연관성 모델에 기반한 오피넌마이닝 시스템의 설계 및 구현

김근형*

Design and Implementation of Opinion Mining System based on Association Model

Keunhyung Kim*

요 약

특정 제품이나 서비스에 대한 네티즌의 의견들은 고객들의 구매 행위에서의 참고대상일 뿐만 아니라 기업 입장에서도 마케팅이나 경영전략을 수립하기 위한 중요한 자료가 될 수 있기 때문에 온라인 고객리뷰를 분석하는 것은 매우 중요하다.

본 논문에서는 비정형(unformatted) 데이터형인 자연어(natural language) 형태로 웹상에 게시된 고객 의견들을 분석할 수 있는 새로운 오피넌마이닝 기법을 제안한다. 기존 데이터마이닝 기법 중의 하나인 연관규칙탐사 기법을 수정하여 오피넌마이닝 과정에 보다 효율적이고 효과적으로 적용하기 위한 방안을 고찰하고 이를 기반으로 실제 시스템을 설계하고 구현하였다.

ABSTRACT

For both customers and companies, it is very important to analyze online customer reviews, which consist of small documents that include opinions or experiences about products or services, because the customers can get good informations and the companies can establish good marketing strategies.

In this paper, we propose the association model for the opinion mining which can analyze customer opinions posted on web. The association model is to modify the association rules mining model in data mining in order to apply efficiently and effectively the association mining techniques to the opinion mining. We designed and implemented the opinion mining systems based on the modified association model and the grouping idea which would enable it to generate significant rules more.

키워드

:온라인고객리뷰, 연관성 모델, 감성문장, 개체지지도, 감성지지도, 연관도, 오피넌마이닝

Keyword

Online customer review, Association model, sensitivity sentence, support of entity, sensitivity support, association, opinion mining

* 정회원 : 제주대학교 교수 (khkim@jejunu.ac.kr)

접수일자 : 2010. 08. 23
심사완료일자 : 2010. 09. 08

Ⅰ. 서 론

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 인터넷상에 게시된 네티즌들의 여론을 수렴하는 데이터 분석기술이다. 웹상에 게시된 네티즌들의 온라인 고객리뷰들, 즉 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 연구 분야이다. 오피넌마이닝 기술은 기존의 자연어처리와 텍스트마이닝 및 데이터마이닝 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 온라인 고객리뷰들로부터 고객의견들을 추출하고 이들을 분석・요약하여 다른 고객들이나 기업에게 유용한 정보 형태로 제공하는 것이다.

기존의 오피넌마이닝 기법에 의한 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. 또한, 긍정과 부정의견을 분류하는 작업은 추가적인 데이터를 필요로 하며 응용영역 의존적으로(domain- dependent) 만들어져야 하므로 매우 번거로운 작업 중의 하나이다.

본 논문에서는 오피넌마이닝 과정에서 데이터마이닝의 연관규칙탐사 기법을 도입함으로써 긍정과 부정의견을 분류하지 않으면서 제품특징에 대한 고객의견을 추출하는 방법을 제안한다. 온라인 고객리뷰의 구조에 적합하게 기존의 연관규칙탐사 모형을 수정하여 오피넌마이닝에 적용하며, 특히 형용사들을 그룹핑(grouping)함으로써 보다 많은 유의미한 규칙들이 추출될 수 있도록 오피넌마이닝 시스템을 설계하고 구현한다.

Ⅱ. 관련 연구

2.1 데이터마이닝

데이터마이닝 분야에서 연관규칙탐사는 많은 연구가 이루어진 분야로써 대량의 데이터(관계형 화일)로부터 속성(변수)들 사이의 연관성을 규칙형태로 추출하는 데이터분석 기술이다[1]. 연관규칙탐사는 미리 정의된 최소지지도와 최소신뢰도를 만족하는 연관규칙(association rule)을 관계형(테이블구조) 화일로부터 추출한다.

P는 매장에 있는 전체 품목 리스트라 하고 T_i는 특정 고객 i에게 판매한 거래품목 리스트라고 하자. 즉, P는 상이한 속성들인 $P_1, P_2, ..., P_n$으로 이루어진 속성(즉, 품목이 됨)들의 집합이고, T_i는 고객 i와 거래한 거래내역(P의 부분집합으로 이루어짐)이 되며 $T_i \subset P$ 가 된다. T_1, T_2, ..., T_n이 모여서 관계형 파일 T를 구성하게 되며 각 $T_i(i = 0,..., n)$는 T의 각 레코드가 된다. 이때, $X \subset P$, $Y \subset P$, $X \cap Y = \varnothing$ 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 $X \cup Y$를 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 $X \cup Y$를 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.

2.2 문서요약

기존의 문서요약 기술은 2가지 유형으로 나누어진다. 하나는 원형틀(template, 원형판)을 채워 넣은 방식이고, 다른 하나는 핵심문장을 추출하는 방식이다[2, 3]. 원형틀을 채워 넣는 방식은 문서안의 핵심 개체(entity)나 사실(fact)을 식별·추출하여 원형틀의 각슬롯(slot)에 할당한다. 이러한 방법은 원형틀이 먼저 만들어져야하기 때문에 해당 도메인(domain)에 대한 사전지식이 필요하며 따라서, 도메인 의존적 기법이라는 한계가 있다. 핵심문장 추출방식은 문서내용 중에서 가장 대표적인 문장이나 단락을 추출함으로써 문서내용을 간략화 한다. 핵심문장 추출방식은 길이가 긴 단일문서의 간략화를 목적으로 하기 때문에 길이가 짧은 대량의 다중문서로 구성된 온라인 고객리뷰를 분석하기 위한 방법으로는 적합하지 않다.

2.3 오피넌마이닝

오피넌마이닝은 상품 평이나 고객리뷰를 요약한다는 측면에서 기존의 문서요약과 유사한 점이 있지만, 온라인 고객리뷰가 대량의 다중문서로 구성되고 마이닝 대상이 상품특성과 의견이라는 측면에서 기존의 문서요약기법과 차이가 있다.

[4]에서는 기계학습 및 자연어처리기술을 활용하여, 온라인고객리뷰 데이터에 대한 감성분석과 분석결과 요약기법을 제시하고 있으며, Opinion Observer라는 시스템을 개발하였다. 미국 카네기멜론 대학교에서는 Redopal 시스템을 개발한 사례가 있으며[5], 이는 고객리뷰 데이터와 사용자 평가점수를 활용하여 요약보고서를 생성하는 기법을 제안하였다. [6]에서는 문장구조와 문장 사이의 관계, 문장성분의 패턴정보 등의 언어규칙을 이용한 통계학적 방법으로 오피넌마이닝에 접근하고 있다. [7, 8, 9, 10]에서는 워드넷을 활용하여 어휘의 긍정이나 부정적 의미를 판단하고 이를 센티워크넷(SentiwordNet)으로 응용하여 감정의 폭을 정량화하는 방법을 제시하고 있다.

[11]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 온라인고객리뷰 구조를 고려하지 않고 그대로 적용하고 있기 때문에 유의미한 연관규칙의 도출 가능성이 줄어들 소지가 있다. 또한, 고객의 주관적 의견을 나타내는 형용사를 그룹핑 하지 않기 때문에 최소지지도(minimum support)와 최소신뢰도(minimum confidence)를 만족시키는 규칙들이 최소화되어 유의미한 정보들이 소실될 가능성이 있다.

Ⅲ. 연관성 모델

이번 장에서는 효율적이고 효과적인 오피넌마이닝을 위하여 연관규칙탐사 모델을 수정한 연관성 모델을 제안한다.

일반적으로 상품이나 서비스의 개체나 속성에 대한 사용자들의 감정표현은 속성과 감정단어가 연속으로 나타나게 된다. 예를 들어, “영화가 재미있고 감동적이예요” 라는 문장에서 “영화”는 개체(entity)에 해당하고 “재미있다”와 “감동적이다” 는 감성어휘가 된다. 온라인 고객리뷰에서 개체와 감성어휘가 연속으로 나타나는 부분을 본 논문에서는 다음과 같이 감성문장(sensitivity sentence) 이라고 정의한다.

<정의1> E가 고객리뷰에 대한 개체집합이고, S가 감성어휘 집합이라고 하자. 이 때, 개체 $e \in E$와 감성어휘 $s \in S$ 가 (e, s)와 같은 형태로 구성될 때, 이를 감성문장이라고 한다. □

예를 들어, “영화가 재미있어요”는 “영화”에 대한 감성문장이 된다. 또한 실제 하나의 문장 내에 두개 이상의 감성어휘가 포함될 수 있는데, 이 경우 해당문장은 두개의 감성문장으로 분리된다. 예를 들어, “영화가 재미있고 감동적이예요” 는 ”영화가 재미있어요“와 ”영화가 감동적이예요“ 로 분리된다.

온라인 고객리뷰 내의 감성문장에 포함된 개체들 중에서 자주 언급되는 개체들은 고객들의 주요 관심대상이므로 상대적으로 중요한 개체들이다. 연관성모델에서는 자주 출현하는 개체들을 우선적인 분석대상으로 설정하기 위하여 개체들의 출현 빈도를 측정하기 위한

개체 지지도를 다음과 같이 정의한다.

<정의2> 개체 지지도(support of entity)

온라인 고객리뷰내의 감성문장들에서 개체 e가 얼마나 자주 출현하는지를 나타내는 척도

$$s(e) = \frac{n(e)}{\text{감성문장들의수}}$$

(n(e) : 감성문장들에서 개체 e의 출현 빈도) □

온라인 고객리뷰 내의 감성문장에 포함된 감성어휘들 중에서 자주 출현하는 감성단어들은 고객의 의견을 파악하는 실마리가 된다. 연관성모델에서는 자주 나타나는 감성단어들을 우선적인 분석대상으로 설정하기 위하여 감성단어들의 출현 빈도를 측정하기 위한 감성 지지도를 다음과 같이 정의한다.

<정의3> 감성 지지도(support of sensitivity)

온라인 고객리뷰내의 감성문장들에서 감성단어 s가 얼마나 자주 출현하는지를 나타내는 척도

$$s(s) = \frac{n(s)}{\text{감성문장들의수}}$$

(n(s) : 감성문장들에서 감성어휘 s의 출현 빈도) □

온라인 고객리뷰에서 고객이 관심을 갖는 개체와 감성단어 사이의 연관성은 의미있는 정보가 될 수 있다. 예를 들어, 디지털카메라와 관련된 온라인 고객리뷰에서 '화질' 이라는 단어와 '좋다' 라는 단어가 동시에 언급되는 빈도가 많을 경우 카메라의 화질에 대한 고객들의 의견은 긍정적임을 의미한다. 개체와 감성단어 사이의 연관성을 측정하기 위한 연관도를 다음과 같이 정의한다.

<정의4> 연관도(degree of association)

온라인 고객리뷰에서 개체 e와 감성어휘 s가 서로 얼마나 관련되어 있는지 나타내는 척도

$$a(e,s) = \frac{s(e \cap s)}{s(e)} \text{ (단, } e \in E, s \in S \text{)} \square$$

Ⅳ. 시스템 설계

'이번 장에서는 연관성 모델을 기반으로 한 오피넌마이닝 시스템을 설계한다.

4.1 오피넌마이닝 과정

그림1에서는 연관성에 기반한 오피넌마이닝 과정을 나타내고 있다.

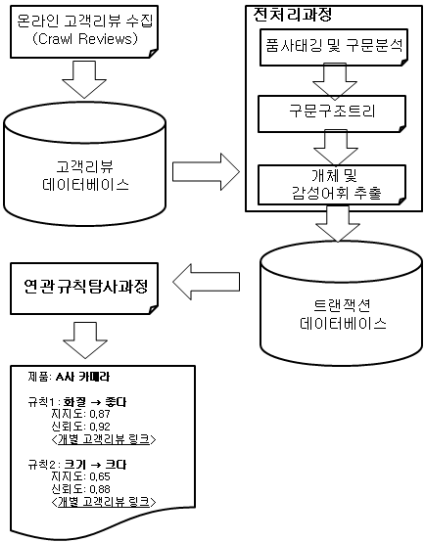

그림 1. 오피넌 마이닝 과정
Fig. 1 Opinion Mining Process

웹사이트 상에 게시된 온라인 고객리뷰들은 텍스트문서 형태로 고객리뷰 데이터베이스에 저장된다. 텍스트문서는 비정형 데이터이므로 전처리과정(preprocessing)을 통하여 정형 데이터인 테이블 파일들로 이루어진 트랜잭션 데이터베이스로 변환된다. 전처리과정 중에서 텍스트문서내의 각 한글문장들은 한국어 구문분석기에 의하여 각 단어들에 품사가 부여된 형태의 구문구조트리로 변환된다. 구문구조트리 파일로부터 개체와 감성어휘에 해당하는 단어들이 추출되어 트랜잭션 데이터베이스에 저장된다.

연관규칙탐사 과정 동안에는 트랜잭션 데이터베이스 내의 테이블화일로부터 각 개체와 감성어휘들에 대한 지지도가 계산되고 또한 개체와 감성어휘 사이의 연관도가 계산되면서 연관규칙들이 생성된다.

4.2 알고리즘

그림2는 개체 및 감성어휘 추출모듈의 개략적인 알고리즘을 나타내고 있다. 구문분석된 고객리뷰들의 집합 T의 각 고객리뷰 t_i로부터 개체 e_{ij}와 감성어휘 s_{ij}를 추출

하고 테이블파일 R의 레코드 i에 삽입하고 있다. R의 각 레코드는 하나의 감성문장에 대응된다.

개체 및 감성어휘 추출 알고리즘

입력: 구문구조트리 집합 T
출력: 테이블 화일 R

```
구문구조트리에서 감성어휘의 수식을 받는 개체와
감성어휘 추출
BEGIN
 while each t_i in T {
    /* t_i는 T안에 있는 각 구문구조트리 */
   if (t_i내에 감성어휘 s_ij가 있다면)
    /* e_ij는 s_ij로부터 수식을 받는 개체 */
        insert (개체, 감성어휘) into R
              values(e_ij, s_ij);
 }
END
```

그림2. 개체 및 감성어휘 추출 알고리즘
Fig.2 algorithm for extracting entities and sensitive vocabulary

개체 지지도 계산 알고리즘

입력: 테이블화일 R
출력: 테이블화일 E

```
 테이블화일 R에 있는 각 명사들의 출현빈도수를
계산하여 테이블화일 E에 입력
 BEGIN
 total = rec# of R; /* 테이블 R의 레코드 개수 */
 while each r_i in R        /* r_i는 R의 각 레코드 */
      if (r_i.개체 <> null){
        select count(*) as n from R
              where 개체 like r_i.개체;
        update R set 개체 = null
                  where 개체 like r_i.개체;
         } /* end of if */
      insert into E(개체, 지지도)
           values(개체속성값, n/total);
 } /* end of while */
 END
```

그림3. 개체 지지도 계산
Fig. 3 calculation for entity support

그림3은 개체 지지도 계산모듈에 대한 개략적인 알고리즘을 나타내고 있다. 테이블파일 R의 각 레코드 r_i는 하나의 감성문장에 대응된다. 테이블파일 R의 각 레코드 r_i의 개체속성 값의 지지도를 계산하기 위해서 SQL의 select명령과 update명령, like연산자를 이용하고 있다. 각 개체의 지지도는 2개의 속성 개체와 지지도로 구성된 새로운 테이블 파일 E에 삽입된다.

감성어휘 그룹핑 및 지지도 계산 알고리즘

입력: 테이블 화일 R, 그룹핑 수준 g
출력: 감성어휘 화일 S

```
그룹핑 수준에 따라 동일 부류의 감성어휘들을
그룹화하고 지지도를 계산함
BEGIN
 total = rec# of R; /* 테이블 R의 레코드 개수 */
 while each r_i in R {
    /* r_i는 R 안에 있는 감성어휘 */
  if (r_i.감성어휘 <> null){
       ss = substring(r_i.감성어휘, g);
         /* g는 그룹화 수준 */
       select count(*) as n from R
           where 감성어휘 like '%'+ss+'%';
        /* n은 그룹에 포함된 감성어휘 수*/
       update R set 감성어휘 = null
           where 감성어휘 like '%'+ss+'%';
       insert (감성그룹, 감성어휘, 지지도)
       into S values(ss, r_i.감성어휘, n/total);
   }/* end of if */
 } /* end of while */
END
```

그림4. 감성어휘의 그룹핑 및 지지도 계산
Fig.4 grouping and support calculation for sensitive vocabularies

그림4는 감성어휘를 그룹핑하기 위한 개략적인 알고리즘을 나타내고 있다. 그룹핑 수준은 감성어휘를 나타내는 문자열의 문자 부합 정도를 의미한다. 그룹핑 수준이 '1'인 경우, 문자열의 첫 번째 문자가 동일하면 동일그룹으로 분류한다. 예를 들면, '지루한'이라는 감성어휘는 '지루했던', '지나친' 등의 감성어휘와 동일그룹이 된다. 그룹핑 수준이 2'인 경우, 문자열의 두번째 문자까지 동일하면 동일그룹으로 분류한다. 예를 들어, 그룹핑

수준이 2이면 '지루한'이라는 감성어휘는 '지나친' 이라는 감성어휘와는 동일그룹이 되지 않는다.

연관도 계산 알고리즘

입력: 테이블 화일 R, E, S
출력: 연관규칙

E의 각 개체와 S의 각 감성어휘 사이의 연관도를 계산하여 유의미한 연관규칙을 생성한다.

```
BEGIN
 while each ri in E
    /* ri는 E의 각 레코드, 1≤ i ≤ rec# of E  */
   if (ri.지지도 ≥ t1) {  /* t1은 최소빈발도,
                        ri.지지도: 개체 지지도  */
       create 뷰 Rv (select * from R
                    where 개체 like ri.개체);
       total = rec# of Rv ;
       while each sj in S
                        /* sj는 S의 각 레코드,
                           1≤ j ≤ rec# of S  */
            if (sj.지지도 ≥ t1) {
                /sj.지지도 : 감성어휘 지지도 */
              select count(*) as n from Rv
                where 감성어휘 like sj.감성어휘;
              a = n / total;
              /* ri.개체와 sj.감성어휘의 연관도 */
              if (a ≥ t2) /* t2는 최소연관도 */
                 「sj.감성어휘→ri.개체 」규칙 생성
            }/* end of if */
         }  /* end of while */
   }  /* end of if */
 } /* end of while */
END
```

그림5. 연관규칙 생성
Fig.5 Generation of association rules

그림5는 개체와 감성어휘 사이의 연관규칙을 생성하는 개략적인 알고리즘을 나타내고 있다. 테이블파일 E의 각 레코드 r_i는 고객리뷰 상의 각 개체 'r_i.개체'에 대한 지지도 'r_i.지지도'를 포함한다. 'r_i.개체'와 연관된 감성어휘의 연관도를 계산하기 위하여 먼저 감성문장을 포함하는 테이블파일 R로부터 'r_i.개체'에 대한 뷰 R_v를 생성한다. R_v로부터 's_i.감성어휘'의 출현빈도를 계산하고 'r_i.개체'와의 연관도 a를 계산한다. a가 최소연관 t_2보다 크면 연관규칙 's_j.감성어휘 → r_i.개체' 가 생성된다.

V. 시스템 구현 및 성능평가

본 논문에서는 4장에서 제시한 알고리즘들을 바탕으로 연관성 기반의 오피넌마이닝시스템을 개발하였다. 프로그램언어로는 비주얼베이직을 사용하였고 DBMS는 MS SQL server 2008을 사용하였다. 한국어구문분석을 위하여 국내의 대표적인 구문분석기[12,13]을 사용하였다. 실험용 데이터는 네이버랩(lab.naver.com)에서 제공하는 영화 "해운대" 40자평 데이터세트를 사용하였다. 데이터세트에는 약 1만개의 고객리뷰가 포함된다. 개발된 시스템은 CPU 1.73GHz, 메모리 1GB, 윈도XP가 탑재된 환경에서 수행되었다.

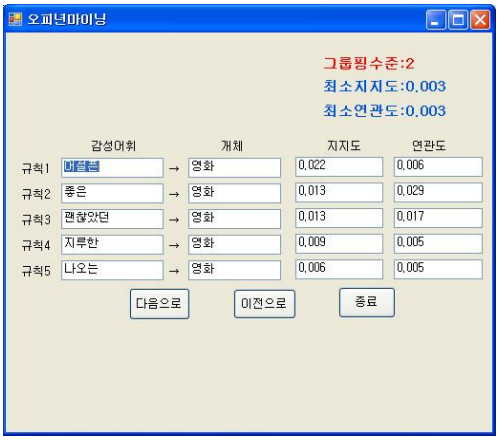

그림6. 시스템 구현 예(최소지지도 : 0.003, 최소연관도: 0.003, 그룹핑 수준 2)
Fig.6 System Implementation(minimum support :0.003, minimum association:0.003, grouping:2)

그림6은 최소빈발도와 최소관련도를 0.003으로 하고 그룹핑 수준을 2로 설정하였을 때 생성된 연관규칙들의 일부를 나타내고 있다. '영화'라는 개체에 대하여 '좋았고' '재미있다' 라는 긍정적 의견들도 있지만, '지루하거나' '어설프다' 라는 부정적인 의견도 존재했다. '나오는' 과 같은 의미없는 의견도 일부 존재했는데, 이는 한국어 구문분석기가 테스트버전이라서 그 정확성이 다소 떨어지기 때문인 것으로 판단된다.

그림7은 최소지지도와 최소연관도의 변화에 따라 생성되는 연관규칙들의 수가 어떻게 달라지는지 보여주는 그래프이다. 최소지지도와 최소연관도를 낮게 설정할 수록 생성되는 연관규칙들의 수는 증가하고 있음을 알 수 있다.

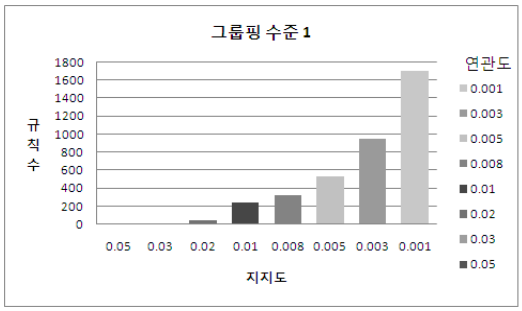

그림7. 지지도와 연관도에 따른 연관규칙들의 수 (그룹핑 수준 1)
Fig.7 number of association rules in variations of support and association degree(grouping:1)

그림8은 그룹핑 수준에 따라 생성되는 연관규칙들의 수가 달라짐을 나타내고 있다. 그룹핑 수준을 높게 (1) 설정할 수록 더 많은 연관규칙들이 생성됨을 알 수 있다.

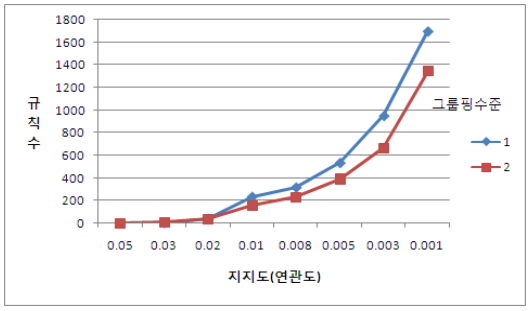

그림8. 그룹핑 수준에 따른 연관규칙들의 수
Fig.8 number of associations in variations of grouping

VI. 결 론

본 논문에서는 데이터마이닝과 텍스트마이닝 분야의 새로운 연구영역인 오피넌마이닝 방법에 대하여 살펴보았다. 웹 2.0 시대에 온라인 고객리뷰들이 더욱 증가될 것이라는 측면에서 오피넌마이닝 기술의 중요성은 더욱 커질 것이다. 오피넌마이닝 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 연관규칙탐사 모델을 오피넌마이닝에 적합하게 수정한 연관성모델을 제안하였으며, 연관성모델을 기반으로 한 오피넌마이닝시스템을 설계하고 구현하였다. 연관성 모델에 기반한 오피넌마이닝시스템은 단순히 긍정과 부정의견으로만 분류하였던 기존의 오피넌마이닝 방법에 비하여 보다 연관규칙형태의 보다 풍부한 고객의견 정보를 도출할 수 있었다. 또한, 감성어휘의 그룹핑 수준을 높일 수록 보다 많은 연관규칙을 생성하였다.

참고문헌

[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] Salton, G. Singhal, A.Buckley, C. and Mitra, M., Automatic Text Decomposition using Text Segments and Text Themes", ACM Conference on Hypertext, 1996.

[3] Boguraev, B., and Kennedy, C.,"Salience-Based Content Characterization of Text Documents", Proc. of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, 1997.

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

[5] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, Chun Jin, " Red Opal: Product-Feature Scoring from Reviews", Proc. of the 8th ACM conference on Electronic commerce, pp.11-15, 2007.

[6] Xiaowen Ding, and Bing Lui, "The Utility of Lingusitic Rules in Opinion Mining", SIGR pp.811-812, 2007.
[7] Courses, E., and Surveys, T., "Using SentiWordNet for multilingual sentiment analysis", Data Engineering Workshop ICDEW 2008.
[8] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", KDD'04, pp.168-177, 2004,
[9] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", KDD'04, 2004, pp.168-177.
[10] Xiaowen Ding, Bing Liu and Philip S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", WSDM'08, 2008, pp.231-239.
[11] 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.
[12] Korean Parser Test Version, http://nlp.kookmin.ac.kr/HAM/kor/download. html.
[13] 강승식, 한국어 형태소분석과 정보검색, 홍릉과학출판사, 2003.

저자소개

김근형(Keun yung Kim)

1990년 2월: 서강대 컴퓨터학과 (공학사)
1992년 2월: 서강대 컴퓨터학과 (공학석사)
2001년 2월: 서강대 컴퓨터학과(공학박사)
2001년 9월 ~ 현재 : 제주대학교 경영정보학과 부교수
※관심분야 : 데이터마이닝, 텍스트마이닝