http://dx.doi.org/10.3745/KIPSTD.2011.18D.4.237

개체연관망 모델에 의한 오피니언마이닝의 확장

김 근 형†

요 약

오피니언마이닝은 대량의 온라인 고객리뷰에서 상품이나 서비스의 속성들에 대한 고객들의 주관적 의견을 긍정과 부정으로 분류하여 요약한다. 그러나, 고객들의 관심사항은 주관적 의견뿐만 아니라 객관적 사실을 통해서도 표현되기 때문에 주관적 의견만을 주요 분석대상으로 하는 기존 오피니언마이닝 기법을 확장할 필요가 있다.

본 논문에서는 주관적 의견뿐만 아니라 객관적 사실도 분석대상으로 하는 개체연관망 모델을 사용하여 기존 오피니언마이닝의 분석능력을 확장한다. 개체연관망 모델은 각 개체에 대한 긍정부정 정도를 표현할 뿐만 아니라 개체들 사이의 연관관계와 상대적 중요성을 나타낼 수 있다.

시스템 구현 결과, 개체연관망 모델에 기반한 오피니언마이닝시스템은 기존 기법에 비하여 보다 풍부한 정보를 추출할 수 있음을 확인하였다.

키워드 : 오피니언마이닝, 개체연관망, 빈발도, 명암도, 연관도

Expansion of Opinion Mining based on Entity Association Network Model

Keunhyung Kim†

ABSTRACT

Opinion Mining summarizes with classifying sensitive opinions of customers in huge online customer reviews for the attributes of products or services by positive and negative opinions. Because the customers represent their interests through subjective opinions as well as objective facts, the existing opinion mining techniques, which can analyze just the sensitive opinions, need to be expanded.. In this paper, We propose the novel entity association network model which expands the existing opinion mining techniques. The entity association model can not only represent positive and negative degree of the sensitive opinions, but also can represent the degree of the associations and relative importances between entities.

We designed and implemented the customer reviews analysis system based on the entity association network model. We recognized that the system can represent more abundant information than the existing opinion mining techniques.

Keywords : Opinion Mining, Entity Association Network, Degree of Frequency, Degree of Shade, Degree of Association

1. 서 론

웹2.0의 등장으로 인터넷에서의 네티즌 역할은 단순한 정보의 사용자에서 생산자(제공자)로 확대되었다. 현재 인터넷 상에는 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.

오피니언마이닝은 대량의 온라인 고객리뷰에서 상품이나 서비스의 속성들에 대한 고객들의 주관적 의견을 긍정과 부정의견으로 분류하여 요약한다[1, 2, 3]. 그러나, 고객들의 관심사항은 주관적 의견뿐만 아니라 객관적 사실을 통해서도 표현되기 때문에 주관적 의견과 객관적 사실을 서로 연관시킬 수 있다면 보다 심도있는 분석을 할 수 있다. 본 논문에서 주관적 의견은 감성적 형용사가 포함된 문장을 의미하며, 객관적 사실은 감성적 형용사를 포함하지 않는 문장으로 한정한다.

예를 들어, 디지털카메라에 대한 온라인 고객리뷰에서 (그림 1)과 같은 2개의 고객리뷰 사례를 보자. <고객리뷰 1>의 경우, 직업이 사진작가인 고객은 카메라의 화질에 관심이 있음을 알 수 있다. 반면, <고객리뷰 2>에서 운동선수인 고객은 카메라의 크기에 관심이 있다. 카메라 제조회사에서는 이러한 정보를 이용하여 크기는 좀 크지만 화질이

† 종신회원 : 제주대학교 경영정보학과 교수
논문접수 : 2011년 1월 24일
수 정 일 : 1차 2011년 5월 6일, 2차 2011년 6월 22일
심사완료 : 2011년 6월 23일

좋은 사진작가용 카메라와 화질은 보통이지만 크기가 작은 운동선수용 카메라 등으로 제품 다양화를 시도할 수 있을 것이다. '사진작가'나 '운동선수'와 같은 개체는 주관적 의견이 반영되지 않는 객관적 사실을 표현하는데 이용되고 있는 반면, '화질'이나 '크기'와 같은 개체는 '좋다' 라든가 '불편하다'와 같은 감성적 형용사와 함께 주관적 의견을 표현하는데 이용되고 있다.

〈고객리뷰 1〉
저의 직업은 사진작가입니다. 카메라의 성능은 뭐니뭐니 해도 화질이 제일 중요하죠. 저는 A사의 카메라를 선호하는데 화질이 좋기 때문입니다.

〈고객리뷰 2〉
저는 운동선수입니다. 카메라 크기는 작을수록 좋습니다. A사 카메라의 크기는 너무 커서 불편한 측면이 있군요.

(그림 1) 온라인 고객리뷰의 예

기존 오피니언마이닝 기법에서는 카메라의 화질과 크기에 대한 고객들의 의견은 추출할 수 있으나 '직업이 사진작가'라는 객관적 사실이나 개체들 사이의 연관성 즉, 사진작가는 카메라의 화질에 관심이 있고, 운동선수는 카메라의 크기에 관심이 있다는 정보는 추출할 수 없다.

본 논문에서는 개체연관망 모델을 사용하여 기존 오피니언마이닝의 분석능력을 확장하고자 한다. 개체연관망 모델을 통하여 다음과 같이 분석능력을 확장할 수 있다. 첫째 감성적 형용사에 의하여 수식되는 주관적 개체뿐만 아니라 감성적 형용사에 의하여 수식되지 않는 객관적 개체들도 분석대상으로 하며, 각 개체에 대한 고객의견 방향을 긍정도, 부정도, 중립도 값으로 세분화하여 표현할 수 있다. 둘째, 주관적 개체와 객관적 개체들을 혼합하여 그 연관성을 분석할 수 있다. 셋째 각 개체들의 출현빈도를 바탕으로 하여 개체들 사이의 상대적 중요성을 분석할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 고찰하고 3장에서는 개체연관망 모델을 제안한다. 4장에서는 개체연관망 모델을 기반으로 한 분석시스템을 설계한다. 5장에서는 구현시스템의 성능을 평가하고 6장에서 결론을 맺는다.

2. 관련 연구

2.1 연관규칙탐사

데이터마이닝 분야에서 연관규칙탐사는 많은 연구가 이루어진 분야로서 대량의 데이터(관계형 화일)로부터 속성(변수)들 사이의 연관성을 규칙형태로 추출하는 데이터분석 기술이다[4]. 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_n으로 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∪Y를 포함하는 레코드 수의 비율을 나타낸다. 즉, 지지도의 의미는 추출된 연관규칙「X→Y」가 얼마나 많은 고객들에게 적용되는 규칙인지를 나타내는 척도로서, 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.

2.2 문서요약

기존의 문서요약 기술은 2가지 유형으로 나누어진다. 하나는 원형틀(template, 원형판)을 채워 넣은 방식이고, 다른 하나는 핵심문장을 추출하는 방식이다[5, 6]. 원형틀을 채워 넣는 방식은 문서안의 핵심 개체(entity)나 사실(fact)을 식별·추출하여 원형틀의 각 슬롯(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 추출함으로서 문서내용을 간략화 한다. 핵심문장 추출방식은 길이가 긴 단일 문서의 간략화를 목적으로 하기 때문에 길이가 짧은 대량의 다중문서로 구성된 온라인 고객리뷰를 분석하기 위한 방법으로는 적합하지 않다.

2.3 오피니언마이닝

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

[7]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[8], 이는 고객리뷰 데이터와 사용자 평가점수를 활용하여 요약보고서를 생성하는 기법을 제안하였다. [9]에서는 문장구조와 문장 사이의 관계, 문장성분의 패턴정보 등의 언어규칙을 이용한 통계학적 방법으로 오피니언마이닝에 접근하고 있다. [1, 2, 10]에서는 워드넷을 활용하여 어휘의 긍정이나 부정적 의미를 판단하고 이를 센티워크넷(SentiwordNet)으로 응용하여 감정의 폭을 정량화하는 방법을 제시하고 있다.

[3]은 오피니언마이닝 과정에서 데이터마이닝의 연관규칙 탐사기법을 적용하여 개체와 감성어휘 사이의 연관규칙을 추출하는 기법을 제안하고 있다. 그러나 개체의 긍정부정 정도를 표현할 수 없으며 개체와 개체사이의 연관성도 추출할 수 없어 정보 표현력의 한계가 있다.

오피니언마이닝 연구의 핵심은 주관적 고객리뷰에 대해서 긍정 혹은 부정적 의견을 자동으로 판단하는 것이다. 그러나, 고객의 관심은 주관적 의견뿐만 아니라 객관적 사실을 통해서도 표현되므로 온라인 고객리뷰의 보다 정확한 분석은 주관적 의견과 객관적 사실 모두를 분석대상으로 설정할 필요가 있다.

3. 개체 연관망 모델

이번 장에서는 온라인 고객리뷰를 보다 정확하고 심층적으로 분석하기 위한 개체 연관망 모델을 새롭게 제안한다.

개체연관망(entity association network)은 온라인 고객리뷰와 같은 대량의 다중문서에 대하여 고객들이 관심을 갖는 개체들과 이들 사이의 연관성을 표현하는 형태로 다중문서를 간략화 시킨다. 온라인 고객리뷰에서 언급되는 개체들은 주로 명사들로 표현된다. 본 논문에서는 일반명사와 고유명사를 개체라고 지칭하기로 한다. 오피니언마이닝은 주관적 의견이 반영된 개체들만을 다루는데 반하여, 개체 연관망에서는 주관적 의견뿐만 아니라 주관적 의견이 반영되지 않은 개체들까지 고려하기 때문에 내포하고 있는 정보량이 더 많아지게 된다.

온라인 고객리뷰에서 자주 언급되는 개체들은 고객들의 주요 관심사항임을 나타내는 것이므로 개체 연관망에서는 자주 출현하는 개체들을 우선적인 분석대상으로 설정한다.

온라인 고객리뷰 상에서 개체들이 얼마나 자주 나타나는지 측정하기 위한 척도가 빈발도이다.

<정의 1> 빈발도(degree of frequency)

온라인 고객리뷰에서 개체 A가 얼마나 자주 출현하는지를 나타내는 척도

$$f(A)=\frac{c(A)}{sizeof(E)}$$

(E: 전체 개체집합, c(A) : 개체 A의 출현 빈도) □

온라인 고객리뷰에서 자주 언급되는 개체들 중에서도 감성어휘들에 의하여 고객들의 주관적 의견이 표현된 개체들이 있다. 감성단어는 “재미있다”, “지루하다”, “멋있다” 등 대부분 형용사 등으로 구성된다. 감성단어에 대한 데이터베이스 구축을 위해서는 감성 단어의 극성이 고려되어야 한다. 예를 들어, “재미있다”는 긍정의 의미이고, “지루하다”는 부정의 의미를 담고 있다. 이러한 감성단어를 바탕으로 고객들의 주관적 의견을 표현하기 위한 척도가 명암도이다. 명암도(degree of shade)는 긍정도, 부정도, 중립도로 이루어지며, (긍정도, 중립도, 부정도)의 형태로 표현된다. 주관적 의견이 얼마나 긍정적인지를 평가할 수 있는 척도가 긍정도이고 얼마나 부정적인지를 나타내는 척도가 부정도이다. 주관적 의견이 얼마나 개입되지 않았는지를 나타내는 척도는 중립도이다.

<정의 2> 긍정도(positive degree)

온라인 고객리뷰에서 개체 A가 감성어휘에 의하여 수식될 때 긍정의 정도를 나타내는 척도

$$s^{+}(A)=\frac{c^{+}(A)}{c(A)}$$

(c(A) : 개체 A의 출현 빈도, $c^{+}(A)$: A를 수식하는 감성어휘 중 긍정적인 어휘의 빈도) □

<정의 3> 부정도(negative degree)

온라인 고객리뷰에서 개체 A가 감성어휘에 의하여 수식될 때 부정의 정도를 나타내는 척도

$$s^{-}(A)=\frac{c^{-}(A)}{c(A)}$$

(c(A) : 개체 A의 출현 빈도, $c^{-}(A)$: A를 수식하는 감성어휘 중 부정적인 어휘의 빈도) □

<정의 4> 중립도(neutral degree)

온라인 고객리뷰에서 개체 A가 감성어휘에 의하여 수식되지 않는 정도를 나타내는 척도

$$s^{0}(A)=\frac{c^{0}(A)}{c(A)}$$

(c(A) : 개체 A의 출현 빈도, $c^{0}(A)$: A를 수식하지 않거나 수식하는 어휘 중 감성어휘가 아닌 빈도) □

온라인 고객리뷰에서 고객들이 관심을 갖는 개체들 사이의 관계성이나 연계성은 의미있는 정보가 될 수 있다. 예를 들어, 디지털카메라와 관련된 온라인 고객리뷰에서 ‘크기’라는 개체와 ‘가격’이라는 개체가 동시에 언급되는 빈도가 많을 경우 카메라의 크기에 관심있는 고객들은 가격에도 관심이 있음을 의미한다.

<정의 5>연관도(degree of association)

온라인 고객리뷰에서 개체 A와 개체 B가 서로 얼마나 연관되어 있는지 나타내는 척도

$$a(A,B)=\frac{c(A\cap B)}{c(A)}$$ (단, c(A) ≥ c(B)) □

<정의 6> 연관성(association)

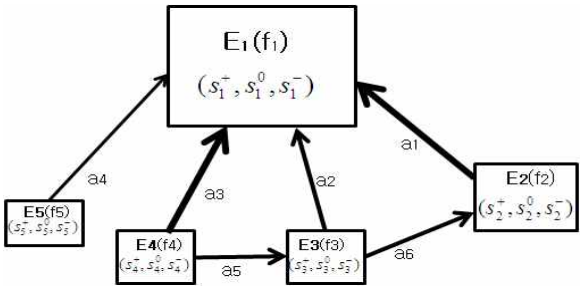
B → A : 개체 A와 개체 B에 대하여 f(A) ≥ t_1, f(B)≥ t_1, f(A) ≥ f(B), a(A, B) ≥ t_2 일 때, B는 A에 연관된다고 한다(단, t_1은 최소빈발도, t_2는 최소연관도) □

<정의 7> 개체 연관망

E에 속하는 모든 개체들 간의 연관성 관계를 나타낸 다이어그램

{ B → A | A ∈ E, B ∈ E, E는 리뷰 안에 있는 개체들의 집합} □

(그림 2)은 개체 연관망을 나타내고 있다. E_i는 개체들이고 f_i는 해당 개체의 빈발도, a_i는 연관도를 의미한다. (s_i^+, s_i^0, s_i^-)에서 s_i^+는 개체 E_i의 긍정도, s_i^-는 개체 E_i의 부정도, s_i^0는 개체 E_i의 중립도를 나타낸다.



(그림 2) 개체연관망의 예

개체연관망은 특정개체의 긍정·부정 정도를 표현할 수 있을 뿐만 아니라 개체 사이의 연관성을 나타낼 수 있으므로 기존 오피니언마이닝 기법을 더 확장한 형태가 된다. 개체 E_1과 E_2의 관심정도는 f_1과 f_2를 통하여 나타낼 수 있다. 개체 E_1과 E_2의 긍정부정 정도는 (s_1^+, s_1^0, s_1^-)와 (s_2^+, s_2^0, s_2^-)의 값으로 표현된다. 개체 E_1과 E_2의 연관도는 a_1이고 개체 E_2가 개체 E_1에 연관되어 있음을 알 수 있다.

4. 시스템 설계

이번 장에서는 개체연관망 모델을 기반으로 하는 온라인 고객리뷰 분석시스템을 설계한다.

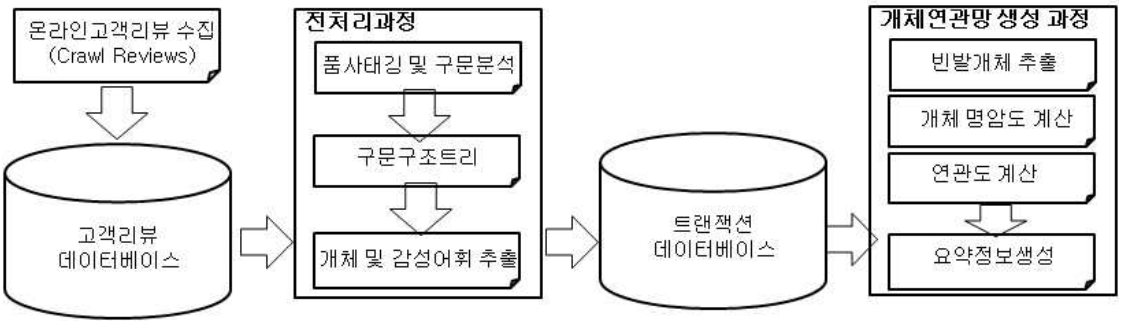
4.1 개체연관망 생성과정

(그림 3)에서는 개체연관망의 생성과정을 나타내고 있다. 웹사이트 상에 게시된 온라인 고객리뷰들은 텍스트문서 형태로 고객리뷰 데이터베이스에 저장된다. 텍스트문서는 비정형 데이터이므로 전처리과정(preprocessing)을 통하여 정형 데이터인 테이블 파일로 변환된다. 전처리과정 중에서 텍스트문서내의 각 한글문장들은 한국어 구문분석기에 의하여 각 단어들에 품사가 부여된 형태의 구문구조트리로 변환된다. 구문구조트리 파일로부터 개체단어와 감성어휘 단어들이 추출되어 트랜잭션 데이터베이스의 테이블 파일들에 저장된다. 고객리뷰들은 그 양이 방대하고 특히 비구조화된(unstructured) 텍스트 데이터 형태로 존재하므로 관계형 테이블(relational table)과 같은 구조화(structured)된 형태로 변환하여 처리하는 것이 바람직하다. 앞으로 온라인 고객리뷰의 게시는 더욱 활성화되어 온라인 고객리뷰의 양은 더욱 방대해질 것으로 예상되므로 대용량의 관계형 테이블 데이터에 대하여 가장 최적화된 처리가 가능한 SQL을 적용하는 것은 그 효율성을 증가시킬 수 있다.

개체연관망 생성과정 동안에는 테이블 파일안의 각 개체들에 대하여 빈발도와 명암도가 계산되고 또한 개체들 사이의 연관도가 계산되면서 개체연관망이 만들어진다.

4.2 알고리즘

(그림 4)에서는 감성단어에 대한 정보가 관리된다는 가정 하에 개체 및 감성어휘 추출모듈의 개략적인 알고리즘을 나타내고 있다. 구문분석된 고객리뷰들의 집합 T의 각 고객리뷰 t_i로부터 개체 e_{ij}와 감성어휘 sij를 추출하고 테이블파일 E과 S에 삽입하고 있다. 테이블화일 E에는 각 고객리뷰에 나타난 개체들이 삽입되고 S에는 감성어휘에 의하여 수식되는 개체들이 삽입된다. E의 각 레코드에는 하나의 고객리뷰가 대응되고 S의 각 레코드에는 하나의 감성문장이 대응된다.



(그림 3) 개체연관망 생성과정

개체 및 감성어휘 추출

입력: 구문분석된 고객리뷰들의 집합 T
출력: 테이블화일 E, S /* E: 전체 개체 포함,
S: 감성어휘에 의하여 수식되는 개체 포함 */

```
구문분석된 고객리뷰들로부터 개체들을 추출하여
테이블화일 E와 S생성
BEGIN
 while each ti in T {
        /* ti는 T안에 있는 각 고객리뷰들 */
    find ei1, ei2,...eim in ti ;
                    /* eij는 ti에 있는 각 개체 */
                    /* m은 E의 속성 수 */
    insert into E values(ei1, ei2,...eim);
    while each eij in ti
      if (eij를 수식하는 감성어휘 sik가 있다면)
                insert (개체, 감성어휘) into S
                    values(eij, sij);
    } /* end of file */
 }/* end of while */
END
```

(그림 4) 개체 및 감성어휘 추출 알고리즘

빈발도 계산 알고리즘

입력: 테이블화일 E
출력: 테이블화일 E_c

```
테이블화일 E에 있는 각 개체들의 출현빈도수를 계산하여
테이블 Ec에 입력
BEGIN
1 total = rec# of E;
2 while each ri in E        /* ri는 E의 각 레코드 */
3   while each eij in ri      /* eij는 ri의 각 속성 값 */
4       if (eij != null){
5           f = 0;
6           while(E의 각 속성 ak에 대하여)
7               select count(*) as t from E
                            where ak like eij ;
                              /* ak는 레코드의 각 속성*/
8               update E set ak = null where ak like eij;
9               f = f+t;  /* f는 명사 eij의 출현 빈도수*/
            } /* end of while */
10          dof = f/total; /*dof : degree of frequency */
11          insert into Ec(개체, 빈발도) values(eij, dof);
        }   /* end of if */
    } /* end of while */
 } /* end of while */
END
```

(그림 5) 빈발도 계산

(그림 5)는 빈발도 계산모듈에 대한 개략적인 알고리즘을 나타내고 있다. 테이블 E의 각 레코드 r_i는 하나의 고객리뷰와 대응되고 e_{ij}는 i번째 고객리뷰 내의 j번째 개체에 대응된다. E의 각 레코드 r_i의 각 개체들 e_{ij}의 빈발도를 계산하기 위해서 E의 각 속성 a_k에 대하여 SQL의 select명령과 update명령, like연산자를 이용하고 있다(7행에서 9행). 빈발도 계산에 사용된 개체들은 다음 빈발도 계산 대상에서 제외시키기 위해서 update명령을 통하여 null로 변환된다. 10행에서 f는 명사 e_{ij}의 출현빈도수를 의미하고 total은 개체의 전체 개수를 나타내므로 dof(degree of frequency)는 e_{ij}의 빈발도가 된다. 11행에서 각 명사의 빈발도는 2개의 속성 '개체'와 '빈발도'로 구성된 새로운 테이블 E_c에 삽입된다.

(그림 6)에서는 명암도 계산 알고리즘을 개략적으로 나타내고 있다. 감성어휘와 해당 개체가 포함된 테이블 S를 입력으로 하여 각 개체의 명암도를 포함하는 테이블 S_c를 생성한다. S로부터 해당 개체의 뷰를 생성한 다음 그 뷰로부터 긍정도와 중립도, 부정도를 계산한다. 명암도 계산이 끝난 개체는 SQL의 update명령에 의하여 null값으로 변경되어 명암도의 중복계산을 방지한다.

명암도 계산 알고리즘

입력: 테이블화일 S
출력: 테이블 화일 S_c

```
테이블 S로부터 각 개체의
   명암도(긍정도, 중립도, 부정도)를 계산
BEGIN
 while each ri in S {
    /* ri는 S의 각 레코드 */
   if (ri.개체 != null){
     create 뷰 Sv (select * from S
                   where (개체 like ri.개체);
    /*ri.개체의 긍정도, 중립도, 부정도 계산 */
     Sv에서 si+(ri.개체) 계산;
     Sv에서 si0(ri.개체) 계산;
     Sv에서 si-(ri.개체) 계산;
     insert into Sc(개체, 긍정도, 중립도, 부정도)
                 values (si+, si0, si-);
     update S set 개체 = null
                 where (개체 like ri.개체);
    } /* end of if */
} /* end of while */
END
```

(그림 6) 명암도 계산 알고리즘

(그림 7)에서는 연관도를 계산하고 개체연관망을 생성하는 개략적인 알고리즘을 나타내고 있다. 입력화일로는 고객리뷰의 모든 개체들을 포함하는 파일 E, 파일 E에 나타나는 각 개체의 횟수를 포함하는 E_c, 그리고 감성어휘에 의하여 수식되는 개체의 명함도가 포함된 파일 S_c 등 3개의 화일을 사용하고 있다. E_c에 있는 각 개체 중 빈발도가 최소빈발도 이상인 개체들의 모든 쌍에 대하여 연관도를 계산한다(2행-12행). 개체 쌍의 연관도가 최소연관도 이상이면 각 개체의 빈발도와 명함도, 개체 쌍 사이의 연관도와 함께 개체연결망에 추가된다(13행-24행).

5. 시스템 구현 및 성능평가

본 논문에서는 앞에서 살펴보았던 개체연관망 모델과 각 알고리즘들을 바탕으로 온라인 고객리뷰 분석시스템을 개발하였다. 프로그램언어로는 비주얼베이직을 사용하였고

연관도 계산 및 개체연관망 생성

입력: 테이블화일 E, E_c, S_c
출력: 개체연관망

E_c 상에 있는 각 개체들 사이의 연관도를 계산하여 개체연관망을 생성한다.

```
BEGIN
1 while each r_i in E_c
            /* r_i는 E_c의 각 레코드, 1≤ i ≤ rec# of E_c  */
2   if (r_i.빈발도 ≥ t_1) {  /* t_1은 최소빈발도 */
3       j = i + 1;
4       create 뷰 E_v
          (select * from E where a_k like r_j.개체 );
                          /* a_k는 E의 레코드의 각 속성 */
5       t_v = rec# of E_v ; /* 뷰 E_v의 전체 레코드 수 */
6       while each r_j in E_c {  /* i+1 ≤ j ≤ rec# of E_c */
          if (r_j.빈발도 ≥ t_1){
7             f = 0;
8             while(E_c의 각 속성 a_k에 대하여){
9                 select count(*) as t from E_v
                                where a_k like r_j.noun;
10                  f = f + t;
                      /*f는 개체 r_j.개체의 출현 빈도수*/
11            } /* end of while */
12            a = f/t_v; /* r_i.개체와 r_j.개체의 연관도 */
            } /* end of if */
13        if (a ≥ t_2){ /* t_2는 최소연관도 */
14            if (r_i.개체 is in S_c) {/* r_s는 S_c의 레코드 */
15                S_i^+=rs.긍정도; S_i^0=r_s.중립도;S_i^-=r_s.부정도;
16              } else {
17                S_i^+ = 0; S_i^0 = 1;S_i^- = 0; }
18            if (r_j.개체 is in S_c) {/* r_s는 S_c의 레코드 */
19                S_j^+=r_s.긍정도; S_j^0=r_s.중립도;S_j^-=r_s.부정도;
20              } else {
21                S_j^+ = 0; S_j^0 = 1;S_j^- = 0; }
22            add r_i.개체 with (r_i.빈발도, (s_i^+, s_i^0, s_i^-)) to
              개체연결망;
23            add r_j.개체 with (r_j.빈발도, (s_j^+, s_j^0, s_j^-)) to
              개체연결망;
24            add (r_j.개체, r_j.개체) with a to 개체연관망;
            } /* end of if */
25        }   /* end of while */
26  }  /* end of if */
27 } /* end of while */
END
```

(그림 7) 연관도 계산 및 개체연관망 생성

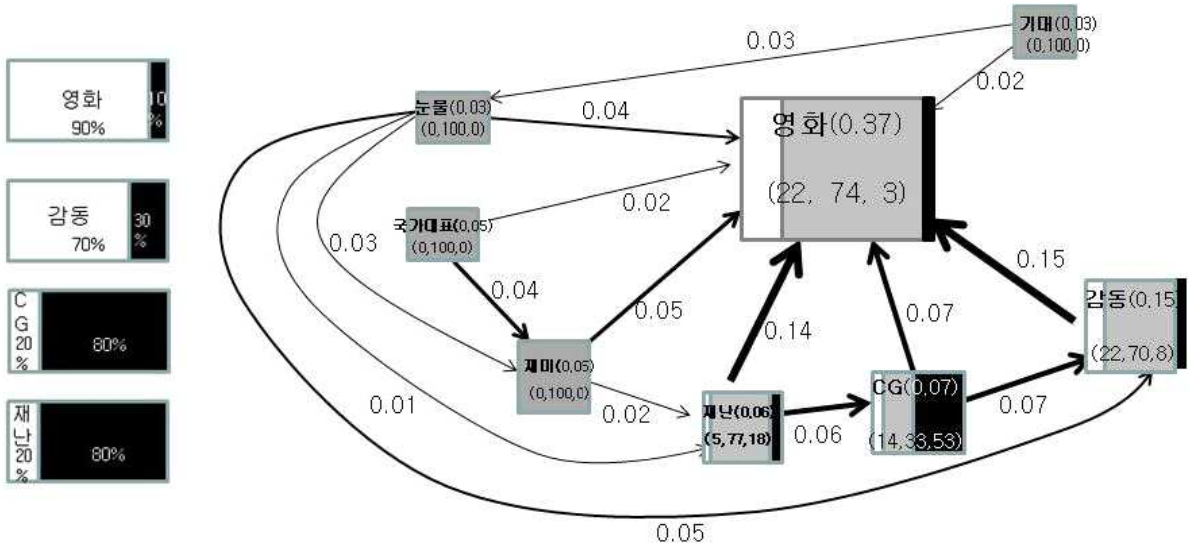
DBMS는 MS SQL server 2008을 사용하였다. 한국어구문분석을 위하여 국내의 대표적인 구문분석기[11,12]를 사용하였다. 실험용 데이터는 네이버랩(lab.naver.com)에서 제공하는 영화 "해운대" 40자평 데이터세트를 사용하였다. 데이터세트에는 약 1만개의 고객리뷰가 포함된다. 개발된 시스템은 CPU 1.73GHz, 메모리 1GB, 윈도XP가 탑재된 환경에서 수행되었다.

정보검색 분야에서 중요한 성능평가 기준으로 사용되는 척도는 정확도(precision)과 재현도(recall)이지만, 기존 오피니언마이닝 기법과 개체연관망 기법을 정확도나 재현도로 비교하기는 어렵다. 왜냐하면, 각 기법이 지향하는 목표가 다르기 때문이다. 기존 오피니언마이닝 기법은 감성어휘에 의하여 수식되는 개체들에 대해서만 긍정 부정 비율을 분석하고자 하는 반면, 개체연관망 기법은 이를 더 확장하여 감성어휘에 의하여 수식되지 않는 개체들도 분석대상으로 하며 개체들 사이의 연관성을 도출하고자 하는 목적을 갖는다.

따라서, 본 논문에서는 정성적 관점과 정량적 관점에서 기존 기법과 개체연관망 기법의 성능을 비교한다. 정성적 관점은 각 기법을 통하여 도출된 요약정보들이 어떻게 다른지 살펴보는 것이며, 정량적 관점은 요약정보를 구성하는 개체들의 수를 분석함으로써 요약정보의 구체화 정도를 비교하는 것이다.

(그림 8)은 "해운대" 영화평 데이터세트에 대하여 기존의 오피니언마이닝 기법[1, 7]과 개체연관망 기법을 적용하였을 때 각각의 분석결과의 일부를 비교하여 나타내고 있다(지면관계상 모든 개체들을 포함시키지 않았음). (그림 8)의 (a)에서 볼 수 있는 바와 같이, 기존 기법은 감성어휘에 의하여 수식되는 개체들만을 분석대상으로 설정하기 때문에 4개의 개체들 즉, '영화', '감동', 'CG', '재난' 에 대해서만 긍정부정 비율을 나타내고 있다. 각 직사각형 내부의 하얀색은 긍정비율, 검은색은 부정비율을 나타낸다. 영화에 대해서는 90% 정도가 긍정적으로 평가하고 있으며 컴퓨터그래픽(C·G)에 대해서는 80% 정도가 부정적으로 평가하고 있다. '영화' 개체를 수식하는 긍정적 감성어휘의 예로는 '재미있는', 완벽한', '괜찮은' 등이 포함되어 있었으며, 'CG' 개체를 수식하는 부정적 어휘의 예로는 '부족한', '어설픈', '엉성한' 등이 포함되어 있었다.

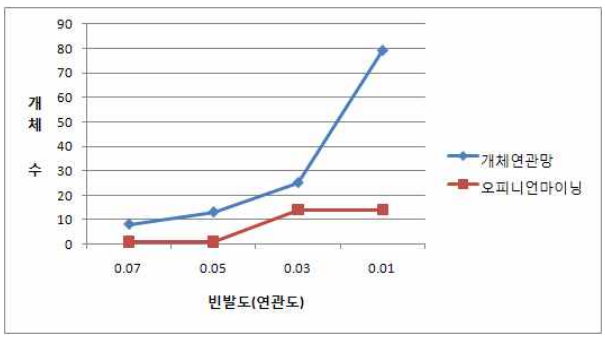
반면, 본 논문에서 제안한 개체연관망 기법에 의한 분석결과는 (그림 8)의 (b)에서 보여주고 있다. 확장된 기법에 의한 분석결과는 감성어휘에 의하여 수식되는 개체들을 포함하고 있을 뿐만 아니라 감성어휘에 의하여 수식되지 않는 개체들 즉, '눈물', '국가대표', '기대', '재미' 등과 같은 개체들도 포함하고 있다. 감성어휘에 의하여 수식되지 않는 개체들이 분석결과에 포함됨으로써 더욱 풍부한 요약정보가 되고 있음을 알 수 있다. '국가대표'라는 개체가 '영화'와 연관되는 것을 볼 때, 고객들이 해운대에 대한 영화평을 하면서 '국가대표'라는 영화와 비교하고 있음을 알 수 있다. 영화평론가 또는 영화제작자는 '해운대'라는 영화에 대하여 '국가대표'라는 영화의 특징 등을 고려하여 좀 더 심도있게 분석할 수 있다. 이러한 정보는 기존의 오피니언마이닝 기법으로는 도출할 수 없는 새로운 요약정보이다. 개체연관망 기법은 감성어휘에 의하여 수식되는 개체들에도 중립도의 개념을 도입함으로써 보다 세밀한 긍정 부정 비율을 제공하고 있다. 예를 들면, '영화' 라는 개체의 경우 기존 오피니언마이닝 기법에 의하면 90%의 긍정비율을 나타내지만, 개체연관망 기법에 의하면 긍정비율이 22%, 부정비율이 3%, 중립비율이 74%가 된다. 이는 기존 기법에서는 중립도를 고려하지 않기 때문에 발생하는 문제로서 의사결정에 커다란 영향을 미칠 수 있다. (그림 8)에서의 개체연관망 기법은 최소빈



(a) 기존 기법 (b) 개체연관망 기법(최소연관도: 0.01)

(그림 8) 기존 기법과의 정성적 비교(최소 빈발도 : 0.03)

발도를 0.03, 최소연관도를 0.01로 설정하였을 때 생성된 개체연관망을 나타내고 있다. 최소빈발도와 최소연관도를 더 높게 설정하면 개체연관망에 포함되는 개체들이 적어져서 보다 단순한 요약정보를 얻을 수 있으며, 더 낮추었을 때는 복잡하지만 보다 풍부한 정보를 제공하는 개체연관망을 도출할 수 있다.



(그림 9) 기존 기법과의 정량적 비교

(그림 9)에서는 기존 오피니언마이닝 기법과 개체연관망 기법에 대하여, 최소빈발도와 최소연관도의 변화에 따라 요약정보에 포함되는 개체들의 수가 어떻게 달라지는지 보여주는 그래프이다. 개체연관망에 포함된 개체수가 기존 오피니언마이닝 기법에 의하여 생성되는 개체보다 더 많음을 알 수 있다. 왜냐하면 개체연관망에는 감성어휘에 의하여 수식되지 않는 개체들도 포함되어 있기 때문이다. 개체들의 수가 많을수록 보다 풍부한 요약정보를 나타낼 수 있지만 불필요한 개체가 포함되어 복잡성을 가중시킬 수도 있으므로 사용자의 기호에 따라 적절한 최소빈발도와 최소연관도를 설정할 필요가 있다.

6. 결 론

웹 2.0 시대에 네티즌들의 경험과 의견 등이 표현된 온라인 고객리뷰나 상품리뷰 등은 매우 많고 더욱 증가하고 있다. 특정 제품이나 서비스에 대한 네티즌의 의견들은 고객뿐만 아니라 기업 입장에서도 마케팅이나 경영전략을 수립하기 위한 중요한 자료가 될 수 있기 때문에 온라인 고객리뷰를 분석하는 것은 매우 중요하다.

본 논문에서는 기존 오피니언마이닝 기법에서 처리할 수 없었던 추가적인 정보를 생성할 수 있는 새로운 기법을 제안하였다. 새로운 기법은 개체연관망 모델을 기반으로 한다. 개체연관망 모델은 기존 오피니언마이닝에서 목표로 했던 개체의 긍정·부정도를 표현할 수 있을 뿐만 아니라 기존 기법에서 처리할 수 없는 개체 사이의 연관성을 나타낼 수 있었다. 또한, 빈발도의 개념을 바탕으로 개체들의 상대적 중요도도 표현할 수 있었다.

개체연관망 모델을 기반으로 시스템을 구현하여 성능을 평가한 결과, 개체연관망에 포함된 개체수가 기존 오피니언마이닝 기법에 의하여 생성되는 개체보다 더 많음을 알 수 있다. 개체들의 수가 많을수록 보다 풍부한 정보를 표현할 수 있으므로 개체연관망이 기존 오피니언마이닝 방법보다 더 많은 정보를 도출할 수 있음을 알 수 있었다.

그러나, 개체연관망에 포함되는 개체들 중에서 유의미하지 않은 개체들이 있을 수 있으며 이들로 인하여 개체연관망 생성시간이 더 길어질 수 있다. 보다 유의미한 개체들만이 개체연관망에 포함되도록 함으로써 개체연관망 생성시간을 단축시키는 연구는 추후 과제로 남겨둔다.

참 고 문 헌

[1] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", KDD'04, 2004, pp.168-177.

[2] Xiaowen Ding, Bing Liu and Philip S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", WSDM'08, 2008, pp.231-239.

[3] 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.

[4] Agrawal, R., Imielinski, T., Swami, A., "Mining association rules between sets of items in large databases", Proc. of ACM SIGMOD, 1993, pp.207-216.

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

[6] 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.

[7] 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.

[8] 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.

[9] Xiaowen Ding, and Bing Lui, "The Utility of Lingusitic Rules in Opinion Mining", SIGR pp.811-812, 2007.

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

[11] Korean Parser Test Version, http://nlp.kookmin.ac.kr/HAM/kor/download. html.

[12] 강승식, 한국어 형태소분석과 정보검색, 홍릉과학출판사, 2003.

김 근 형

e-mail : khkim@jejunu.ac.kr

1990년 서강대 전자계산학과(학사)

1992년 서강대 전자계산학과(석사)

2001년 서강대 컴퓨터학과(박사)

2001년~현 재 제주대학교 경영정보학과 교수

관심분야 : 데이터베이스, 데이터마이닝, 텍스트마이닝