http://dx.doi.org/10.5392/JKCA.2012.12.04.076

관계형 다차원모델에 기반한 온라인 고객리뷰 분석시스템의 설계 및 구현

Study on Designing and Implementing Online Customer Analysis System based on Relational and Multi-dimensional Model

김근형*, 송왕철**
제주대학교 경영정보학과*, 제주대학교 컴퓨터공학과**

Keun-Hyung Kim(khkim@jejunu.ac.kr)*, Wang-Chul Song(kingiron@gmail.com)**

요약

오피니언마이닝 기법은 대량의 고개리뷰들에 나타나는 핵심개체 또는 속성들에 대하여 고객들이 느끼는 긍정 또는 부정의 정도를 계산할 수 있지만, 그 분석능력이 단순하다는 한계가 있다. 본 논문에서는 온라인 고객리뷰들에 대하여 다차원적으로 분석할 수 있는 기법을 제안하였다. 기존의 OLAP기법을 텍스트 데이터형에 적용할 수 있도록 수정하였다. 다차원 분석모델은 명사축과 형용사축, 문서축으로 구성되는 3차원 공간 개념을 4개의 관계형 테이블로 실체화 한 것이다. 다차원 분석모델은 기존의 오피니언마이닝, 정보요약, 클러스터링 알고리즘들을 융합할 수 있는 새로운 틀이라는 점에서 그 가치가 있다. 본 논문에서 제안한 다차원 분석모델과 알고리즘들을 실제로 구현하여 온라인 고객리뷰에 대한 복잡한 분석을 수행할 수 있음을 확인하였다.

■ 중심어 : | 다차원 분석모델 | 오피니언마이닝 | 정보요약 | 클러스터링 | 연관규칙탐사 | 관계형 모델 |

Abstract

Through opinion mining, we can analyze the degree of positive or negative sentiments that customers feel about important entities or attributes in online customer reviews. But, the limit of the opinion mining techniques is to provide only simple functions in analyzing the reviews. In this paper, we proposed novel techniques that can analyze the online customer reviews multi-dimensionally. The novel technique is to modify the existing OLAP techniques so that they can be applied to text data. The novel technique, that is, multi-dimensional analytic model consists of noun, adjective and document axes which are converted into four relational tables in relational database. The multi-dimensional analysis model would be new framework which can converge the existing opinion mining, information summarization and clustering algorithms. In this paper, we implemented the multi-dimensional analysis model and algorithms. we recognized that the system would enable us to analyze the online customer reviews more complexly.

■ keyword : | Multi-dimensional Analysis Model | Opinion Mining | Information Summarization | Clustering | Association Rules Mining | Relational Model |

* 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. 2011-0011446).
접수번호 : #120305-009
접수일자 : 2012년 03월 05일
심사완료일 : 2012년 04월 19일
교신저자 : 김근형, e-mail : khkim@jejunu.ac.kr

I. 서 론

오늘날 인터넷이 활성화되고 전자상거래를 이용하는 사람들이 증가하면서 인터넷을 통한 제품에 대한 경험과 지식의 공유가 활발해졌다. 제품에 대한 경험이나 지식에 대한 의견이라 할 수 있는 고객리뷰(Customer Review)들은 소비자들이 상품을 구매할 때 많은 영향을 미칠 뿐만 아니라 기업들에게도 새로운 마케팅전략을 수립하는데 중요한 자료로 활용될 수 있다. 인터넷의 일상화에 따라 온라인 고객리뷰는 그 양이 계속 증가할 것이며 기업이나 고객의 입장에서도 그 중요성 또한 더욱 커질 것으로 전망된다.

최근에 연구되고 있는 오피니언마이닝(opinion mining)은 웹사이트에 게시되어 있는 온라인 고객리뷰들을 분석 대상으로 하는 텍스트마이닝(Text Mining)의 한 분야로서 고객의견에 대한 긍정(positive)과 부정(negative)의 분포 등을 분석할 수 있다[1-6]. 기존의 연구들은 리뷰의 대상이 되는 제품이나 제품속성에 대응하는 정확한 명사를 찾는 방법이라든가, 긍정/부정을 나타내는 감성단어를 정확하게 분류하는 방법 등과 관련된 것들이었다[4-6][8][12][13]. 덕분에 오피니언마이닝의 분석 정확도는 많이 개선되었지만, 그 분석능력이 단순하기 때문에 그 유용성에 한계가 있다. 예를 들어, 기존의 오피니언마이닝 기법은 디지털카메라의 화질에 대하여 긍정적 의견과 부정적 의견의 분포만을 일회성으로 계산할 수 있다. 디지털 카메라의 화질에 대한 부정적 의견이 70%로 분석되었다고 했을 때, 분석자는 부정적 의견을 나타낸 고객리뷰들의 특징을 추가적으로 분석함으로서 보다 의미있고 가치있는 정보를 도출할 수 있다. 결국 온라인 고객리뷰를 다양한 관점에서 상호대화방식으로(interactively) 분석할 수 있다면 보다 의미있고 가치있는 결과를 도출할 수 있다.

OLAP(OnLine Analytical Processing)은 정형화된 데이터를 대상으로 한 다차원 분석기술로서, 이미 상용화되어 많은 기업들이 의사결정지원을 위하여 유용하게 활용하고 있는 검증된 기술이다. 기존의 OLAP가 정형화된 데이터(structured data)에 적용될 수 있는 분석기법인데 반하여, 본 논문에서는 비정형화된 텍스트 데이터(unstructured data)에 OLAP기술을 적용함으로써 오피니언마이닝 뿐만 아니라 고객리뷰들의 특징추출 및 클러스터링 기능까지 제공할 수 있는 새로운 분석방법론을 제안하고자 한다.

2장에서는 온라인 고객리뷰의 분석과 관련한 선행연구들을 검토하고 한계점이나 문제점을 분석하며, 3장에서는 온라인 고객리뷰의 다차원분석을 위한 새로운 모델을 제안한다. 4장에서는 다차원분석 모델에 기반한 데이터베이스와 알고리즘을 설계하며, 5장에서는 실제 시스템을 구현하여 그 기능을 살펴보며 6장에서 결론을 맺는다.

II. 관련연구 검토 및 분석

[1]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[2], 이는 고객리뷰 데이터와 사용자 평가점수를 활용하여 요약보고서를 생성하는 기법을 제안하였다. [3]에서는 문장구조와 문장 사이의 관계, 문장성분의 패턴정보 등의 언어규칙을 이용한 통계학적 방법으로 오피니언마이닝에 접근하고 있다. [4-6]에서는 워드넷을 활용하여 어휘의 긍정이나 부정적 의미를 판단하고 이를 센티워드넷(SentiwordNet)으로 응용하여 감정의 폭을 정량화하는 방법을 제시하고 있다.

[7]은 오피니언마이닝 과정에서 데이터마이닝의 연관규칙탐사기법을 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.

[8]에서는 이모티콘을 활용하여 텍스트의 긍정적 또는 부정적 감정을 인식하도록 하였는데, 이모티콘 기반의 감정분류 성능은 70~80% 사이의 정확도를 보였다.

SNS(Social Networks Services)에는 매 순간 엄청난 수의 사용자가 이용하기 때문에 긍정/부정 오피니언의 변화가 지속적으로 일어날 수 있으며, 이와 관련하여 최신의 데이터에 기반하여 효율적으로 분석결과를 업

데이트하는 방법이 제안되었다[9]. [10]에서는 제품속성단어와 제품속성을 수식하는 감성단어 사이의 의존관계를 통하여 핵심 감성단어를 자동 추출하는 방법을 제안하고 있으며, [11]에서는 제품속성단어와 제품속성을 수식하는 감성단어 사이의 의존관계에 HITS (Hyperlink-induced topic search) 알고리즘을 적용하여 제품속성의 랭킹을 결정하는 기법을 제안하고 있다. 이러한 기존의 오피니언마이닝 기법들은 온라인 고객리뷰들에 대한 요약보고서를 일회성으로 생성하기 때문에 분석자와의 상호대화방식에 의한 다양하고 깊이 있는 분석을 제공할 수 없다는 한계가 있다.

온라인 고객리뷰들도 결국은 사용자들이 작성해 놓은 문서들의 모음이라고 할 수 있는데, 기존의 문서요약(summarization) 기술은 2가지 유형으로 나누어진다. 하나는 원형틀(template, 원형판)을 채워 넣는 방식이고, 다른 하나는 핵심문장을 추출하는 방식이다[12][13]. 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 추출함으로써 문서내용을 간략화 한다. 핵심문장 추출방식은 길이가 긴 단일문서의 간략화를 목적으로 하기 때문에 길이가 짧지만 대량의 문서로 구성된 온라인 고객리뷰를 분석하기 위한 방법으로는 적합하지 않다. 본 논문에서는 온라인 고객리뷰라는 짧지만 많은 수로 구성된 문서들에 존재하는 핵심명사들을 추출하고 그 핵심명사들 사이의 연관도를 표현하는 방식으로 다중문서의 요약정보를 도출하고자 한다.

온라인 고객리뷰를 분석할 때 유사한 고객리뷰끼리 분류하여 클러스터링(clustering)함으로써 각 클러스터들에 대한 추가적인 분석의 발판을 마련할 수 있다. 클러스터링은 분류키워드 또는 분류주제가 알려지지 않은 상태에서 문서들을 분석하여 유사한 내용의 문서들을 묶은 후에, 필요에 따라 분류된 문서들을 대표하는 키워드들이나 주제를 추출한다. 수집된 문서들이 매우 많을 때, 클러스터링 후에 생성된 클러스터들은 수백 또는 수천 개가 되어 데이터분석자에게 또 다른 부담을 줄 수 있다[14-16]. 따라서 대부분의 클러스터링 알고리즘들은 생성되는 클러스터들의 최대 숫자나 크기, 클러스터에 속하는 문서들 사이의 유사도 등을 입력으로 받아서 보다 융통성 있게 클러스터들을 생성할 수 있다.

클러스터링 분석은 다른 분석 작업을 위한 준비과정으로서의 역할이 크며 일반적으로 클러스터링 후에 추가적인 분석작업이 뒤따른다. 이는 클러스터링 분석이 다른 분석기법과 결합될 때 그 유용성이 커질 수 있다는 의미가 된다.

본 논문에서는 분석자와 상호대화 방식으로 클러스터링과 문서요약, 오피니언마이닝을 연계시킬 수 있는 다차원분석 모델과 알고리즘을 개발하고자 한다.

Ⅲ. 다차원 분석모델

온라인 고객리뷰의 다차원분석을 위한 다차원 분석모델은 3차원 축을 기반으로 하며, 이는 성형스키마(star schema) 형태의 관계형 테이블들로 구체화시킬 수 있다.

1. 3차원 축

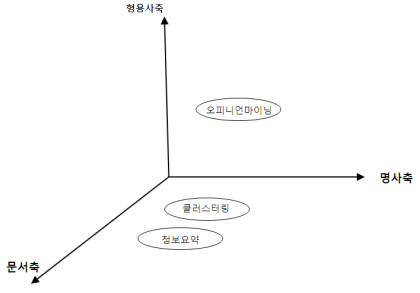


그림 1. 3차원 공간에서 가능한 분석 유형

위의 [그림 1]과 같이 3개의 축(형용사축, 명사축, 문서축)으로 구성된 3차원공간에서 다양한 분석을 수행할 수 있다. 형용사축과 명사축의 2차원 공간에서는 오피니언마이닝 분석이 가능하며, 명사축과 문서축을 중심으로 하여 클러스터링이나 정보요약을 수행할 수 있다. 또한, 형용사축, 명사축, 문서축으로 이루어지는 3차

원 공간에서는 다양한 상호대화분석이 가능하다. 예를 들면, 특정 클러스터의 요약정보와 오피니언마이닝 등의 결과를 만들어낼 수 있다.

2. 성형스키마

3차원 축은 관계형 OLAP 모델인 성형스키마 형태를 갖는 테이블들의 모임으로 변환될 수 있다. [그림 2]는 [그림 1]의 3차원 축과 대응되는 성형스키마를 나타내고 있다. 차원테이블들과 사실테이블사이의 다양한 SQL 조인(join) 처리를 통하여 다양한 다차원 분석을 수행할 수 있다.

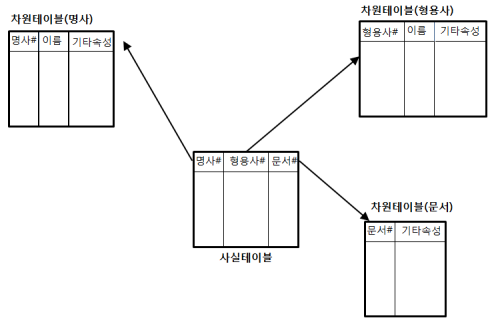


그림 2. 성형스키마 구조

Ⅳ. 다차원 분석시스템설계

[그림 3]은 온라인 고객리뷰를 다차원적으로 분석하는 과정을 나타내고 있다. 웹사이트 상에 게시된 온라인 고객리뷰들은 비정형 데이터형인 텍스트문서 형태로 고객리뷰 데이터베이스에 저장되어 있으며 보다 효율적인 처리를 위하여 관계형 파일(relational file)로 변환될 필요가 있다. 관계형 파일로 변환하기 위하여 텍스트문서내의 각 문장들은 구문분석기에 의하여 각 단어들에 품사가 부여된 형태의 구문구조트리로 변환된다. 구문구조트리 파일로부터 명사와 형용사에 해당하는 단어들이 추출되어 테이블 형태의 관계형 파일들에 저장된다. 관계형 파일들은 관계형 기반의 다차원 모델로 구성되며 분석알고리즘을 통하여 다양한 분석을 수행할 수 있다.

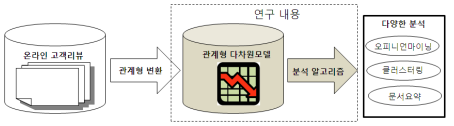


그림 3. 시스템 개요

1. DB설계

[그림 4]는 다차원분석을 위한 데이터베이스 스키마(schema)를 나타내고 있다. F(사실테이블)의 데이터는 고객리뷰의 구문분석 결과인 구문구조트리로부터 감성형용사와 그 수식되는 명사들이 추출되면서 삽입된다. F테이블의 기본키는 (문서번호, 형용사번호, 명사번호)로 이루어진 합성속성이며, 따라서 하나의 고객리뷰 문서는 여러 개의 레코드들에 대응된다. A(형용사테이블)은 형용사에 대한 정보를 포함하는데, 구문구조트리로부터 추출된다. A의 명암도 속성은 해당 형용사의 긍정/부정 정도를 -1과 1사이의 값으로 나타낸 것이며(-1에 가까울수록 부정적인 감성표현, 1에 가까울수록 긍정적인 감성표현), 감성형용사가 아닌 경우의 명암도 속성값은 0이 된다. 명암도나 동의어 속성값은 온톨로지(ontology)나 수작업으로 설정된다.

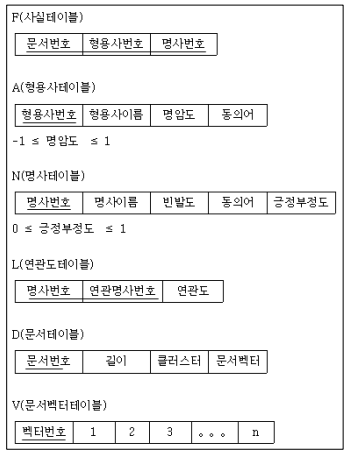


그림 4. 데이터베이스 스키마

N(명사테이블)은 고객리뷰들에 출현하는 명사들에 대한 정보를 포함하는데, 역시 구문구조트리로부터 추출된다. N의 빈발도 속성은 해당 명사가 고객리뷰에서 얼마나 자주 출현하는지를 나타내는 척도로서, 전체 고

객리뷰 수에 대하여 해당명사가 나타나는 고객리뷰 수의 비율로 표현된다. 긍정부정도는 해당명사가 감성형용사들에 의하여 얼마나 긍정적으로 또는 부정적으로 수식되고 있는지를 나타내는 척도로서, 그 긍정 또는 부정의 정도를 -1과 1사이의 값으로 나타낸 것이다. -1에 가까울수록 부정적이며 1에 가까울수록 긍정적이다.

L(연관도테이블)은 고객리뷰들에 자추 출현하는 핵심명사들 사이의 연관도 정보를 포함한다. L의 기본키는 (명사번호, 연관명사번호)로 이루어진 합성속성이다. 연관도 속성은 해당명사가 '연관명사'와 함께 얼마나 자주 동일 고객리뷰에 출현하면서 서로 연관되어 있는지를 알려주는 척도이다.

D(문서테이블)은 각 고객리뷰 문서에 대한 정보를 포함한다. D의 클러스터 속성은 해당 고객리뷰 문서가 속하는 클러스터를 의미한다. 문서벡터 속성은 명사테이블 N에 나타나는 명사들을 축으로 한 벡터공간(vector space)에서 해당문서가 어느 위치에 대응되는지를 나타내는 값이다. V의 벡터번호 속성은 D의 문서벡터 속성과 대응된다.

2. 분석 알고리즘 설계

[그림 4]의 각 테이블들에 대하여 또는 사실테이블과 차원테이블 사이의 조인(join)과정을 통하여 다양한 분석 알고리즘이 적용된다. [표 1]에서는 적용 가능한 분석 알고리즘들을 개략적으로 설명하고 있다.

표 1. 분석알고리즘

알고리즘 유형	내 용	사용 테이블
오피니언 마이닝	온라인 고객리뷰에 나타나는 핵심개체들에 대한 명암도를 계산한다. 출현빈도 많은 명사들에 대하여 해당명사를 수식하는 감성형용사들의 명암도 합계에 의하여 계산된다.	N, A, F
정보요약	고객리뷰 문서들에서 빈발하게 출현하는 핵심명사들 사이의 연관도를 추출한다. 최소빈발도와 최소연관도는 분석자가 설정할 수 있다.	N, F, L
클러스터링	IDF(Inverse Document Frequency)방법[16]을 이용한다. IDF 방법은 특정문서 d에 속하는 특정단어 w의 가중치를 계산할 때 w가 d에서 나타난 빈도수 f_w^d 을 f_w^t와 곱한다. f_w^t은 w가 모든 문서에 적용되어 계산되는 것으로서, w가 모든 문서에 골고루 나타난다면 f_w^t값은 작아지고 그렇지 않으면 f_w^t값은 커진다.	D, N, F, V

[그림 5]에서는 오피니언마이닝 알고리즘을 구체적으로 보여주고 있다. 명사테이블 N의 빈발도 속성과 형용사테이블 A의 명암도 속성 그리고 사실테이블 F(형용사와 명사 사이의 수식관계 나타남)를 입력으로 하여 자주 출현하는 k개의 핵심명사들에 대한 긍정부정도를 계산한다.

오피니언마이닝 알고리즘

입력: 명사테이블 N(빈발도),
형용사테이블 A(명암도),
사실테이블 F
(명사와 형용사의 수식관계)
출력: 명사테이블 N(긍정부정도)

핵심명사에 대한 긍정부정도 계산

BEGIN
1. N으로부터 최소빈발도를 만족하는
명사 k개 선택
2. for *명사i* ($1 \le i \le k$)
3. *명사i* 의 긍정부정도 =

$$\frac{\sum_{j} \text{명사}i\text{를 수식하는 형용사}j\text{의명암도}}{F\text{에서 명사}i\text{의 출현횟수}};$$

do
END

그림 5. 오피니언마이닝 알고리즘

[그림 6]에서는 정보요약 알고리즘을 구체적으로 나타내고 있다. 빈발하게 출현하는 핵심명사들 사이의 연관도를 표현함으로써 고객리뷰의 전체적인 윤곽을 나타내는 방법이다. 최소빈발도 이상 출현하는 임의의 명사i와 명사j에 대하여(2행과 4행), 명사i가 출현하는 고객리뷰들중에서 명사i와 명사j가 동시에 출현하는 고객리뷰들이 어떤 비율로 나타나는지를 표현한 것이 명사i와 명사j의 연관도가 된다(5행).

[그림 7]에서는 클러스터링 알고리즘을 구체적으로 나타내고 있다. 1행에서 5행에서는 명사테이블 N에 나타나는 각 명사들을 축으로 하는 벡터공간에 각 고객리뷰 문서들을 위치시키고 있다. 각 고객리뷰 문서에 대응하는 벡터값은 IDF 기법을 이용하여 계산된다(4행).

n차원 벡터공간에 표현된 각 고객리뷰 문서들은 처음에 임의로 선택된 k개의 벡터값(6행)을 기준으로 클러스터링된다(7행). 이후 각 클러스터의 중심 벡터값이 계산된 후 새로운 중심벡터값을 기준으로 각 고객리뷰 문서들을 재클러스터링된다. 클러스터의 중심벡터값이 변하지 않을 때까지 이러한 과정은 반복된다(7행-9행).

정보요약 알고리즘

입력: 최소빈발도, 최소연관도,
명사테이블 N,
사실테이블 F(문서가 포함하는 명사들)
출력: 연관도테이블 L(연관명사와 연관도)

핵심명사들 사이의 연관성 계산

BEGIN
1.for each *명사i* (1≤i ≤ N의 레코드수)
2. if (빈발도(*명사i*) ≥ 최소빈발도)
3. for each *명사j*
(i+1≤j ≤ N의 레코드수)
4. if (빈발도(*명사j*) ≥ 최소빈발도)
5. 연관도(*명사i, 명사j*) =
$$\frac{\text{명사}i\text{와 명사}j\text{를 동시에 포함하는 문서수}}{\text{명사}i\text{를 포함하는 문서수}}$$
end if
6. if (연관도(*명사i, 명사j*)
≥ 최소연관도)
7. *명사i*의연관도=연관도(*명사i,명사i*);
8. *명사i*의 연관명사 = *명사j*;
else
9. *명사i*의연관도=0;
10. *명사i*의 연관명사 = null;
end if
do
end if
do
END

그림 6. 정보요약 알고리즘

클러스터링 알고리즘

입력: 문서테이블 D
명사테이블 N
(명사축에 의한 벡터공간 생성)
사실테이블 F
(문서가 포함하는 명사들)
출력: 문서테이블 N(클러스터)
문서벡터테이블 V

유사한 고객리뷰 문서들끼리 군집화 함

BEGIN
1.for each *문서i* (1≤i ≤ D의레코드수)
2. for each *명사j*((1≤j≤N의 레코드수)
3. if (*명사j*가 *문서i*에 포함)
4. V_{ij} = (*문서i*에서*명사j*의출현횟수)×
$$\log\frac{D\text{의 레코드수}}{\text{명사}j\text{를 포함하는 문서수}}\ ;$$
else
5. V_{ij} = 0;
end if
do
do
6. 임의의 k개의 문서벡터 V_i, V_{i+1},..., V_{i+k} 선택하여 각각을 k개의 클러스터
C_1, C_2,,,,C_k의 중심벡터값으로 설정
7. for each *문서i* (1≤i ≤ D의레코드수)
*문서i*의 벡터값과 C*j* (1≤j≤k)의
중심벡터값 비교하여 가장 가까운
클러스터 선택
do
8. k개의 클러스터 C_1, C_2,,,,C_k의
중심벡터값을 업데이트
9. if (k개 클러스터의 중심값이 변함)
go to 7;
else
D의 클러스터 속성값 지정;
end if
END

그림 7. 클러스터링 알고리즘

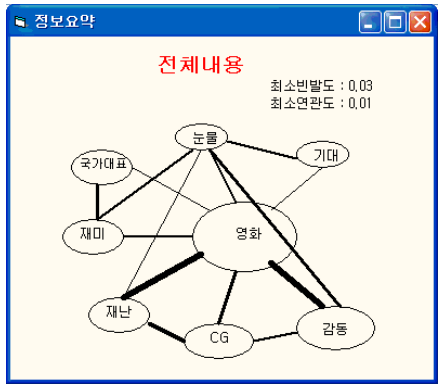
V. 시스템 구현 및 평가

본 논문에서는 4장에서 제안한 데이터베이스 스키마와 알고리즘들을 바탕으로 다차원모델 기반의 온라인 고객리뷰 분석시스템을 개발하였다. 실험용 데이터는 네이버랩(lab.naver.com)에서 제공하는 영화 "해운대" 40자평 데이터세트를 사용하였다. 여기에는 약 1만개의 고객리뷰가 포함된다. 한국어 구문분석을 위하여 구문분석기는 KLT[17][18]을 사용하였고, DBMS는 SQL Server 2008, 프로그래밍언어는 SQL Server 2008의 T-SQL과 비주얼베이직을 사용하였다. 개발된 시스템은 CPU 1.7GHz, 메모리 8GB, 윈도XP가 탑재된 환경에서 실행되었다.

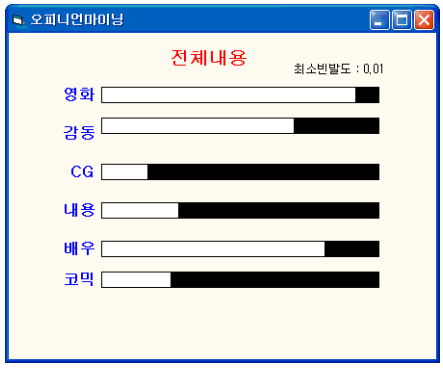
정보검색분야에서 중요한 성능평가 기준으로 사용되는 척도는 정확도(precision)과 재현도(recall)이지만, 기존 오피니언마이닝 기법과 다차원분석기법의 정확도와 재현도는 동일할 것으로 판단된다. 왜냐하면, 다차원분석기법에서도 오피니언마이닝을 위한 알고리즘으로 기존 기법[6]을 사용하고 있기 때문이다.

따라서, 본 논문에서는 정성적 관점에서 기존기법과 다차원분석기법을 비교하고자 한다. 정성적 관점의 분석을 통하여 각 기법으로부터 도출된 정보들이 어떻게 다른지, 기존기법에 비하여 다차원분석 기법이 얼마나 어떻게 더 구체적으로 정보분석을 가능하게 하는지 살펴보고자 한다.

[그림 8]은 기존기법[6][19]를 이용하여 고객리뷰를 분석한 결과를 나타내고 있다. [그림 8]의 (a)에서는 최소빈발도를 0.03, 최소연관도를 0.01로 하였을 때, 전체 고객리뷰에 대한 요약결과를 보여주고 있다. 고객리뷰 내의 중요한 명사들과 명사들 사이의 연관도를 나타내는 방식으로 정보요약을 표현하고 있다. 각 명사들은 타원형으로 표현되고 타원형의 크기는 해당명사의 빈발도를 의미하며, 타원사이에 연결된 직선의 두께는 두 명사 사이의 연관정도를 나타낸다. 직선의 두께가 넓을수록 더 큰 연관도를 갖는다. 정보요약에서 고객들은 '감동'이라는 단어와 'CG(컴퓨터그래픽)', '국가대표', '재난' 등의 단어를 주로 많이 사용하는 것으로 나타났다. 오피니언마이닝 결과를 볼 때, 영화에 대해서는 호평하고 있으나 'CG'(컴퓨터그래픽)과 '내용'에 대해서는 불만족하는 고객들이 많이 있음을 알 수 있다. 그러나 내용에 대한 불만족이 컴퓨터그래픽이 부실해서인지, 이야기 내용전개에 대한 미숙함 때문인지 명확하게 판단하기가 어렵다.



(a) 정보요약결과

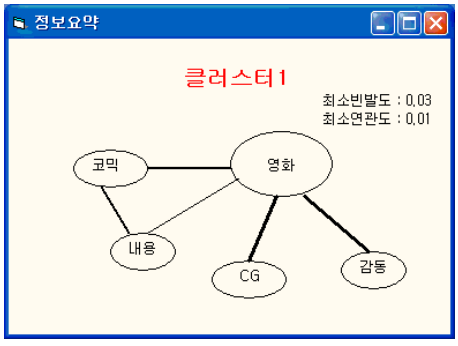


(b) 오피니언마이닝 결과

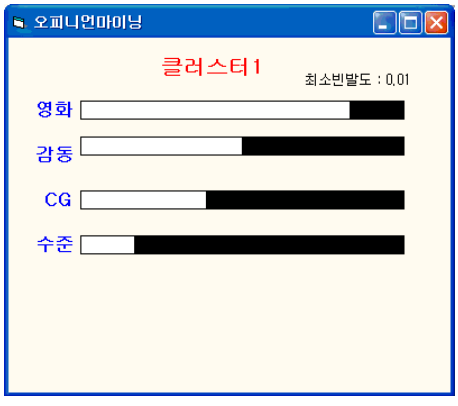
그림 8. 기존기법을 통하여 도출된 정보유형

[그림 9]와 [그림 10]은 다차원모델 기법을 이용하여 고객리뷰를 분석한 결과를 나타내고 있다. 먼저 클러스터링을 통하여 고객그룹을 나눈 후 각 클러스터에 대한 정보요약과 오피니언마이닝 결과를 나타내고 있다. [그림 9]는 클러스터1에 대한 분석결과이고, [그림 10]은 클러스터5에 대한 결과를 나타내고 있다. 클러스터1에 속하는 고객들은 'CG'와 '감동' 등의 단어를 주로 사용하고 있으며, '감동'에 대한 품평은 보통이지만, 'CG'에 대한 평가는 좋지 않음을 알 수 있다. 반면, 클러스터5에 속한 고객들은 '감동'이라는 단어와, '국가대표', '내

용'이라는 단어를 주로 사용하고 있으며, 오피니언마이닝 결과로서는 '감동'과 '내용'에 대한 평가가 좋지 않음을 알 수 있다. 클러스터5에 속한 고객들은, '국가대표'라는 영화와 비교함으로써, '감동' 부분과 '내용' 부분에 대하여 부정적으로 평가하는 것으로 예측해 볼 수 있다. 즉, CG(컴퓨터그래픽)의 부실함 때문에 영화 내용에 대하여 부정적으로 평가한다기보다는, '국가대표' 라는 영화와 비교함으로써 상대적으로 '내용'에 대하여 부정적인 평가를 내릴 개연성이 있다. 결국, '해운대'라는 영화는 CG의 부실함보다는 '국가대표'에 비하여 이야기 전개방식의 미숙함 때문에 '내용'에 대한 평가가 부정적임을 알 수 있다. 이러한 분석결과는 기존의 방법[6][19]으로는 도출하기 어렵다는 측면에서 다차원모델의 효과성을 가름해 볼 수 있는 부분이다.

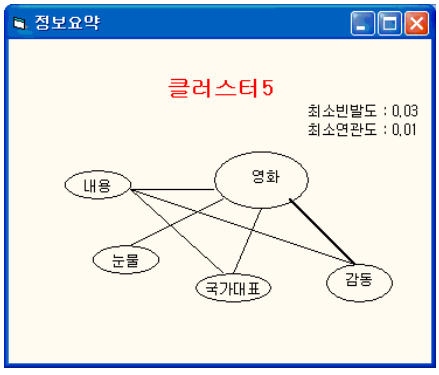


(a) 정보요약결과

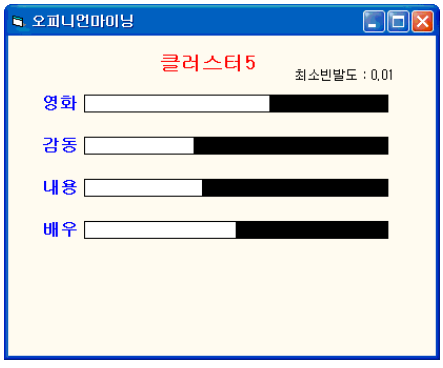


(b) 오피니언마이닝 결과

그림 9. 다차원모델 기법을 통하여 도출된 정보유형1



(a) 정보요약결과



(b) 오피니언마이닝 결과

그림 10. 다차원모델 기법을 통하여 도출된 정보유형2

Ⅵ. 결 론

인터넷 상에서의 참여와 공유, 확대 재생산을 표방하는 웹 2.0시대를 맞이하여 인터넷 사용자들이 생성하는 온라인 고객리뷰들인 비정형 데이터(unstructured data)는 매우 많고 또한 계속 증가하고 있다. SNS가 급성장하면서 온라인고객리뷰들은 더욱 증가될 것이라는 측면에서 비정형데이터의 분석기술에 대한 수요는 더욱 커질 것이다. 본 논문에서는 기존 오피니언마이닝 기법과 클러스터링기법, 연관규칙탐사 기법 등을 융합함으로써 보다 차원높은 고객리뷰 분석이 가능한 다차원 분석모델을 제안하였다. 다차원 분석모델은 기존의 정형데이터에 적용되었던 OLAP 기법을 텍스트 데이터형에 적용할 수 있도록 수정한 것이다. 관계형 기반의 다차원 분석모델은 명사축과 형용사축, 문서축으로 이루어

진 3차원 공간 개념을 사실테이블, 명사테이블, 형용사 테이블, 문서테이블과 같은 관계형 테이블들로 실체화 한 것이다. 이러한 테이블들에 오피니언마이닝 알고리즘과 연관규칙탐사 알고리즘, 클러스터링 알고리즘 등을 적용하여 다양한 분석 기능을 제공할 수 있었다.

다차원 분석모델은 기존의 다양한 기법들을 융합하기 위한 토대임을 확인하였다는 점에서 그 의의가 있지만, 본 논문에서 적용했던 알고리즘들은 단순한 것이라는 점에서 그 한계가 있다. 추후, 온라인 고객리뷰에 대하여 다차원 기반의 상호대화적인 복잡한 분석이 가능한 알고리즘들에 대하여 지속적인 연구를 수행할 것이다.

참 고 문 헌

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

[2] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red Opal: Product-Feature Scoring from Reviews," Proc. of the 8th ACM conference on Electronic commerce, pp.11-15, 2007.

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

[4] X. W. Ding, "A Holistic Lexicon-Based Approach to Opinion Mining," Proc. of the international conference on web search and web mining, pp.231-240, 2008.

[5] E. Courses and T. Surveys, "Using SentiWordNet for multilingual sentiment analysis," Data Engineering Workshop ICDEW, pp.102-110, 2008.

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

[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, pp.270-274, 2009.

[8] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," In Proceedings of the European Language Resources Association, pp.1321-1326, 2010.

[9] Ismael S. Silva, JanainaGomide, Adriano Veloso, Wagner Meira Jr., and Renato Ferreira, "Effective Sentiment Stream Analysis with Self-Augmenting Training and Demand-Driven Projection," SIGIR, pp.475-484, 2011.

[10] Guang Giu, Bing Liu, J. J. Bu and Chun Chen, "Expanding DomainSentiment Lexicon through Double Propagation," Proc. of 21th IJCAI-09, pp.1199-1204, 2009.

[11] Lei Zhang, Bing Liu, S. H. Lim, and Eamonn O'Brien-Strain, "Extracting and Ranking Product Features in Opinion Documents," Proceedings of the 23rd International Conference on Computational Linguistics, pp.1462-1470, 2010

[12] G. Salton, A. Singhal, C. Buckley, and M. Mitra, "Automatic Text Decomposition using Text Segments and Text Themes," ACM Conference on Hypertext, pp.1-13, 1995.

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

[14] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol.34, No.1, pp.1-47, 2002.

[15] K. R. Larsen and D. E. Monarchi, "A Mathematical Approach to Categorization and Labeling of Qualitative Data: The Latent Categorization Method," Sociological Methodology, Vol.34, No.1, pp.349-392, 2004.

[16] Manu Konchady, Text Mining Application Programming, Thomson Charles River Media, 2006.
[17] http://nlp. kookmin.ac.kr /HAM/kor/download.html
[18] 강승식, *한국어 형태소분석과 정보검색*, 홍릉과학출판사, 2003.
[19] 김근형, "종속성 네트워크 기반의 온라인 고객리뷰 분석시스템의 설계 및 구현", 한국콘텐츠학회논문지, 제10권, 제11호, pp.30-37, 2010.

저 자 소 개

김 근 형(Keun-Hyung Kim) 정회원

- 1990년 2월 : 서강대학교전산학과(공학사)
- 1992년 2월 : 서강대학교 전산학과(공학석사)
- 2001년 2월 : 서강대학교컴퓨터학과(공학박사)
- 2001년 9월 ~ 현재 : 제주대학교 경영정보학과 교수

<관심분야> : 오피니언마이닝, 데이터마이닝

송 왕 철(Wang-Chul Song) 종신회원

- 1989년 2월 : 연세대학교전자공학과(공학사)
- 1991년 2월 : 연세대학교 전자공학과(공학석사)
- 1995년 2월 : 연세대학교전자공학과(공학박사)
- 1996년 ~ 현재 : 제주대학교 컴퓨터공학과 교수

<관심분야> : 망관리, 모바일 애드혹 네트워크 등