둡 기반의 HDFS(Hadoop Distributed File System)를 사용하고, 축적된 대용량 리뷰데이터의 빠르고 효율적인 처리를 위해 하둡 맵리듀스를 기반으로 한 감성 분석함수들을 구축하여 분석에 활용하였다[1].

본 논문의 구성은 다음과 같다. 1장 서론에 이어 2장에서는 오피니언마이닝에 대한 관련 연구들을 살펴보고, 3장에서는 시스템의 전체적인 설계와 만족도 분석을 위한 감성 분석 기법들에 대해 설명한다. 4장에서는 설계한 시스템에 대한 구현과 평가를 하며, 마지막으로 5장에서 결론을 맺는다.

## 2. 관련연구

### 2.1 오피니언마이닝

오피니언마이닝은 텍스트를 분석하여 글쓴이의 의견이나 평가, 태도, 감정 등을 추출해 내는 기법[2]으로, 텍

스트의 정확한 오피니언마이닝을 위해서는 문서에 나타난 의견의 극성을 분석하는 감성분석이 가장 핵심적이다. 현재 감성 분석 기법에는 크게 의견의 의미방향을 분류하는 분야와 언어적 자원을 구축하는 분야가 있다. 본 논문에서는 두 개의 분야를 모두 적용하여 감성분석을 한다.

### 2.2 감성 분석

이 절에서는 감성분석의 기법 중 의미방향을 분류하는 관점에서 관련연구들을 분석한다. 먼저 [3]은 화장품, [4]는 고객의 소리, [5]은 대학교 강의에 대해 미리 구축한 각 도메인별 단어사전과 입력 리뷰 텍스트들을 매칭하여 텍스트 내의 극성을 분석한다. 하지만 [3-5]에서 제안한 기법들은 어휘의 의미 강도를 고려하지 않아 극성 정도에 대한 명확한 기준을 알 수 없다는 한계가 있어 세밀한 감성분석이 어렵다는 단점을 가지고 있다. 한편, [6-7]은 기존연구들과는 달리 어휘의 부사 규칙, 부정어 규칙을 두어 의미적 강도를 파악할 수 있게 하였다. 하지만 [6]의 경우 상품의 속성문장들이 정형화된 문장규칙을 가지지 않는 경우에는 적용할 수 없고, [7] 또한 연구에서 정의된 문장규칙에 맞게 작성되지 않은 리뷰데이터에는 적용하기 어렵다는 한계를 가지고 있다. 따라서 본 연구에서는 [6-7]의 알고리즘을 참고해서 사용하되, 자유도가 높은 한국어 리뷰의 특징을 고려하여 속성이나 문장규칙에 대한 고려를 하지 않고 감정을 표현하는 단어에 초점을 맞추어 알고리즘을 다시 설계하였다. 또한 기존의 감성 분석 알고리즘의 결과는 단순히 긍정/부정에 대한 의견분류에 그친데 반해, 본 논문에서 설계한 알고리즘은 도출된 극성 값들을 통계적 척도에 적용하여 리뷰데이터에 내포되어 있는 작성자의 만족정도를 세밀하게 보여주었다.

### 2.3 감성 사전

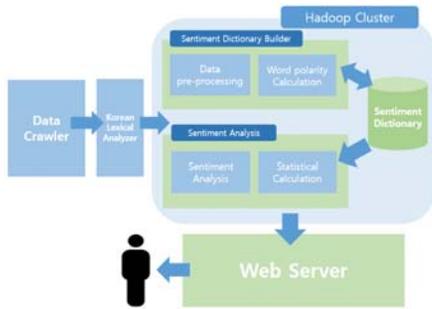
감성 분석을 위한 언어적 자원을 구축하는 연구에는 [7-9]가 있다. 먼저 [7]에서는 ‘회노애락’에 대한 동의어 사전을 이용해 사전을 구축하였다. 하지만 이는 신조어에 대한 감성단어를 알 수 없으며, 도메인의 특징을 고려하지 않아 정확성이 떨어진다는 단점을 가지고 있다. [8]에서는 평점이 포함된 리뷰에서 평점을 기반으로 긍정/부정을 판단하여 사전을 구축한다. 하지만 [8]은 평점이 없는 리뷰의 경우 사전구축이 어렵다는 한계를 가지고

있다. 한편, [9]에서는 리뷰에 나타나는 모든 단어 사이의 밀접도를 구해 그래프를 구성하고, 미리 정의된 감성 어휘(seed어휘)들을 그래프 전체에 전파하는 방식으로 모든 어휘의 감성을 추론하여 사전을 구축한다. [9]에서 제안하는 방식은 도메인 특징을 고려한 사전을 구축할 수 있다는 장점이 있으므로 본 논문에서는 [9]를 적용하여 콘텐츠 별 감성사전을 구축한다. 또한 시간이 지남에 따라 신조어, 은어로 인해 시드어휘가 바뀔 수 있다는 가능성을 고려하여 시드어휘를 갱신시키는 방법을 추가로 제안한다.

### 3. 콘텐츠 만족도 분석 시스템

#### 3.1 시스템 설계

콘텐츠 리뷰데이터의 오피니언마이닝을 위해 본 논문에서 설계한 시스템의 구조는 그림 1과 같다.



(그림1) 시스템 구조도

(Figure 1) System Architecture

전체 시스템은 대용량의 SNS 리뷰데이터를 수집하기 위한 데이터 크롤러(Data Crawler), 수집된 데이터를 형태소 단위의 표현으로 분류하기 위한 형태소 분석기, 콘텐츠 도메인별 감성사전을 구축하기 위한 사전구축기, 리뷰데이터의 오피니언 마이닝을 통해 만족도를 도출하기 위한 감성분석기와 마지막으로 사용자에게 분석결과를 보여주는 결과화면으로 나뉜다. 시스템 각 기능에 대한 자세한 설명은 다음과 같다. 먼저 데이터 크롤러에서는 분석의 대상이 될 콘텐츠를 선정하고, 이를 키워드로 하여 웹 사이트로부터 대량의 리뷰데이터를 주기적으로 파싱해 온다. 이렇게 파싱된 대량의 리뷰데이터들은 형태소 분석 단계를 거쳐 하둠 분산 파일 시스템에 병렬 저장되며, 저장된 데이터들은 첫 번째로 MapReduce기반의 사

전 구축기(Sentiment Dictionary Builder)에 의해 극성이 분류되어 사전데이터로 구축된다. 이에 대한 자세한 설명은 3.2절에서 한다. 두 번째로는 MapReduce기반의 감성 분석 함수 (Sentiment Analysis)에 의해 입력된 리뷰데이터의 감성이 분석되는데, 자세한 설명은 3.3절에서 한다. 마지막 시각화 단계에서는 리뷰데이터의 분석 결과를 웹 인터페이스를 통해 사용자에게 보여주며 그 실험 결과는 4절에서 설명한다.

#### 3.2 콘텐츠 별 감성 사전 구축

리뷰데이터의 정확한 감성분석을 위해서는 각 도메인에 맞는 감성사전의 구축이 필수적이다. 본 연구에서 목표하는 콘텐츠의 만족도 분석은 다양한 콘텐츠를 대상으로 하는 만족도 분석이므로 콘텐츠의 도메인이 정해지면 그에 따라 자동적으로 감성사전이 구축될 수 있도록 해야 한다. 본 논문에서는 이러한 콘텐츠 별 감성사전을 구축하기 위하여 [9]의 그래프 기반 특정분야 감성사전 구축 방법을 이용하였다.

$$V = \{x_1, \dots, x_i, x_{i+1}, \dots, x_n\}, x_i \in X, n = |X|$$

$$E = \{(x_i, x_j) \mid x_i, x_j \in X, i \neq j\}$$

(그림 2) 단어그래프 정의

(Figure 2) Definition of The word graph

그림 2은 [9]에서 제안하는 단어 그래프의 정의로 그래프의 정점은 단어, 간선은 단어와 단어 사이의 관계로 정의한다. 그림2의 정점 집합 V는 감성을 알고 있는 소수의 감성단어 집합(SEED집합)과 감성을 알지 못하는 단어집합의 두 개의 하위집합으로 구성되어 있으며, 위 그래프에서 간선의 가중치는 두 정점 사이의 밀접도인 PMI(Point-wise Mutual Information)의 값으로 계산된다. 간선은 밀접도가 0보다 클 때 정의되며, 그 값이 0보다 작으면 두 정점사이의 관계가 없다고 판단되어 간선이 정의되지 않는다. 결과적으로 이러한 정의를 기반으로 하여 단어 그래프가 생성되며, [9]에서 제안한 감성 추론식을 통해 모든 단어의 감성이 구해져 단어사전이 구축된다. 한편, 간선의 수가 많다는 것은 도메인 내에서 그 단어와 관계를 맺는 단어의 수가 매우 많다는 뜻으로, SEED 집합의 단어들은 간선의 수가 다른 어휘들보다 항상 많게 나타나게 된다. 만약 지속적으로 누적되는 데이터들 속에서 기존 SEED 단어들의 간선의 수와 같거나 더 많은 간선의 수를 갖는 단어가 출현한다면, 이는 도메인 내에서 SEED

단어로서의 영향력을 가지고 있다고 할 수 있다. 따라서 본 연구에서는 [9]의 연구를 확장하여 계속해서 갱신되는 단어 그래프들 속에서 기존의 SEED어휘들 보다 간선의 수를 많이 가지게 되는, 즉 다른 단어들과의 연관성이 기존의 SEED보다 더 높은 단어들을 시드 단어집합에 추가로 갱신하여 그래프 구성의 세밀성을 높였다.

### 3.3 콘텐츠 만족도 분석

리뷰데이터로부터 콘텐츠의 만족도를 분석하기 위해서는 리뷰데이터에 대한 세밀한 감성분석이 핵심적이다. 본 논문에서는 리뷰에 나타나 있는 감성표현단어들에 중점을 두고, 본 연구의 목적에 맞게끔 [6-7]의 알고리즘들을 확장하여 콘텐츠 만족도 도출을 위한 감성분석 알고리즘을 설계하였다(그림 3). 설계한 알고리즘에서는 감정 표현 단어의 의미강도를 부여하기 위해 부사어 규칙, 부정어 규칙을 적용하여 극성 분석을 세밀화 하였다. 또한, 기존의 연구들에서 감성분석 시 극성의 결과만을 도출한다는 한계를 보완하여 도출된 극성 결과 값을 통계적 기법에 적용하여 만족도를 계산하였다. 만족도의 계산은 특정 대상에 대한 개인의 태도를 측정하는 통계적 기법인 리커트 5점 척도[10]를 사용하였으며, 알고리즘의 결과로 나온 긍정 / 부정의 극성값들은 리커트 5점 척도에 적용되어 매우만족(5) > 만족(4) > 보통(3) > 불만족(2) > 매우불만족(1)의 결과를 도출하게 된다. 알고리즘에 대한 자세한 설명은 다음과 같다. 먼저, 알고리즘의 입력데이터는 형태소단위의 단어리스트(W))로 구성된 하나의 리뷰데이터이며, 단어리스트내의 모든 단어들은 3.2에서 구축한 감성사전과 매칭 되어 감성 분석이 이루어진다. 또한 알고리즘 내의 score함수에 의해 긍정어휘는 +1의 값을, 부정어휘는 -1의 값을 갖게 되며(line:6) 극성 분석의 세밀화를 위해 부사에 가중치(weight)를 두어 각 감성의 정도가 깊어지게 하였다(line:4). 예를 들어, “정말”, “너무”와 같은 강한 강도의 강조 부사어에는 2의 weight값을, “조금”, “약간”과 같은 약한 강도의 부사어에는 0.5의 weight값을 주고. “안”과 같은 부정부사어에는 -1의 weight값을 주었다. 한편 “않다”와 같은 부정동사가 나올 경우에는 기존의 극성 값이 반대 극성으로 바뀌므로 -1의 값을 곱해 스택에 저장되게 하였다(line:7-8). 이렇게 계산된 알고리즘내의 모든 감성 값들은 스택형태의 자료구조에 의해 관리되며, 모든 단어의 감성분석이 완료되면 스택 내에 저장되어있는 양수 값들은 긍정 극성의 합을 저장하는  $SL_{positives}$ 에(line:12), 음수 값 들은 부정극성 합에 절대 값을 주어

$SL_{negatives}$ 에(line:13) 저장되고, 결과적으로 이들은 리커트 5점 척도(line:14)에 적용되어 만족도를 도출하게 된다.

```

Algorithm 2. Sentiment Analysis
Input: Wordlists of Review W[]
Output: LikertScale of Review
1: set Score = 1;
2: for each word in W do
3:   if isAdverb(word) then
4:     weight ← DistinctAdverbWeight(word);
5:   if isSentimentWord(word) then
6:     Stack ← Score(word) * weight;
7:   if isNegativeVerb(word) then
8:     pop element and push element multiplied by -1;
9:   end if
10: end for
11:
12:  $SL_{positives}$  ← CalculatePositiveScoreLine(stack);
13:  $SL_{negatives}$  ← CalculateNegativeScoreLine(stack);
14: LikertScale ← (  $SL_{positives}$  * 5 ) / (  $SL_{positives}$  +  $SL_{negatives}$  );
    
```

(그림3) 만족도 도출을 위한 감성분석 알고리즘

(Figure 3) Algorithm 1: Sentiment Analysis For Calculate Degree of Satisfaction

## 4. 시스템 구현 및 결과

본 논문에서 제안한 시스템은 hadoop-2.6.0 버전이 설치된 1대의 Master서버와 2대의 Slave 서버를 기반으로 구현되었다. 본 연구의 실험에는 맛집 콘텐츠를 실험 대상으로 선정하고, TV프로그램에서 방영된 맛집 1,163개의 정보를 기반으로 약 6GB의 리뷰데이터를 수집하여 실험에 활용하였다. 또한 리뷰데이터의 형태소 분석은 국민대학교 연구실에서 제공하는 KLT 형태소 분석기 [11]를 사용하였다.

### 4.1 감성 어휘 사전 구축

맛집 리뷰 오피니언마이닝을 위한 감성사전 구축을 위해서는 먼저 시드 단어를 정의해야 한다. 시드 단어의 정의는 음식점과 관련해 감성이 분명한 10개의 단어를 골라 시드단어로 정의하였다. 그림 4은 처음에 정의된 시드집합이다.

긍정 SEED (+1)	부정 SEED (-1)
맛있다/V	맛없다/V
좋다/V	싫다/V
저렴/N	비싸다/V
친절/N	남기다/V

(그림 4) 초기 시드 단어 집합

(Figure 4) Dictionary of Pre-defined Seed Words

다음으로는 형태소 분석과정을 거친 리뷰데이터들 중 감성단어의 대상이 되는 용언(V), 명사(N), 부사(a)와 품사를 알 수 없으나 신조어나 은어가 될 가능성이 있는 단어(K)들을 추출하고, 감성어휘가 아닌 단어들 예를 들면 용언으로 태깅되어 추출되지만 감성어휘가 아닌 “이다”, “가다” 등과 같은 불 용어들은 따로 전처리를 하였다. 그리고 [9]에서 제안한 방법에 적용하여 단어 그래프를 구성하고, 그래프를 통해 시드단어를 중심으로 감성을 추론하여 이를 토대로 단어사전을 구축 하였다. 그림 5은 갱신된 시드집합과 간선의 수를 나타낸다. 시드집합의 갱신은 구축된 단어 그래프와 사전을 이용하여 3.2에서 제안한 방법을 토대로 시드집합을 갱신하였다. 초기 시드집합인 그림 4과 비교해 보았을 때, 그림 5에서 “깔끔하다”, “다양”, “추천”, “아쉽다”, “짜다”, “느끼하다” 등의 어휘들이 시드집합에 추가된 것을 볼 수 있다. 추가된 어휘들은 단어 그래프 생성 시에 초기 시드집합의 각 극성에서 가장 낮은 간선의 수를 가지고 있는 긍정시드어휘 “친절” 17047개, 부정시드어휘 “남기다”의 10248개의 간선의 수보다 높은 간선의 수를 가지고 있으므로 도메인 내에서 많은 단어들과 밀접 성을 가지는 것으로 판단되어 시드집합에 포함되었다.

긍정 SEED (+1)	Degree	부정 SEED (-1)	Degree
좋다/V	317502	아쉽다/V	25492
맛있다/V	305350	짜다/V	25255
깔끔하다/V	41513	맛없다/V	21244
다양/N	38610	비싸다/V	20912
추천/N	34913	느끼하다/V	14312
부드럽다/V	31896	못먹다/V	13868
저렴/N	22290	싫다/V	13448
친절/N	17047	남기다/V	10248

(그림 5) 갱신된 시드 단어 집합

(Figure 5) Dictionary of Updated Seed Words

#### 4.1 만족도 분석 결과

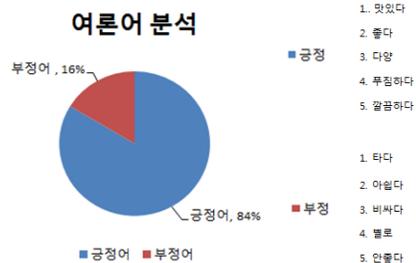
##### “팔색삼겹살” 만족도



(그림 6) 만족도 분석 결과

(Figure 6) Result of Satisfaction Analysis

그림 6은 맛집 ‘팔색삼겹살’의 1027개 리뷰 분석 결과 화면으로, 각 리뷰들은 다음과 같은 과정을 거쳐 만족도가 도출된다. 먼저 ‘팔색삼겹살’을 키워드로 하여 수집된 1027개의 리뷰 데이터들은 형태소 분석과 전 처리 과정을 거쳐 HDFS에 저장되며, 저장된 데이터들은 3.3절의 하둡 맵리듀스를 기반으로 하는 감성 분석 알고리즘을 통해 ‘팔색삼겹살’을 key로 ‘팔색삼겹살’에 대한 리뷰데이터들을 value로 하여 분산적으로 분석을 수행하게 된다. 그림 6 ‘팔색삼겹살’의 경우, 1027개의 리뷰데이터들로부터 도출된 만족도는 총 76.1%이며, 먼저 왼쪽의 원형 도표를 보면, ‘팔색삼겹살’의 이용자들 중 만족의 경향을 나타내는 이용자들은 53%, 보통의 경향을 나타내는 이용자들은 44%, 불만족의 경향을 나타내는 이용자들은 4%로 해당 맛집에 대해 불만족의 경향을 나타내는 이용자들 보다 만족의 경향을 나타내는 이용자들이 14배 정도 많다는 것을 알 수 있다. 또한 이에 대한 자세한 분석 내용을 보기 위해 왼쪽 도표를 살펴보면 1027개의 리뷰들 중 매우 만족은 355개, 만족은 199개, 보통은 469개, 불만족은 25개, 매우 불만족은 9개로 총 1027명의 이용자들 중 과반수이상인 554명의 이용자들이 해당 음식점에 대해 대체적으로 만족했다는 것을 알 수 있다. 한편, 그림 7은 그림 6에서 나타난 매우만족(5) > 만족(4) > 보통(3) > 불만족(2) > 매우 불만족(1)의 리뷰들에 나타난 긍정, 부정어 여론을 보여주는 결과로 왜 이용자들이 그림 6와 같은 만족도를 나타냈는지에 대한 근거를 좀 더 자세히 볼 수 있다.



(그림 7) 여론어 분석 결과

(Figure 7) Result of Public Opinion Analysis

그림 7은 맛집 ‘팔색삼겹살’의 리뷰에서 도출된 만족도와 관련해 사용자들이 SNS에 실제로 표현한 감성어휘들의 여론을 보여준다. 왼쪽의 도표는 감성분석이 수행될 때 매칭된 감성어휘들을 대상으로 긍정과 부정어휘의 비율을 계산한 결과를 나타내며, ‘팔색 삼겹살’의 경우

총 3723개의 감성어휘들이 출현했고, 그 중 긍정어휘는 중복포함 3120개, 부정어휘는 중복포함 603개로 여론어의 비율은 긍정 84%, 부정 16%가 되어 왼쪽 도표와 같이 나타난다. 한편, 오른쪽의 표는 긍정, 부정어휘들을 각각 워드카운트 하여 각 감성에서 가장 많이 언급된 상위 5개의 감성어들을 보여준다. 먼저 오른쪽 표를 보았을 때 ‘팔색삼겹살’에 대한 가장 큰 만족요인은 기본적인 긍정어 ‘맛있다’, ‘좋다’, 외에 ‘다양’, ‘푸짐하다’, ‘깔끔하다’가 되며, 이용자들은 이러한 긍정 여론어들을 통해 ‘음식이 맛있고’, ‘메뉴가 다양하며’, ‘양이 푸짐하다’ 등이 ‘팔색삼겹살’ 가장 큰 장점이라는 것을 파악 할 수 있다. 한편, 불만족의 요인은 ‘타다’, ‘아쉽다’, ‘비싸다’ 등으로 이용자들은 ‘고기가 탸다’ ‘가격이 비싸다’는 이유가 해당 음식점의 가장 큰 단점이라고 파악 할 수 있다. 결과적으로 그림 6~7과 같은 만족도 분석 결과를 통해 시스템의 이용자들은 해당 콘텐츠의 만족도 분석 결과와 함께 리뷰 속에 내포되어있는 감성 여론의 결과를 좀 더 자세하게 볼 수 있다.

#### 4. 결 론

본 논문에서는 한정된 특정 콘텐츠가 아닌, 다양한 콘텐츠 리뷰에 적용되어 오피니언 마이닝이 가능한 콘텐츠 만족도 분석 시스템을 설계 및 구현하였다. 또한 본 시스템의 기준이 되는 콘텐츠 도메인에 따른 감성사전 구축 방법과 만족도 도출을 위한 감성 분석 알고리즘의 단계를 구체적으로 제시하였다. 그리고 본 연구에서 설계한 시스템을 실험하기 위해 맛집 콘텐츠를 선정하여 약 6GB의 리뷰데이터들을 기반으로 맛집 도메인에 대한 감성사전을 구축하고, 제안한 알고리즘에 적용해 특정 맛집의 감성 분석 결과와 만족도를 도출해 냈다.

본 연구의 콘텐츠 만족도 분석 시스템은 리뷰에 내포되어 있는 작성자의 의견 정도를 추출하여 자세한 만족도를 분석 할 수 있다는 점, 다양한 콘텐츠들을 대상으로 분석이 가능하다는 점, 대용량의 리뷰데이터들을 빠르고 효율적으로 처리할 수 있다는 점에서 매우 실용적인 분석시스템으로써의 가치가 있다. 본 시스템을 통해 콘텐츠 이용자들은 이용하고자 하는 콘텐츠에 대한 자세한 의견과 전체적인 만족도를 미리 파악할 수 있어 보다 효율적인 의사결정을 할 수 있을 것이며, 콘텐츠 제공자들은 기존 설문조사 방식과 비교해 제공한 콘텐츠에 대한 이용자들의 방대한 의견정보를 좀 더 빠른 시간 내에 명확하고 자세하게 분석 할 수 있어 현재 개선이 필요한 문

제의 발견과 지속적인 품질유지 등의 차후 마케팅 전략을 세우는 데 많은 도움이 될 것으로 기대한다.

향후 연구로는 감성 사전 구축에 있어 중립의 의미를 갖는 어휘를 처리하는 방법을 연구하여 감성 서술어의 의미를 확장 할 필요가 있으며, 긍정과 부정단어가 고루 쓰이지 않은 말뭉치에 대한 명확한 기준을 찾는 것이 필요하다. 또한 대용량의 리뷰데이터로부터 작성자의 만족도를 도출하는 것뿐만 아니라, 이를 통해 콘텐츠에 대한 욕구, 기대감등을 추론할 수 있는 기능까지 추가된다면 좀 더 지능적인 만족도 분석 시스템이 될 수 있을 것이다.

#### 참 고 문 헌 (Reference)

- [1] Hadoop, [Online] Available <http://hadoop.apache.org/>
- [2] S.M Kim, E Hovy "Extracting opinions, opinion holders, and topics expressed in online news media text", Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text, Australia, 2006 <http://www.isi.edu/natural-language/people/hovy/papers/06ACL-WS-opin-topic-holder.pdf>
- [3] So-young Lee, Kwanho In, Ung-mo Kim "Analysing Product Reviews of Cosmetics Using Opinion Mining" Proc. of the Korea Computer Congress 2013, pp. 362- 264 <http://www.dbpia.co.kr/Article/NODE02217118>
- [4] Yoo-sin Kim, Seung-Ryul Jeong "Intelligent VOC Analyzing System Using Opinion Mining" Journal of Intelligent Informaion Systems 19(3), 2013.9, pp.113-125 [http://ocean.kisti.re.kr/download/volume/kiiss/JJSHBB/2013/v19n3/JJSHBB\\_2013\\_v19n3\\_113.pdf](http://ocean.kisti.re.kr/download/volume/kiiss/JJSHBB/2013/v19n3/JJSHBB_2013_v19n3_113.pdf)
- [5] Seung Youp Lee, Kwan Ho In, Ung Mo Kim "Analyzing University Bulletin Board Data by using Opinion Mining" Journal of KOREA INFORMATION SCIENCE SOCIETY 39(2C), 2012.11, pp. 104-106 [www.dbpia.co.kr/Article/NODE02048307](http://www.dbpia.co.kr/Article/NODE02048307)
- [6] Jae-Young Chang "A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in On-Line Shopping Mall" Journal of Society for e-Business Studies 14(4), 2009.11, pp. 19-33 [http://society.kisti.re.kr/sv/SV\\_svpsbs03VR.do?method=detail&cn2=KJGRBH\\_2009\\_v14n4\\_19&menuid=1&subid=11](http://society.kisti.re.kr/sv/SV_svpsbs03VR.do?method=detail&cn2=KJGRBH_2009_v14n4_19&menuid=1&subid=11)

- [7] Jong-Hyuk Lee, Won-Sang Lee, Jae-Won Park, Jae-Hyun Choi "The Blog Polarity Classification Techbique using Opinion Mining" Journal of Digital Contents Society 15(4), 2014.8, pp.559-568  
<http://dx.doi.org/10.9728/dcs.2014.15.4.559>
- [8] Yean-Ju Oh, Soo-Hoan Chae "Movie Rating Inference by Construction of Movie Sentiment Sentece using Movie comments and ratings" Journal of Korean Society for Internet Information 16(2), 2015.4, pp. 41-48  
[www.dbpia.co.kr/Article/NODE06291845](http://www.dbpia.co.kr/Article/NODE06291845)
- [9] Jung-Ho Kim, Yean-Ju Oh, Soo-Hoan Chae "The Construction of a Domain-Specific Sentiment Dictionary Using Graph-based Semi-supervised Learning Method" Korean Journal of the science of emotion & sensibility Vol.18 No.1, 2015 pp.97-104  
[http://210.101.116.28/W\\_files/kiss61/1g600854\\_pv.pdf](http://210.101.116.28/W_files/kiss61/1g600854_pv.pdf)
- [10] LikertScale, [Online], Available  
[http://ko.wikipedia.org/wiki/%EB%A6%AC%EC%BB%A4%ED%8A%B8\\_%EC%B2%99%EB%8F%84](http://ko.wikipedia.org/wiki/%EB%A6%AC%EC%BB%A4%ED%8A%B8_%EC%B2%99%EB%8F%84)
- [11] Korean Lexical Analyzer, [Online], Available  
[nlp.kookmin.ac.kr](http://nlp.kookmin.ac.kr)

## ● 저 자 소 개 ●



### 김 문 지 (MoonJi Kim)

2013 ~ 현재 숙명여자대학교 컴퓨터학과 학사과정  
관심분야 : 빅데이터 분석, 클라우드 컴퓨팅 환경  
E-mail : km0420jj@gmail.com



### 송 은 정 (EunJeong Song)

2013 ~ 현재 숙명여자대학교 컴퓨터학과 학사과정  
관심분야 : 빅데이터 분석, 클라우드 컴퓨팅 환경  
E-mail : srsr123@sookmyung.ac.kr



### 김 윤 희 (YoonHee Kim)

1991년 숙명여자대학교 전산학과(학사)  
1996년 Syracuse University 전산학과(석사)  
2000년 Syracuse University 전산학과(박사)  
1991년~1994년 한국전자통신 연구원 연구원  
2000년~2001년 Rochester Institute of Technology 컴퓨터공학과 조교수  
2001년~2004년 숙명여자대학교 컴퓨터학과 조교수.  
2004년~2009년 숙명여자대학교 컴퓨터학과 부교수  
2009년~현재 숙명여자대학교 컴퓨터학부 교수  
관심분야 : 그리드 컴퓨팅 환경(PSE), 워크플로우 제어, 그리드/클라우드 관리  
E-mail : yulan00@gmail.com