데이터마이닝에 기반한 온라인 고객리뷰 분석 방법론

A Study on Methodology for Analyzing Online Customer Reviews by Using Data Mining

김 근 형* · 김 두 경**
(Kim, Keun-Hyng · Kim, Doo-Kyung)

목 차

Ⅰ. 서 론

웹2.0의 등장으로 인터넷에서의 네티즌 역할은 단순한 정보의 사용자에서 생산자(제공자)로 확대되었다. 현재 인터넷상에는 네티즌들이 생산한 수많은 온라인 콘텐츠(Online Contents)들이 존재한다. 온라인 콘텐츠에는 다양한 유형들이 있으나 그 중에서도 온라인 커뮤니티(community)를 통하여 네티즌들의 다양한 의견이나 경험, 지식 등을 표현한 비정형화된 텍스트데이터(Unformatted Text Data) 형태의 온라인 고객리뷰들(Online Customer Reviews)이 방대하게 존재하고 있으며 더욱 증가되고 있는 추세이다. 웹 2.0의 국내 주요 사례로 싸이월드, 네이버 지식인, 다음카페 등이 있지만 IT 선진국인 미국

* 제주대학교 경상대학 경영정보학과 부교수
** 제주대학교 경상대학 경영정보학과 교수

에서는 더욱 활성화 되어 있다. 미국의 최대 온라인 관광 사이트인 TripAdvisor (www.tripadvisor.com)에는 관광객들이 직접 기록해 놓은 수십만 건의 관광정보와 경험·지식들이 비정형 데이터 형태로 게시되어 있고 많은 관광객들이 유용한 정보로 활용하며 확대 재생산하고 있다[김근형, 2009]. 국내에서도 웹2.0 마인드를 표방하는 전문 사이트들이 개설되어 운영되고 있으며 네티즌들의 참여가 점차 확대되고 있는 추세이기 때문에 네티즌들의 의견, 경험, 지식 등을 포함하는 온라인 고객리뷰는 더욱 증가될 것이다.

대부분의 사람들은 제품이나 서비스를 구매할 때 그 제품이나 서비스에 대하여 더 많은 정보를 얻기를 원한다. 의사결정 과정에서 다른 사람의 의견이 영향을 미치고 다른 사람이 추천하는 상품을 구매하려고 한다. 특정 제품이나 서비스에 대한 네티즌의 의견들은 마케팅이나 CRM(Customer Relationship Management)관점에서 볼 때 고객들뿐만 아니라 기업에게도 매우 중요한 자료이다.

오피넌마이닝(opinion mining)은 인터넷 상에 게시된 네티즌들의 여론을 수렴하는 데이터 분석기술이다. 웹상에 게시된 네티즌들의 제품 품평(品評)들이 계속적으로 증가되고 있고 이러한 고객리뷰들을 기계적으로 분석할 수 있는 기술의 중요성 때문에 오피넌마이닝은 데이터마이닝 분야에서 그 의미가 더욱 커지고 있는 새로운 연구분야이다. 오피넌마이닝 기술은 기존의 자연어처리(natural language processing)와 텍스트마이닝(text mining) 및 데이터마이닝(data mining) 등의 기술들이 융합된 것으로써, 비정형 데이터형태의 고객의견들을 분석하여 긍정적 의견(positive opinion)과 부정적 의견(negative opinion) 으로 단순화시켜 요약할 수 있다. 오피넌마이닝의 궁극적인 목적은 대량의 제품 품평들로부터 고객의견들을 추출하고 이를 요약 분석하여 다른 고객들이나 기업에게 유용한 정보 형태로 제공하는 것이다. <그림1>은 네티즌의 제품 품평에 대한 온라인 고객리뷰 예를 나타내고 있다.

고객리뷰1

저는 일반적으로 A사 카메라를 선호합니다. 왜냐하면, A사 카메라는 화질이 좋고 매우 견고하며 우수한 기능들이 있기 때문입니다. 나는 A사 카메라의 품질이 매우 좋다고 생각합니다.

고객리뷰2

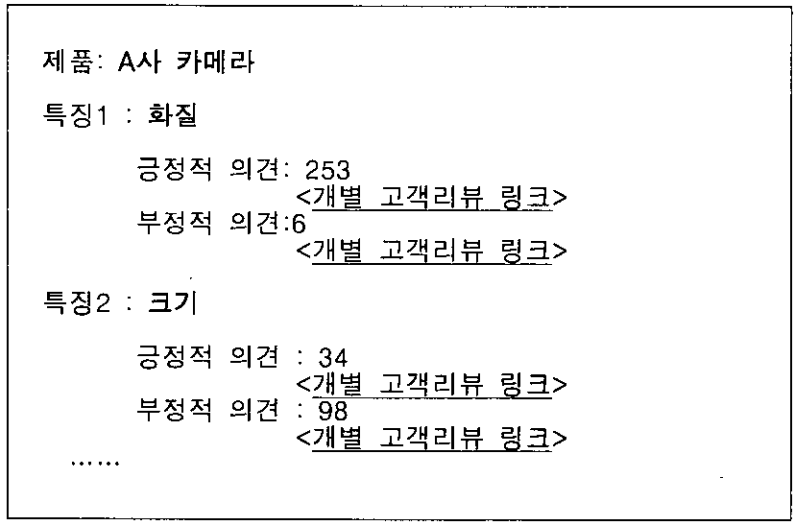
A사 카메라의 화질은 좋지만 크기가 너무 커서 불편한 측면이 있다.

<그림1> 제품 품평 예

<그림1>은 단 2건의 온라인 고객리뷰를 나타내고 있지만, 몇 천 건 또는 몇 만 건에 이르는 대량의 온라인 고객리뷰들을 수작업으로 분석하는 것은 매우 어려운 일이다. 대량의 온라인 고객리뷰들을 컴퓨터에 의하여 자동적으로 신속하게 분석해야 할 필요성 때문에 오피넌마이닝 기술이 각광받게 된 것이다. <그림1>에서 '카메라'는 제품이름을 나타내고 있고 '화질', '기능', '품질' 등은 제품의 특징들(features)을 의미한다. 이들 제품과 제품특징들에 대하여 '화질이 좋고', '견고하며', '우수한 기능', '불편한', 'A사의 카메라가 매우 좋다' 등과 같은 구나 절들은 고객들의 주관적 의견 즉, 긍정적 의견 또는 부정적 의견 등을 의미한다. 특정 제품이나 서비스의 특징들에 대한 대량의 고객 의견들을 분석하여 요약할 수 있다면, 이는 고객 측 뿐 만 아니라 제품을 생산하는 기업 측에도 중요한 자료가 될 수 있다.

<그림2>는 오피넌마이닝 기술에 의하여 대량의 온라인 고객리뷰를 자동 분석하여 요약한 결과의 한 예이다[Minqing Hu, 2004; Xiaowen Ding, 2008]. <그림2>에서는 'A사 카메라'의 '화질'에 관심을 갖는 259명의 네티즌 의견들 중에서 253명은 긍정적인 평가를 내렸고 단지 6명만이 부정적 평가를 내렸음을 나타내고 있다. 긍정적 의견을 나타내는 문장들을 자세히 보려면 '<개별 고객리뷰 링크>'를 클릭하여 더 구체적으로 조회해 볼 수 있다.

<그림2>와 같은 분석결과는 제품의 각 특징들에 대한 고객의견 분포를 전반적으로 파악할 수는 있지만, 고객들이 보다 중요하게 생각하는 특징들은 무엇이고 그 특징들 사이의 상대적 중요성은 어떤지 등에 대한 정보는 보여줄 수 없다. W.Y.Kim(2009)은 오피넌마이닝 과정에서 데이터마이닝의 연관규칙탐사기법을 적용하여 보다 중요한 특징들과 그 특징들 사이의 상대적 중요성을 파악할 수 있는 오피넌마이닝 기법을 제안하고 있다. 그러나 연관규칙탐사기법을 적용할 때 고객의 주관적 의견을 나타내는 형용사를 그룹핑 하지 않기 때문에 최소지지도(minimum support)와 최소신뢰도(minimum confidence)를 만족시키는 규칙들이 최소화되어 유의미한 정보들이 소실된다.



<그림2> 온라인 고객리뷰 분석 예

본 논문에서는 품사태깅(Part-Of-Speech Tagging) 기법을 이용하여 제품특징들과 고객의견들이 어떻게 추출되는지 고찰한다. 또한, 오피년마이닝 과정에서 연관규칙탐사(association rule mining) 기법을 적용하여 제품특징과 고객의견 사이의 연관성을 분석하는 방법론을 고찰하고자 한다. 특히, 오피년마이닝 과정에서 연관규칙탐사기법을 적용할 때 형용사들의 동의어 관계를 이용하여 그룹핑함으로써 보다 많은 유의미한 규칙들을 추출할 수 있는 새로운 오피년마이닝 방법을 제안하고자 한다.

<표 1> 품사 리스트의 예

품 사	축약기호	예
일반 명사	NN	카메라, 품질
명사구	NP	A사 카메라, 우수한 화질
형용사	JJ	우수한, 견고한
동사	VB	선호하다, 생각하다
접속사	CC	그리고, 왜냐하면
부사	AB	일반적으로

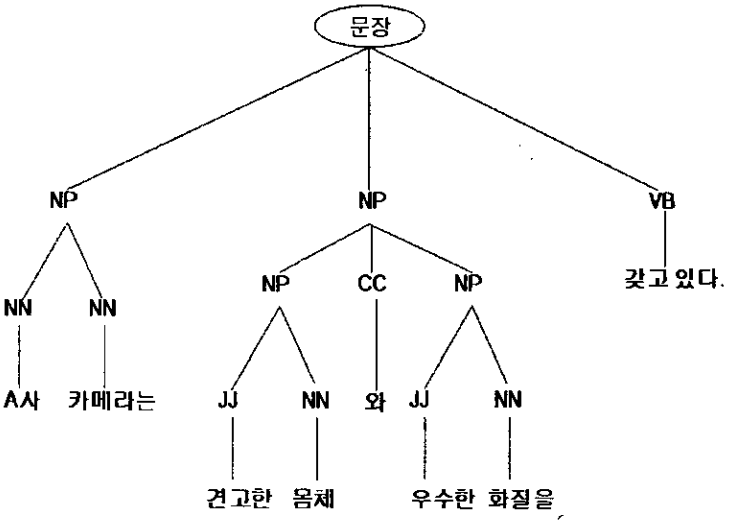
Ⅱ. 이론적 배경

2.1. 품사태깅

품사태깅은 문장(sentence)에 있는 각 단어들에 대하여 명사, 동사, 형용사, 부사, 대명사 등과 같은 각 품사들을 대응시키는 과정이다. 태깅 알고리즘으로의 입력은 자연어(natural language) 문장을 구성하는 단어 열과 <표 1>과 같은 품사 리스트이다. 태깅 알고리즘의 출력은 각 단어가 속할 수 있는 최적의 품사들이다. 스탠포드 태거(Stamford Tagger, 2008)는 영어문장의 품사 태깅이 가능한 공개 소프트웨어이다.

구문분석기(parser)는 품사태깅 후에 문장의 구문을 분석하여 적절한 자료구조로 표현한다. 스탠포드 파서(Stamford Parser, 2008)는 영어문장의 구문을 분석할 수 있는 공개 소프트웨어이며, KLT(http://nlp.kookmin.ac.kr/HAM/kor/download.html)는 한글문장에 대한 품사태깅과 구문분석이 가능한 연구용 공개 소프트웨어이다.

<그림3>은 구문분석기를 통하여 고객리뷰 중의 한 문장 『A사 카메라는 견고한 몸체와 우수한 화질을 갖고 있다』를 분석하여 구문구조트리(sentence structure tree)로 표현한 예를 나타내고 있다.



<그림3> 구문구조트리

2.2. 연관규칙탐사 개요

데이터마이닝 분야에서 연관규칙탐사는 많은 연구가 이루어진 분야로써 대량의 데이터(관계형 화일)로부터 속성(변수)들 사이의 연관성을 규칙형태로 추출하는 데이터분석 기술이다[Agrawal, 1993]. 연관규칙탐사는 미리 정의된 최소지지도와 최소신뢰도를 만족하는 연관규칙(association rule)을 관계형(테이블구조) 화일로부터 추출한다.

P는 매장에 있는 전체 품목 리스트라 하고 Ti는 특정 고객 i에게 판매한 거래품목 리스트라고 하자. 즉, P는 상이한 속성들인 P1, P2, ..., Pn으로 이루어진 속성(즉, 품목이 됨)들의 집합이고, Ti는 고객 i와 거래한 거래내역(P의 부분집합으로 이루어짐)이 되며 Ti ⊂ P 가 된다. T1, T2, ..., Tn이 모여서 관계형 파일 T를 구성하게 되며 각 Ti(i = 0,..., n)는 T의 각 레코드가 된다. 이때, X ⊂ P, Y ⊂ P, X∩Y=∅ 일 때 연관규칙은 「X→Y」 형태로 표현되며 관계형 파일 T로부터 추출된다.

연관규칙의 유의미성 검증을 위한 2가지 중요한 척도는 지지도(degree of support)와 신뢰도(degree of confidence)이다. 연관규칙 「X→Y」 의 지지도란 T의 전체 레코드 수에 대하여 X∪Y를 포함하는 레코드 수의 비율을 나타낸다. 즉, 지지도의 의미는 추출된 연관규칙 「X→Y」 가 얼마나 많은 고객들에게 적용되는 규칙인지를 나타내는 척도로써, 지지도가 높은 연관규칙일 수록 보다 많은 고객들이 관심을 가지는 중요한 품목들을 포함하게 된다. 반면, 연관규칙 「X→Y」 의 신뢰도는 T에서 X를 포함하는 레코드 수에 대하여 X∪Y를 포함하는 레코드 수의 비율을 나타낸다. 즉, 신뢰도의 의미는 추출된 연관규칙 「X→Y」 에서 X에 포함되는 품목들과 Y에 포함되는 품목들이 얼마나 강한 연관성을 갖는지를 나타내는 것이다.

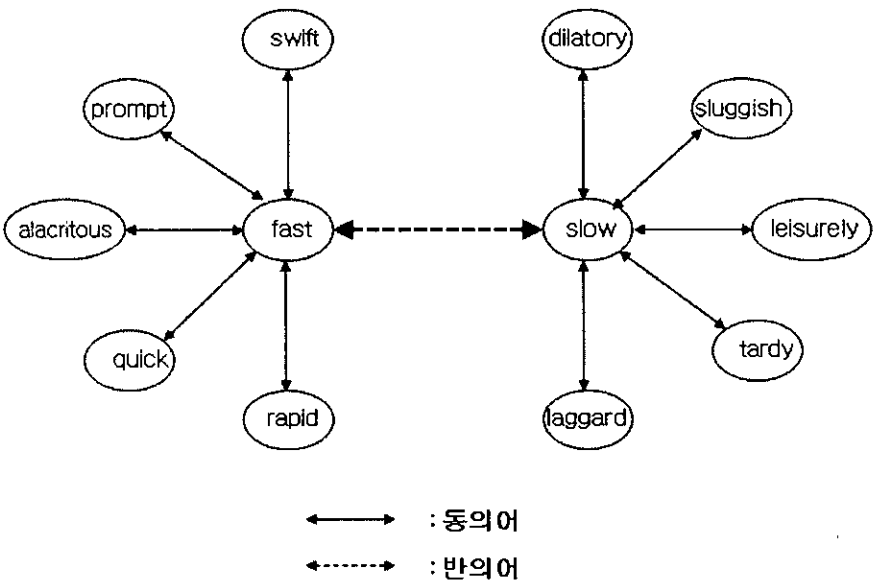
추출된 연관규칙은 미리 설정된 최소지지도와 최소신뢰도를 만족해야 데이터마이닝 분석자에게 유의미한 규칙이 될 수 있다.

2.3. 형용사의 그룹핑

워드넷(WordNet, http://wordnet.princeton.edu/wordnet/)은 프린스턴대학의 심리학 교수인 조지 A.밀러가 주도하는 인지과학연구소에서 개발한 영어의 의미 어휘목록 데이터베이스이다. 워드넷은 영어단어를 'synset'이라는 유의어 집단으로 분류하여 간략하고 일반적

인 정의를 제공하고, 이러한 어휘목록 사이의 다양한 의미관계를 기록한다. 그 목적은 2가지이다. 하나는 사전(단어집)과 시소러스(유의어·반의어 사전)의 배합을 만들어 보다 직관적으로 사용할 수 있고 자동화된 본문분석과 인공지능 응용을 뒷받침하려는 것이다. 아쉽게도 워드넷과 같이 방대하면서도 체계적으로 구축된 한국어 시소러스는 없지만 특정 도메인(특정 제품이나 서비스 관련 영역 등)에 특화된 시소러스의 개발은 어렵지 않을 것으로 판단된다.

워드넷에서 형용사들은 양극성 집단(bipolar clusters)으로 조직된다[Minqing Hu, 2004]. <그림 4>는 양극성으로 형용사가 그룹핑된 예를 나타내고 있다.



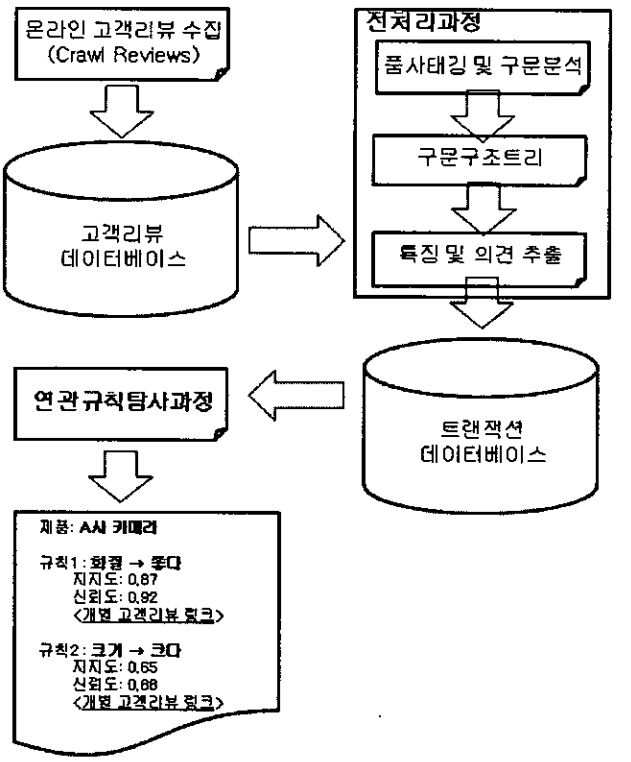
<그림4> 형용사들의 양극성 그룹핑 예

일반적으로 어떤 형용사의 동의어들은 비슷한 의미를 갖고 반의어들은 반대의 의미를 갖는다. Minquing[Minqing Hu, 2004]은 온라인 고객리뷰 상에 나타나는 형용사들을 그룹핑할 때 긍정적 집단과 부정적 집단의 두 부류로만 분류하기 때문에 종자(seed)가 될 초기 형용사(seed lists)들을 사람이 직접 설정해주어야 하는 번거로움이 발생한다. 뿐만 아니라 고객의 주관적 의견을 양극화로 너무 한정하는 것은 분석결과의 부정확성을 초래할 가능성이 크다.

```
1.  Assign = FALSE;
2.  Procedure Grouping_Adjectives(in_adj)
3.  begin
4.      for each group Gi in group_list
5.        if (in_adj is synonym with adjective s in group Gi ){
6.              append in_adj to Gi ;
7.                      out_adj = Gi(0);
8.                      Assign  = SUCCESS;
9.                      break;
10.     }
11.     if (Assign = FALSE){
12.                     generate new group Gnew;
13.                     append Gnew to group_list;
14.                     out_adj = Gnew(0); }
15. return out_adj;
16. end
```

<그림5> 형용사의 그룹핑 알고리즘

본 연구에서는 온라인 고객리뷰 상에 나타나는 형용사들을 시소러스의 동의어 관계를 바탕으로 n개의 그룹으로 분류하고자 한다. <그림 5>는 고객리뷰 상의 형용사들을 그룹핑하기 위하여 본 논문에서 새롭게 제안하는 알고리즘을 나타내고 있다. 특정 도메인과 관련된 형용사 시소러스는 이미 구축되어 있고 알고리즘이 실행될 때 참조되는 것으로 가정한다. 온라인 고객리뷰상의 각 문장에 대한 구문구조트리를 탐색하면서 새로운 형용사 'in_adj'를 만날 경우, 기존에 분류된 형용사 그룹들인 group_list의 각 group에 속한 형용사들과 동의어 관계인지를 체크한다(5행). 기존 형용사 그룹들에서 동의어 관계가 발견되면 해당 그룹에 'in_adj'를 추가하고 해당 그룹의 대표 형용사 'out_adj'를 리턴한다(6행, 7행). 기존 형용사 그룹들에서 동의어 관계가 발견되지 않으면 새로운 형용사 그룹을 만들고 새롭게 생성된 그룹을 group_list에 추가한다(13행).



<그림6> 오피년마이닝 과정 개요

Ⅲ. 오피년마이닝 방법론

본 논문에서 오피년마이닝 과정은 2단계로 이루어진다. 단계1에서는 비구조화된 데이터의 전처리과정을 수행하여 구조화된(structured) 형태의 트랜잭션 데이터베이스를 생성하는 과정이다. 단계2 에서는 트랜잭션 데이터베이스의 관계형 화일 데이터를 대상으로 연관규칙탐사기법을 적용하여 제품특징과 고객의견 사이의 연관규칙을 추출하는 과정이다. <그림6>은 오피년마이닝 방법을 개략적으로 나타내고 있다.

3.1. 전처리과정

전처리과정(preprocessing)에서는 고객리뷰 데이터베이스의 각 고객리뷰들에 대하여 구

문구조트리가 생성되며, 각 구문구조트리로부터 명사에 대응되는 '특징'과 형용사에 대응되는 '의견'을 추출하여 트랜잭션 화일을 생성한다. 트랜잭션 화일 T의 스키마(schema)는 T(특징, 의견)으로 구성된다. <그림 7>은 구문구조트리를 입력으로 하여 구조화된 트랜잭션 화일을 생성하는 알고리즘을 나타내고 있다. <그림7>에서 고객리뷰로부터 발견되는 형용사는 앞에서 제안했던 Grouping_Adjectives() 프로시저에 의하여 그룹핑되면서 트랜잭션 화일에 추가되고 있음을 알 수 있다(6행).

```
1.  Procedure Generating_Transaction(구문구조트리i)
2.  begin
3.      구문구조트리를 depth-first-search 방식으로 탐색한다;
4.      if (JJ 노드를 발견하면)
5.              if (JJ노드의 오른쪽 형제노드가 NN이면){
6.                  의견 = Grouping_Adjectives(JJ와 대응하는 형용사) ;
7.                  특징 = NN과 대응하는 명사 ;
8.                  트랜잭션 Ttemp(특징, 의견) 생성;
9.                  트랜잭션 Ttemp(특징, 의견) 을 트랜잭션화일 T에 추가; }
10. end
```

<그림7> 트랜잭션 생성 알고리즘

3.2. 연관규칙탐사 및 분석결과

스키마가 T(특징, 의견)으로 구성된 관계형화일인 트랜잭션 화일 T에 대하여 연관규칙탐사 알고리즘을 적용하면서 최소지지도와 최소신뢰도를 만족하는 규칙들을 생성한다. 규칙의 지지도가 높을수록 보다 많은 고객들이 관심을 갖고 있다는 의미이며, 신뢰도가 높을수록 그 규칙의 정확성이 높다는 의미이다. <그림8>에서 규칙 [화질 → 좋다]의 지지도가 [크기 → 크다] 의 지지도보다 높다. 이것은 고객들이 카메라의 특징들인 화질과 크기 중에서 화질에 더 높은 관심이 있음을 의미하는 것이다. 규칙 [화질 → 좋다]의 신뢰도가 0.92 라는 의미는 카메라 화질에 대한 고객의견 중에서 92%가 좋다라는 긍정적 반응을 보였음을 나타내는 것이다.

제품: A사 카메라

규칙1 : 화질 → 좋다

지지도: 0.87
신뢰도: 0.92
<개별 고객리뷰 링크>

규칙2 : 크기 → 크다

지지도: 0.65
신뢰도: 0.88
<개별 고객리뷰 링크>

.....

<그림8> 연관규칙탐사에 기반한 고객리뷰 분석 결과 예

Ⅳ. 결 론

경영정보를 생성하기 위한 기존의 정보분석 환경은 대부분 테이블 형태의 정형화된 데이터구조를 가정하였다. 데이터마이닝은 테이블 구조의 정형화된 데이터를 대상으로 한 정보분석 기술 중의 하나이다. 그러나 웹 2.0 시대에는 수많은 네티즌들이 기업경영과 관련한 다양한 의견들을 비정형화된 데이터로 생성하여 게시하기 때문에 비정형화된 데이터를 대상으로 한 정보분석 기술 중의 하나인 텍스트마이닝의 수요가 커질 것으로 판단된다.

본 논문에서는 데이터마이닝과 텍스트마이닝 분야의 새로운 연구영역인 오피년마이닝 방법에 대하여 살펴보았다. 웹 2.0 시대에 온라인 고객리뷰들이 더욱 증가될 것이라는 측면에서 오피년마이닝 기술의 중요성은 더욱 커질 것이다. 오피년마이닝 기술은 기술적 수요가 클 뿐 아니라 기술적 한계에 봉착한 자연어처리기법의 응용영역을 온라인 고객리뷰로 축소시켜 자연어 처리의 정확성을 높일 수 있으며, 학문적인 성과가 큰 기존의 데이터마이닝 기술을 접목할 수 있어 그 성공가능성이 매우 큰 연구영역이다.

본 논문에서는 새로운 오피년마이닝 방법을 제안하였다. 단순히 긍정과 부정의견으로만 분류하였던 기존의 오피년마이닝 방법을 개선하여 연관규칙탐사기법과 형용사 그룹핑 기법을 접목시킴으로써 보다 풍부한 고객의견 정보를 도출할 수 있는 방안을 제시하였다.

본 논문에서 제안한 아이디어는 논리적 관점에서만 고찰되었고 실제 시스템을 구현하여 그 성능을 실험적으로 검증하기 위한 추가적인 연구가 필요하다.

참 고 문 헌

1. Agrawal, R., Imielinski, T., Swami, A., "Mining association rules between sets of items in large databases", Proc. of ACM SIGMOD, 1993, pp.207-216.
2. Korean Parser Test Version, http://nlp.kookmin.ac.kr/HAM/kor/download.html.
3. Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", KDD'04, 2004, pp.168-177.
4. Stanford Tagger Version 1.6.2008. http://nlp.stanford.edu/software/taggerr.shtml
5. Stanford Parser Version 1.6.2008.
http://nlp.stanford.edu/software/lex-parser.shtml
6. Xiaowen Ding, Bing Liu and Philip S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", WSDM'08, 2008, pp.231-239.
7. W.Y.Kim, J.S. Ryu, K.I.Kim, U.M.Kim, "A Method for Opinion Mining of Product Reviews using Association Rules", ICIS, 2009, pp.270-274.
8. 김근형, "온라인 고객리뷰 분석을 통한 시장세분화에 텍스트마이닝 기술을 적용하기 위한 방법론", 한국콘텐츠학회논문지, 2009.