온라인 고객리뷰 분석을 통한 시장세분화에 텍스트마이닝 기술을 적용하기 위한 방법론

Methodology for Applying Text Mining Techniques to Analyzing Online Customer Reviews for Market Segmentation

김근형*, 오성열**
제주대학교 경영정보학과*, 제주산업정보대학 세무회계과**

Keun-Hyung Kim(khkim@cheju.ac.kr)*, Sung-Ryoel Oh(jejuoh99@daum.net)**

요약

본 논문에서는 텍스트마이닝 기술을 이용하여 온라인 고객리뷰를 분석하기 위한 방법론을 제안하였다. 온라인 고객리뷰를 보다 효율적이고 효과적으로 분석할 수 있도록 시장세분화의 개념을 도입하였다. 즉, 제안한 방법론은 텍스트마이닝 분야에서 시장세분화의 개념에 부응하는 기술들이라 할 수 있는 범주화와 정보추출 기법의 사용을 포함한다. 특히, 통계적으로 보다 견고한 분석결과를 도출할 수 있도록 전통적 통계분석기법중의 하나인 교차분석방법을 제안하는 방법론에 포함하였다. 제안한 방법론의 타당성을 확인하기 위하여 양질의 온라인 고객리뷰가 있는 웹사이트를 선정하여 실제로 온라인 고객리뷰들을 분석하여 보았다.

■ 중심어 : | 방법론 | 온라인고객리뷰 | 텍스트마이닝 | 교차분석 | 시장세분화 |

Abstract

In this paper, we proposed the methodology for analyzing online customer reviews by using text mining technologies. We introduced marketing segmentation into the methodology because it would be efficient and effective to analyze the online customers by grouping them into similar online customers that might include similar opinions and experiences of the customers. That is, the methodology uses categorization and information extraction functions among text mining technologies, matched up with the concept of market segmentation. In particular, the methodology also uses cross-tabulations analysis function which is a kind of traditional statistics analysis functions to derive rigorous results of the analysis. In order to confirm the validity of the methodology, we actually analyzed online customer reviews related with tourism by using the methodology.

■ keyword: | Methodology | Online Customer Reviews | Text Mining | Cross-Tabulation Analysis | Market Segmentation |

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

접수번호 : #090421-002
접수일자 : 2009년 04월 21일
심사완료일 : 2009년 05월 08일
교신저자 : 김근형, e-mail : khkim@cheju.ac.kr

Review)들은 소비자들이 상품을 구매할 때 많은 영향을 미칠 뿐만 아니라 기업들에게도 새로운 마케팅전략을 수립하는데 중요한 자료로써 활용될 수 있다. 과거에도 기업들이 고객리뷰의 중요성을 인식하고 이를 획득하여 마케팅전략수립에 활용하려 하였으나[1] 고객리뷰 획득의 어려움으로 인하여 연구의 한계를 보여 왔다. 그러나 인터넷의 발달과 더불어 보다 쉽고 적은 비용으로 온라인 고객리뷰(online customer reviews) 획득이 가능해졌다. 웹 2.0 시대에는 인터넷 사용자가 직접 콘텐츠를 제작하여 서로 공유하는 온라인 커뮤니티(community)가 더욱 활성화될 것이고 이로 인한 온라인 고객리뷰들은 더욱 증가할 것이다. 그럼에도 불구하고 온라인 고객리뷰 등의 콘텐츠를 작성하거나 공유하는 방법에는 많은 관심을 기울이고 있지만 이들을 분석하여 좋은 정보나 지식을 생산하기 위한 방법에 대한 연구는 그에 미치지 못하고 있다.

온라인 고객리뷰들은 그 양이 방대하고 특히 비구조화된(unstructured) 텍스트 데이터 형태로 존재하므로 그 데이터의 분석이 쉽지 않다. 전통적인 통계분석방법이나 데이터마이닝 등 기존의 데이터 분석방법들은 분석대상 데이터가 관계형 테이블(relational table)과 같은 구조화(structured)된 형태를 가정하고 있으므로 온라인 고객리뷰와 같은 비구조화된 데이터에는 적합하지 않다. 이러한 어려움 때문에 최근의 어떤 연구에서는 소량의 온라인 고객리뷰를 수작업 형태로 분석한 경우도 있었다[2]. 그러나 소량의 온라인 고객리뷰만을 분석하는 것은 그 결과의 의미가 퇴색되고 대량의 고객리뷰를 분석할 수 있어야 그 분석결과에 대한 효과성이 있을 것이다. 비구조화된 대량의 문서들을 분석하기 위한 정보기술 중의 하나가 바로 텍스트마이닝(text mining)이다. 즉, 텍스트마이닝은 대량의 고객리뷰를 분석하기에 적합한 정보기술인 것이다.

요즘 각광받고 있는 텍스트마이닝 기술은 비구조화된 대량의 텍스트데이터를 분석할 수 있는 정보기술로써 정보추출(Information Extraction), 문서 분류 및 범주화(Text Classification and Clustering) 등 그 연구분야가 다양하다. 텍스트마이닝과 관련한 지금까지의 연구들은 기술을 중심으로 한 연구들이 주를 이루었고[3-15], 텍스트마이닝 기술을 체계적으로 응용하기 위한 연구들[16][17]은 상대적으로 적었다. 텍스트마이닝 기술을 응용하는 몇몇 연구들 중에서는 텍스트마이닝의 특정 분야만을 이용하는 경우가 대부분이었고 텍스트마이닝의 다양한 분야를 조합하여 시너지를 창출하고자 하는 연구들은 드물었다. 특히, 텍스트마이닝 기술을 전통적인 통계분석방법과 결합하여 그 분석결과의 타당성을 확고히 할 뿐만 아니라 그 응용의 가치를 배가시키는 연구는 거의 없는 실정이다.

한편, 고객리뷰는 소비자의 생각을 나타내는 데이터이므로 비슷한 생각을 하는 소비자들을 묶어서 분석을 하면 그 분석결과의 효과성을 높일 수 있다. 시장세분화(market segmentation)은 비슷한 특성을 갖는 소비자들을 그룹화하여 그 요구사항이나 니즈를 파악하고 목표 그룹에 맞는 마케팅 전략을 수립 및 실행하는 경영기법이다. 텍스트마이닝 분야 중 분류 및 범주화 기술에 의하여 세분시장들을 생성해 낼 수 있으며 정보추출 기술에 의하여 각 세그먼트들의 특징을 뽑아낼 수 있다. 텍스트마이닝 기술을 시장세분화에 응용하면 보다 효과적인 데이터 분석 결과를 얻을 수 있다.

본 논문에서는 텍스트마이닝 기술과 전통적인 통계분석방법 그리고 시장세분화 개념을 적용하여 온라인 고객리뷰를 분석하는 방법론을 제안하고자 한다. 온라인 고객리뷰를 분석하기 위하여 텍스트마이닝의 문서분류 및 범주화 기술과 정보추출 기술을 어떻게 조합하여 응용하면 될 것인지, 특히 전통적인 통계분석 방법 중의 하나인 교차분석방법을 어떻게 텍스트마이닝 기술과 접목시킬 수 있는지 그 방법론을 제안할 것이다. 또한, 제안한 방법론을 이용하여 관광분야의 고객리뷰를 분석하고 가설검증을 하는 응용사례를 제시하고자 한다. 본 논문은 비즈니스분야 뿐만 아니라 사회과학분야의 연구에서 텍스트마이닝 기술이 어떻게 응용될 수 있는지 그 새로운 응용방안을 제시하는데 의의가 있다.

II. 이론적 배경

1. 온라인 고객리뷰와 시장세분화

고객리뷰는 고객이 특정제품 및 서비스에 대하여 가지고 있는 지식으로서 사용경험 등이 이에 해당된다[2]. 고객리뷰는 마케팅 분야에서 주로 연구되어온 구전(word of mouth)의 한 종류로서 구전은 소비자들의 구매의사결정에 중요한 역할을 한다는 것이 오래전부터 밝혀졌다[18]. 온라인 고객리뷰는 인터넷상에서 구전을 하기 위하여 소비자들이 직접 작성한 인터넷 문서로서, 이미 문서화가 되어 있으므로 연구자가 개입하지 않고서도 현상을 있는 그대로 관찰 및 연구할 수 있는 좋은 조건을 제시하고 있다.

온라인 고객리뷰를 작성하는 소비자들은 다양하지만 유사한 주제 또는 소비 선호도를 나타내는 비슷한 소비자 그룹들이 존재할 수 있다. 온라인 고객리뷰를 분석할 때 특정 기준을 중심으로 유사한 고객리뷰들을 구분하여 분석하면 조직의 역량에 적합한 분석결과를 얻을 확률이 높다. 즉, 시장세분화(Market Segmentation)의 개념을 도입하여 고객리뷰를 분석하는 것이 보다 효과적이고 유의미한 분석결과를 얻을 수 있을 것이다.

시장세분화의 출발은 개인 소비자들의 요구사항이 다양한 반면 이러한 개인들의 요구들을 일일이 만족시킬 수 있는 상품이나 서비스를 제공하는 것은 불가능함을 인식하는 것이다. Smith[19]는 시장세분화에 대하여 이질적인 시장(heterogeneous market)을 여러 개의 작은 동질적인 시장(homogeneous market)으로 구분하는 것이라고 정의하면서 그 당시의 새로운 경영전략으로 소개하였다. 시장세분화를 위해서는 유사한 개인특성(personal characteristics)을 갖는 소비자 그룹들이 발굴되어야 한다. 이러한 개인특성들을 세분화기준(segmentation criteria)이라고 하는데, 인구통계적 변수(socio-demographical variable, 예를 들면 청년 관광객과 노년 관광객), 행동변수(behavioral variables, 예를 들면 스키관광객과 휴양관광객), 심리통계적 변수(psychographic variables, 예를 들면 관광동기) 등이 세분화기준이 될 수 있다.

시장세분화의 장점은 상품이나 서비스를 특정 소비자그룹의 니즈(needs)에 맞춤으로서 (1)전체 소비자를 대상으로 한 경쟁으로부터 특정 소비자만을 대상으로 한 경쟁으로 그 경쟁강도가 약화될 수 있고 (2)전체 소비자가 요구하는 상품이나 서비스를 제공하기 위한 비용에서 특정 소비자만을 위한 상품이나 서비스를 제공하는 비용으로 절감될 수 있으며 (3)세분된 시장에 가장 적합한 커뮤니케이션 수단으로 가장 효과적인 메시지를 전달할 수 있고 (4) 따라서 특정 소비자그룹의 니즈에 적합한 상품이나 서비스를 효율적이고 효과적으로 제공할 수 있게 됨으로써 소비자 만족도를 극대화시킬 수 있다[20].

단계1: 세분화 기준을 결정한다.
단계2: 응답소비자들을 각 세그먼트들에 할당함에 의하여 그룹핑한다.
단계3: 구분된 소비자 그룹들인 세그먼트들을 유의미하게 구분하는 특성(personal characteristics)을 파악하여 세그먼트들을 분석한다.
단계4: 세분된 세그먼트들의 유용성을 경영적 관점에서 평가한다.

그림 1. 시장세분화의 절차

결국 시장세분화의 목적은 기업이나 조직에 가장 가치있는 소비자 그룹을 찾는 과정이라고 할 수 있는데, 그 절차는 [그림 1]과 같다[20]. [그림 1]의 단계3에서 특성들에 의한 세그먼트들끼리의 차별성(differentiation)을 분석할 때 차별화 속성들을 명목변수들(nominal variables)로 설정하고 각 변수들끼리 t-검정(t-test), 카이자승검정(Chi-square tests), 로지스틱 회귀분석(logistic regressions) 등을 이용할 수 있다[20].

시장세분화를 위한 방법은 크게 사전세분화(priori segmentation 또는 commonsense segmentation)와 사후세분화(posteriori segmentation 또는 data driven segmentation)로 나누어서 접근할 수 있다[21]. 사전세분화는 경영자 또는 분석자가 미리 세분화기준을 결정하여 데이터를 수집하고 세그먼트별 특성들을 분석하는 경우이고, 사후세분화는 세분화기준이 미리 결정되지 않고 데이터수집과 분석이 먼저 이루어진 후에 세그먼트 및 세분화기준이 결정되고 세그먼트 즉 세분시장의 유용성 등이 평가되는 경우이다. 즉, 사전세분화에서는 그림1의 단계1동안에 세분화 기준과 세분시장의 유용성 등이 미리 평가되지만 사후세분화에서는 [그림 1]의 단계1과 단계4가 통합되어 평가되는 것으로 생각할

수 있다.

시장세분화의 방법은 사전세분화에서 사후세분화로 이동하여 왔다[21]. 왜냐하면, 하나의 변수에 의하여 시장을 세분화하는 것보다는 다양한 소비자 정보를 바탕으로 여러 관점에서 세분시장을 도출할 수 있는 사후세분화방법이 더 효과적인 분석결과를 가져다 줄 수 있기 때문이다.

설문지 데이터와 같은 정형화된 데이터를 수집하여 사후세분화를 수행할 때는 요인분석(Factor Analysis)이나 클러스터링분석(Clustering Analysis)을 통하여 세그먼트들을 도출한 후에 t-검정(t-test), 카이자승검정(Chi-square tests), 로지스틱 회귀분석(logistic regressions) 등 전통적인 통계분석방법을 이용하면서 세그먼트 별 차별성을 파악할 수 있다.

그러나 온라인 고객리뷰와 같은 비구조적인 데이터를 이용하여 사전세분화 또는 사후세분화를 수행하기 위해서는 텍스트마이닝과 같은 새로운 정보기술을 이용해야 한다. 온라인 고객리뷰를 통계적으로 견고하게 분석하여 사전 시장세분화 또는 사후 시장세분화를 수행하기 위해서는 텍스트마이닝의 다양한 분야와 전통적인 통계분석방법 등을 조합·응용하는 기술이 필요하다.

2. 텍스트마이닝 기술

2.1 텍스트마이닝 개요

텍스트마이닝은 다양한 정보원천(different written resources)으로부터 자동적으로 정보를 추출함으로써 이전에 알려지지 않았던 새로운 정보를 발견하는 정보기술이라고 정의할 수 있다[22]. 텍스트마이닝은 데이터마이닝과 비슷한 개념이지만, 데이터마이닝이 관계형 데이터베이스나 XML과 같은 구조화된 데이터들만을 처리할 수 있는 반면, 텍스트마이닝은 텍스트문서, e-메일, HTML화일과 같은 비구조화 또는 반구조화된 데이터를 처리할 수 있다. 온라인 고객리뷰는 일종의 텍스트문서로서 텍스트마이닝 기술에 의하여 분석 처리될 수 있다.

[그림 2]는 텍스트마이닝 과정을 나타내고 있다. 비구조화된 형태의 문서들(collected documents)이 수집되면 전처리(preprocessing)과정을 거쳐서 텍스트분석(Analyze Text)이 수월한 형태로 변환된다. 전처리작업은 텍스트분석을 위한 알고리즘의 성능을 결정짓는 중요한 요소로서 현재 자연언어처리(natural language processing)와 기계학습(machine learning) 등을 응용하여 활발하게 연구되고 있는 분야이다. 전처리를 거친 텍스트데이터는 컴퓨터가 처리할 수 있는 다양한 방식으로 표현되는데 일반적으로 문서의 단어들이 벡터공간(vector space)속에서 표현된다.

전처리과정을 거친 데이터들을 대상으로 하여 정보추출(information extraction), 범주화(categorization), 문서요약(summarization) 등 다양한 텍스트분석들이 이루어질 수 있다. 텍스트분석 결과는 조직에 유용한 정보 또는 지식의 형태로 나타날 수 있다.

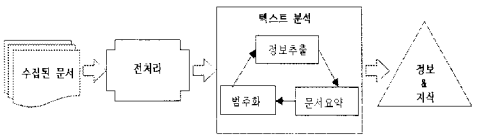

그림 2. 텍스트마이닝 과정

2.2 전처리 과정

텍스트마이닝에서 전처리(preprocessing)는 일반적인 텍스트 데이터들을 컴퓨터가 처리하기 쉽도록 변환하는 작업으로써, 특정단어와 관련된 문서들을 신속하게 검색할 수 있도록 인덱스(index) 화일을 만드는 과정이다. 인덱스를 만드는 방법은 여러 가지가 있지만 FRB(Frequency-Based), IDF(Inverse Document Frequency), LSI(Latent Semantic Indexing) 등이 대표적이다. 문서검색의 정확도를 높이기 위하여 FRB, IDF, LSI 순으로 그 기술이 개선되어 왔다[11].

FRB는 가장 단순한 형태의 인덱싱 방법으로써 문서 안에서 빈번히 나타나는 단어들을 그 문서를 대표하는 중요한 단어로 파악하고 가중치를 높게 주는 개념이다. 따라서 FRB의 인덱싱 방법은 문서들에 포함된 단어들이 특정문서에 몇 번 나타나는지 그 빈도수(frequency)를 벡터공간에 나타낸다. 벡터공간은 문서들과 단어들이 축(dimension)이 되는 2차원공간이라 할 수 있다. 빈도수는 긴 문서에 있는 단어들로의 쏠림현상(skewed)

을 방지하기 위하여 [표 1]에서 처럼 정규화된 빈도수(normalized frequency)형태로 표현된다. 왜냐하면 긴 문서는 짧은 문서에 비하여 특정단어의 출현 빈도수가 높을 가능성이 크기 때문에 결국은 긴 문서가 관련 문서로 검색될 확률이 클 것이다. 따라서 특정단어의 빈도수는 소속 문서의 전체 단어 수에 상대적으로 계산될 필요가 있다. 정규화된 빈도수는 특정단어가 특정문서에서 나타나는 빈도수를 그 특정문서에 포함된 모든 단어의 수로 나눈 것으로써 0과 1사이의 값이 된다. [표 1]에서 FRB 인덱싱의 경우 'human'이라는 단어의 정규화된 빈도수는 문서 D1과 D4에서 0.167이다.

웹페이지를 검색하기 위한 검색엔진(search engines)에서의 인덱싱 개념도 FRB와 비슷하다. 검색엔진에서의 가장 일반적인 인덱싱 방법은 역화일구조(inverted file organization)이다[11]. 역화일구조에서는 특정문서에 나타나는 특정단어의 빈도수뿐만 아니라 그 특정단어가 문서의 어떤 위치에 나타나는지 그 오프셋(offset) 정보도 인덱스화일 안에 포함된다. 그러나 인덱스 파일 안에는 특정문서에서 중요한 단어가 무엇인지에 대한 정보뿐만 아니라 다른 문서와 구분을 해주는 단어가 무엇인지에 대한 정보도 포함되어야 보다 정확한 문서검색이 가능하다. FRB방법은 특정문서에서 중요한 단어가 무엇인지에 대한 정보만 계산하고, 다른 문서와 구분을 해주는 단어가 무엇인지에 대한 정보는 포함하지 않는다.

IDF는 FRB를 보완한 인덱싱 방법으로써 특정문서에서 중요한 단어가 무엇인지 뿐만 아니라 다른 문서와 구분을 해주는 단어가 무엇인지에 대한 정보를 포함하기 위한 계산을 한다. IDF 방법은 수식(1)처럼 특정문서 i에 속하는 특정단어 m의 가중치를 계산할 때 m이 i에서 나타난 빈도수 c_{im}을 f_m과 곱한다. f_m은 m이 모든 문서에 적용되어 계산되는 것으로써, m이 모든 문서에 골고루 나타난다면 f_m값은 작아지고 그렇지 않으면 f_m 값은 커진다. 일반적으로 아래 식처럼 f_m값이 계산된다 [11].

$$f_m = logN - logd_m + 1 \quad (1)$$

(N :문서들의 수, d_m : m을 포함하는 문서들의 수)

[표 1]의 IDF 인덱싱의 경우 'human'과 'interface'는 문서 D1에서 동일한 빈도로 나타나지만 'interface'가 'human'보다 문서 D1을 다른 문서로부터 구분하는데 더 많이 기여함을 나타낸다.

표 1. 인덱스화일의 예

인덱싱 방법	문서 / 단어	D1	D2	D3	D4	D5	D6	D7
FRB	human	0.167	0.000	0.000	0.167	0.000	0.000	0.000
	interface	0.167	0.000	0.000	0.000	0.000	0.000	0.000
	intersection	0.000	0.000	0.000	0.000	0.000	0.250	0.000
IDF	human	0.129	0.000	0.000	0.149	0.000	0.000	0.000
	interface	0.174	0.000	0.000	0.000	0.000	0.000	0.000
	intersection	0.000	0.000	0.000	0.000	0.000	0.288	0.000
LSI	human	0.085	0.000	0.085	0.085	0.000	0.000	0.000
	interface	0.011	0.000	0.011	0.011	0.000	0.000	0.000
	intersection	0.000	0.072	0.000	0.000	0.072	0.072	0.072

IDF가 FRB를 개선하여 문서검색의 정확도를 높이기는 했지만 유사어 또는 개념이나 주제 등에 의한 의미적인 문서검색에는 한계가 있다. 'car'에 의하여 검색되는 문서들은 'vehicle'에 의해서도 검색되어야 하지만 IDF나 FRB는 단어의 정확한 일치를 평가하기 때문에 'car'와 'vehicle'을 다른 단어로 처리한다.

LSI 인덱싱은 문서들이 공유(co-occurrence)하는 단어들을 파악하여 동일한 주제(topic)나 개념(concpet)으로 인식함으로써 검색단어와 정확하게 일치하지 않더라도 개념이나 주제에 의하여 문서검색이 가능할 수 있게 한다. 예를 들어, 'car'와 'vehicle'이 'clutch', 'tires', 'drivers'등의 단어들과 같이 나타나면 LSI에서는 'car'와 'vehicle'을 유사한 것으로 간주한다. LSI 인덱싱에서는 동일한 주제나 개념을 발견하기 위하여 SVD(Singular Value Decomposition)방법을 이용하여 단어-문서-행렬(Words- Documents-Matric)을 3개의 다른 행렬들의 곱으로 변환한다[23].

[표 1]에서 'intersection'이라는 단어는 문서 D6에서만 나타나지만 LSI 인덱싱을 이용하면 'intersection'과 관련된 문서로써 D2, D5, D6, D7이 검색될 수 있다. LSI는 문서들이 공유하는 단어들을 특정주제로 통합하여 새로운 차원(dimension)들로 구성되는 LSI공간(LSI space)에 문서들을 표현하기 때문에, IDF나 FRB에서는 유사하지 않았던 문서들이 LSI공간에서는 유사한 문서로 인식되어 보다 정확한 문서검색이 이루어지는

것이다.

2.3 정보추출

정보검색(Information Retrieval)이 특정단어와 관련된 문서들을 찾는 것임에 반하여 정보추출(Information Extraction)은 특정문서 안에서 유용한 정보 즉, 사람이름(person), 장소이름(place), 전화번호, 날짜(date), 화폐단위(currency) 등 문서내의 개체들(Entities) 및 이들 사이의 관계성(Relationship)을 식별하여 검색해주는 기술이다. 따라서 정보추출을 협의의 의미로 개체추출(Entity Extraction)이라고 부르기도 한다. 개체추출기(Entity Extractor)는 문장(sentence)내의 잠재적인 개체단어(entity instances)를 식별하여 개체타입(entity types)으로 분류하는 기능을 제공하기 때문에 개체추출은 분류문제(classification problem)의 한 영역이라고 할 수 있다.

개체추출기는 2가지 메타데이터(meta data)를 이용하여 개체추출작업의 정확성을 높인다. 그 메타데이터 중의 하나는 개체리스트이고 다른 하나는 태그된 개체들(tagged entities)를 포함하는 훈련데이터(training data set)이다[11].

개체리스트는 일반적으로 많이 사용되는 사람이름이나 장소이름, 화폐단위 등을 포함하는 구조화된 데이터베이스이다. 개체추출기는 문서내의 개체들을 식별할 때 개체리스트를 참조하여 대응하는 개체타입을 결정한다. 그러나 특정개체가 둘 이상의 개체리스트에 포함되어 있을 경우에는 그 개체의 타입을 결정하기가 쉽지 않다. 이때는 휴리스틱(heuristic)한 방법을 이용할 필요가 있다. 태그된 개체들을 포함하는 훈련데이터에 의하여 개체추출기를 훈련시키는 것이 그 한 방법이다. 훈련데이터에 의하여 훈련된 개체 추출기는 훈련데이터 내에서의 개체 특징(feature)뿐만 아니라 개체타입이 발생되는 순서 등을 학습하게 되어, 둘 이상의 개체리스트에 포함된 개체들의 개체타입을 결정하는 문제 등을 해결할 수 있게 된다. 훈련데이터에 의하여 개체추출기를 훈련시키는 기법은 HMM(Hidden Markov Model) 등과 같은 기계학습(Machine Learning)기술이 이용된다.

2.4 범주화

범주화(Categorization)는 수집된 문서들 중에서 유사한 내용의 문서들을 그룹화하여 분류하는 기술로써, 비구조적(unstructured)으로 모여 있는 문서들(collected documents)을 구조적(structured)으로 조직화(organization) 하는 과정이라 할 수 있다[24]. 텍스트마이닝 도구(tools)나 알고리즘들은 구조적으로 조직화된 문서들에 대하여 보다 수월한 분석을 수행할 수 있기 때문에 범주화 과정은 효율적이고 효과적인 텍스트마이닝 분석을 위한 중간단계(intermediate step)라고도 할 수 있다. 일반적으로 문서들의 범주화를 위한 유사도 측정방법(similarity measure)은 통계적인 관점에서 단어들의 빈도수와 문서에서의 동시발생 빈도수(co-occurrence frequency)가 주로 사용된다. Resnik[25]는 문서들 사이의 단어들에 대한 동시발생 빈도수는 그 문서들이 서로 관련되어 있는 증거라고 언급하고 있다. 또한, Hoskinson[26]는 가장 빈번하게 나타나는 단어들에 부합하는 개념을 분류해내기 위하여 단어들의 빈도수와 문서에서의 동시발생 빈도수의 조합을 사용하고 있다.

범주화 과정은 분류를 위한 주제(theme)가 미리 알려지는지의 여부에 따라 크게 감독형(supervised)과 비감독형(unsupervised)으로 나눌 수 있다. 감독형 범주화는 분류를 위한 주제가 미리 정해지는 경우로써 문서분류(document classification)라고도 한다. 비감독형 문서분류는 분류를 위한 주제가 미리 정해지지 않는 경우로써 클러스터링(clustering)이라고도 한다.

감독형 범주화 즉, 문서분류는 주어진 키워드 또는 주제에 따라 문서를 분류하는 기법으로서, 주어진 키워드 집합에 따라 특정문서가 해당 범주로 분류될지 아닐지를 결정하게 된다. 보다 지능적으로 문서분류를 하기 위해서는 미리 범주가 결정된 훈련데이터를 통하여 문서분류기를 학습시킬 수도 있다[24].

클러스터링은 분류키워드 또는 분류주제가 알려지지 않은 상태에서 문서들을 분석하여 유사한 내용의 문서들을 묶은 후에, 필요에 따라 분류된 문서들을 대표하는 키워드들 또는 주제를 추출한다. 수집된 문서들이 매우 많을 때, 클러스터링에 의하여 생성된 클러스터

(cluster)들은 수백 또는 수천 개가 되어 데이터분석자에게 또 다른 부담을 줄 수 있다[11]. 따라서, 대부분의 클러스터링 알고리즘들은 생성되는 클러스터(cluster)들의 최대 숫자나 클러스터 크기, 클러스터에 속하는 문서들 사이의 유사도 등을 입력으로 받아서 보다 융통성 있게 클러스터들을 생성할 수 있다.

문서분류 및 클러스터링은 문서들을 관련 내용별로 자동적으로 조직화함으로써 사용자가 많은 양의 문서들을 좀 더 편리하게 접근할 수 있도록 해준다.

3. 교차분석

교차분석(cross-tabulation analysis)은 전통적인 통계분석방법중의 하나로써 명목자료(nominal variable)를 이용하여 두 변수간의 상호관련성을 알아보고자 할 때 이용된다[27][28]. 교차분석에서 이용되는 통계량은 카이자승(Kai Square, X^2)으로써 이는 기대빈도와 실제빈도의 차이에 의해서 계산된다. 기대빈도와 실제빈도상의 일치정도를 적합도(goodness of fit)라고 하며 이는 X^2 값에 의해서 판단된다. X^2 값이 작을수록 적합도가 높으며 커질수록 적합도가 떨어진다.

두 변수의 상호관련성 또는 차이 여부에 대한 검증은 두 변수의 각 범주(category)가 갖는 도수분포와 두 변수의 분포가 결합되어 나타나는 결합분포를 이용하여 기대빈도를 계산하고, 두 변수의 범주가 결합되어 형성되는 셀에 나타나는 실제빈도와의 차이를 계산하는 방법에 의해서 두 변수가 상호 독립적인지 아니면 서로 관련성이 있는지를 분석하게 된다.

대부분의 텍스트마이닝 분석방법은 단어들의 발생빈도수를 이용하므로 교차분석은 텍스트마이닝분석을 한 단계 승화시켜 응용하는데 이용될 수 있다.

Ⅲ. 방법론 개발을 위한 개념적 모델링

앞에서 살펴보았던 시장세분화의 개념과 텍스트마이닝의 기술분야, 그리고 텍스트마이닝 기술과 전통적인 통계분석방법중의 하나인 교차분석방법은 어느 정도 연관성이 있다. 아래의 [그림 3]은 이러한 연관관계를 나타내고 있다.

그림 3. 시장세분화와 텍스트마이닝 그리고 교차분석의 연관성

앞에서 살펴본 것처럼, 시장세분화는 이질적인 시장을 여러 개의 작은 동질적인 시장으로 구분하는 것이라고 하였는데, 이는 텍스트마이닝에서의 범주화 개념과 일치한다. 시장세분화 과정 중에서 세분화기준을 미리 결정하여 세분화분석을 하는 사전세분화는 텍스트마이닝 범주화의 감독형 범주화에 대응시킬 수 있고, 세분화기준이 미리 결정되지 않는 사후세분화의 경우는 텍스트마이닝의 클러스터링에 대응된다. 또한, 세분화된 시장들인 세그먼트별 특성을 분석하는 과정은 텍스트마이닝의 정보추출 개념과 대응시킬 수 있다. 세그먼트 특성이나 차별성을 도출하기 위하여 선택된 속성을 의미하는 명목변수는 세그먼트 내의 특정 개체타입에 대응시킬 수 있다. 텍스트마이닝 정보추출영역의 개체추출기능을 이용하여 세그먼트 내의 특정개체타입들에 속하는 개체 인스턴스들(entity instances)의 발생 여부 또는 빈도수 등을 계산함으로써 세그먼트별 특성이나 차별성 등을 분석할 수 있다. 세그먼트내의 개체타입에 속하는 개체 인스턴스들의 발생 빈도수에 의한 세그먼트별 차별성의 통계적 유의성을 판별하기 위해서는 교차분석방법을 도입하여야 한다. 교차분석방법은 명목척도의 데이터에 한하여 통계적인 검증을 수행할 수 있는 전통적인 통계분석방법중의 하나인데, 특정 개체타입을 명목척도의 변수로 설정하고 개체타입에 속하는 개체인스턴스들을 적절히 구분하여 명목데이터화 할 수 있다. 예를 들면, 관광객의 휴가여행과 관련하여 '비용관심도'라는 명목변수를 '화폐단위 개체타입'에 대응시키고, 실제 온라인 고객리뷰 상에서 '화폐단위 개체인스턴스들' 이라 할 수 있는 '$', 'Dollar', 'cheap', 'expensive' 등의 단어가 발생하면 휴가 비용에 '관심'이 있는 것으로 인식하고, '화폐단위 개체 인스턴스'가 존재하지 않으면 휴가비용에 '무관심'한 것으로 인식할

수 있다.

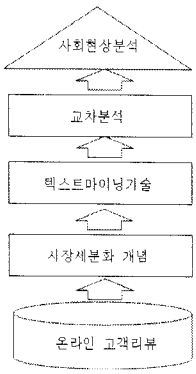

그림 4. 온라인 고객리뷰의 분석을 위한 계층적 개념도

[그림 4]는 시장세분화의 개념과 텍스트마이닝 기술, 교차분석기법 등을 도입하여 온라인 고객리뷰를 보다 효율적이고 효과적으로 분석하기 위한 계층적 개념도를 나타내고 있다. 대량의 온라인 고객리뷰가 존재하고 이를 통계적으로 보다 견고하게 분석하고자 할 때, 시장세분화의 개념을 바탕으로 텍스트마이닝 기술을 적용하고, 텍스트마이닝 분석에 의한 결과 데이터를 교차분석방법에 의하여 다시 분석함으로써 통계적 유의미성을 갖는 사회현상을 발견할 수 있다는 개념이다.

[그림 5]는 온라인 고객리뷰를 분석하기 위한 구체적 방법을 개념적으로 나타내고 있다. 타원형으로 표시된 대량의 온라인 고객리뷰를 분석하기 위하여 첫째, 텍스트마이닝의 범주화 기법을 이용하여 각 범주들 즉 세분시장을 생성한다. 시장세분화의 목표 또는 방법에 따라 감독형 범주화나 비감독형 범주화 방법을 사용할 수 있다. 둘째, 시장세분화의 목표에 따라 세분시장의 특성 또는 차별성을 파악하기 위한 명목변수들을 결정하고 이 명목변수들과 대응할 개체타입을 결정한다. 셋째, 텍스트마이닝의 정보추출 기법을 이용하여 각 범주별로 해당 개체타입들의 빈도수를 계산하고, 넷째 범주별-개체타입별 빈도수를 바탕으로 교차분석을 수행하여 빈도수 차이의 통계적 유의미성을 판별한다.

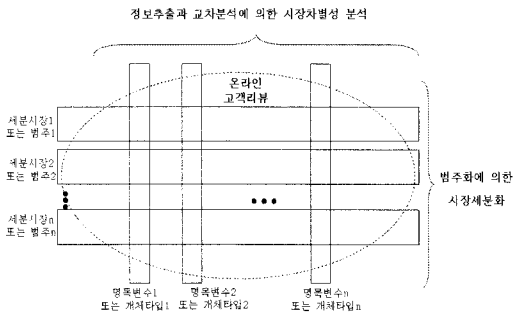

그림 5. 온라인 고객리뷰 분석을 위한 개념적 모델링

Ⅳ. 온라인 고객리뷰 분석을 위한 방법론

본 절에서는 시장세분화의 개념과 텍스트마이닝 기술, 교차분석 방법 등을 이용하여 온라인 고객리뷰를 보다 견고하게 분석하기 위한 방법론을 5단계 절차에 따라 단계별로 제안한다.

단계 1 : 시장세분화에 의한 온라인 고객리뷰의 분석목표 및 방법을 설정한다.

여기에서의 분석목표는 고객리뷰를 범주화하여 각 범주들 사이의 특성 및 차별성을 파악하는 형태여야 한다. 분석목표를 명확히 표현하기 위하여 가설의 형태로 표현하는 것이 바람직하다. 또한, 사전세분화인지, 사후세분화인지에 대한 결정도 이루어져야 한다. 사전세분화의 경우에는 세분화기준과 세분시장 비교를 위한 명목변수들이 결정되어야 한다. 사후세분화의 경우에도 세분시장 비교를 위한 명목변수들이 결정되어야 한다. 현재의 텍스트마이닝 기술로는 선택 가능한 명목변수들이 텍스트마이닝 도구의 정보추출기능이 지원하는 개체타입들로 한정된다.

단계 2 : 분석목표에 적합한 온라인 고객리뷰데이터를 수집한다.

소량의 온라인 고객리뷰를 수집하고자 할 때에는 수작업으로도 가능하지만 대량의 온라인 고객리뷰데이터를 수집하기 위해서는 뉴스모니터(News Monitor)(http://textmine.sourceforge.net/cluster.html)와 같은

자동수집기(automated collector)를 이용하는 것이 바람직하다. 자동수집기를 이용할 경우 모집단 전체를 수집할 수도 있으므로 표본추출의 필요성이 없다. 이것은 텍스트마이닝을 이용한 온라인 고객리뷰 데이터 분석의 또 다른 장점이라고 할 수 있다.

단계 3 : 온라인 고객리뷰데이터를 텍스트마이닝 기술에 의하여 범주화한다.

단계1에서 사전세분화로 결정되었으면 결정된 세분화기준에 대응하는 키워드들을 기반으로 감독형 범주화 즉 문서분류를 실행한다. 사후세분화로 결정되었으면 클러스터링 방법으로 문서분류를 한다. 이때, 필요에 따라 생성될 클러스터들의 개수, 유사도 등의 인자들(parameters)을 입력으로 준다. 사후세분화의 경우 텍스트마이닝 도구가 클러스터를 대표하는 핵심키워드들을 추출하는 기능이 있다면 이 기능을 이용하여 세분화기준을 추론한다.

단계 4 : 명목변수에 대응하는 개체타입을 결정하고, 정보추출기법에 의하여 각 범주별로 개체인스턴스들의 발생빈도수를 계산한다.

명목변수에 대응하는 개체타입을 결정한 다음, 개체타입에 포함될 개체인스턴스들을 결정한다. 개체타입에 포함될 개체인스턴스들을 확장시키기 위해서는 개체추출기를 추가적으로 훈련시켜야 한다.

단계 5 : 범주별-개체타입별 빈도수를 바탕으로 교차분석을 수행하고 빈도수 차이의 통계적 유의미성을 판별하여 세분시장의 특성 및 차별성을 분석한다.

범주유형과 개체타입에 각각 대응하는 두 명목변수 간의 상호관련성 및 독립성을 교차분석을 이용하여 검증한다. 교차분석결과 X^2값이 커지면 기대치와 실제치 간의 차이가 크다는 것을 의미하며 따라서 두 변수의 상관도가 높다는 것을 의미한다. 교차분석에서 이용되는 X^2검증은 자유도를 고려하여 X^2값이 기각역, 즉 '상호독립적이다' 라는 귀무가설이 기각되는 영역에 들어가는지의 여부를 판정해야 한다.

V. 방법론의 응용

본 절에서는 앞 절에서 제안한 방법론의 타당성을 확인하기 위하여 그 응용사례를 제시하고자 한다. 그러나 아직 우리나라에는 한글 텍스트 문서를 분석할 수 있는 오픈소스(open source) 기반의 텍스트마이닝 도구가 없고 또한 양질의 한글판 온라인 고객리뷰를 구하는 것이 어렵기 때문에 부득이 영문 텍스트마이닝 툴과 영문 텍스트 문서를 이용하였다. 미국 관광객들에게 가장 잘 알려져 있고 또 가장 많이 이용하는 관광정보사이트인 TripAdvisor(www.tripadvisor.com)[29]의 온라인 고객리뷰데이터를 앞 절에서 제안했던 방법론 절차를 따라가면서 분석해보았다. 텍스트마이닝 도구는 두 시스템을 이용하였는데, 하나는 콜로라도대학(볼더)의 LSA그룹에서 개발하여 연구자들에게 웹기반 인터페이스를 통한 클러스터링 분석을 허용하는 웹사이트(http://semantics. colorado.edu)이고, 다른 하나는 오픈소스 기반의 연구용 공개시스템인 TextMine (http://textmine.sourceforge. net/)을 설치하여 그 기능들 중 정보추출기능을 이용하였다.

단계 1 : *분석목표 및 방법 설정*

분석목표는 가설형태인 "어트랙션형(Attraction) 관광객과 휴양형(Rest) 관광객사이에 휴가비용에 대한 관심도의 차이가 있다"로 설정하였다. 과거의 연구결과[30]에 의하면, 어트랙션형 관광객과 휴양형 관광객들 사이에는 휴가비용에 대한 관심정도가 다른 것으로 나타났다. 고객리뷰를 범주화하기 위한 방법으로 사전세분화 방법과 사후 세분화방법 양쪽을 이용하기로 하였다. 사전세분화의 경우 세분화기준으로 휴가유형을 선택하고 2개의 관광유형인 어트랙션형 관광과 휴양형 관광을 범주유형으로 설정하였다. 세분시장 비교를 위한 명목변수로는 '비용관심도'로 설정하였다.

단계 2 : *온라인 고객리뷰데이터의 수집*

TripAdvisor 사이트의 고객리뷰들에 대하여 어트랙션과 휴양 관련 고객리뷰들 270건을 수작업으로 수집하였다. 사례분석 용도이기 때문에 고객리뷰 수를 많지 않게 하였다.

단계 3 : *범주화*

사전세분화와 사후세분화에 의한 범주화를 실행하였다. 사전세분화의 경우 범주유형1인 어트랙션형 관광에 대응하는 키워드들을 'attraction', 'amusement', 'shopping', 'theme park' 들로 설정하였고, 범주유형2인 휴양형 관광에 대응하는 키워드들로는 'sight-seeing', 'rest', 'relaxation', 'history culture', 'beautiful scene'들로 설정하였다. 역화일 또는 FRB 인덱싱 방법을 사용하는 감독형 범주화시스템은 270건의 고객리뷰들 중 134건의 고객리뷰들을 '어트랙션형' 으로 분류하였고 136건의 고객리뷰들을 '휴양형' 으로 분류하였다. 사후세분화의 경우 LSI 인덱싱 방법을 이용하는 비감독형 범주화시스템(http://semantics.colorado.edu)으로 범주화를 하였다. 생성될 클러스터 개수를 2로 설정한 결과, 133건의 문서들이 범주유형1(클러스터1)으로 범주화되었고 나머지 137건의 문서들이 범주유형2(클러스터2)로 범주화되었다. 비감독형 범주화시스템은 아직 클러스터를 대표하는 핵심키워드들을 추출하는 기능이 없기 때문에 세분화기준을 추론할 수는 없었다.

단계4 : *각 범주유형별로 정보추출기능에 의한 개체타입의 발생빈도수 계산*

TextMine의 정보추출기능은 'Person', 'place', 'organization', 'number', 'currency', 'dimension', 'time', 'technical term', 'miscellaneous' 등 9개의 개체타입을 식별할 수 있다. 따라서 명목변수 '비용관심도'는 'currency' 개체타입에 대응시켰고 'currency' 개체타입에 포함되는 개체인스턴스들은 '$', 'Dollar' 등 미국 화폐단위뿐만 아니라 전 세계의 모든 화폐단위를 포함하고 있다. 고객리뷰 문서안에 'currency' 개체타입의 인스턴스가 1개 이상 포함되어 있으면 그 고객리뷰는 비용관심도가 있는 것으로 판별하였고, 'currency' 개체타입의 인스턴스가 포함되어 있지 않으면 비용관심도가 없는 것으로 간주하였다. 비용관심도에 대한 보다 정확한 분석을 위해서는 화폐단위에 대한 인스턴스뿐만 아니라 비용과 관련된 단어들 예를 들면, 'money', 'expensive', 'cheap' 등의 단어들도 인식하여야 하지만 TextMine에서는 이러한 단어들을 현재로서는 인식할 수 없다. 사전세분화의 경우 각 범주유형별 'currency' 개체타입의 빈도수는 아래의 [표 2]처럼 나타났고, 사후세분화의 경우는 [표 3]처럼 나타났다.

표 2. 사전세분화에 의한 각 범주별 'currency' 개체타입 빈도수

범주유형 \ 비용관심	무관심	관심	합계
어트랙션형	80	54	134
휴양형	86	50	136
합계	166	104	270

표 3. 사후세분화에 의한 각 범주별 'currency'개체타입 빈도수

범주유형 \ 비용관심	무관심	관심	합계
클러스터1	75	58	133
클러스터2	91	46	137
합계	166	104	270

단계 5 : *범주유형별-개체타입별 빈도수를 바탕으로 교차분석을 수행*

범주유형(휴가유형)과 개체타입(비용관심도)에 각각 대응하는 두 변수간의 상호관련성을 알아보기 위하여 교차분석을 수행하였다. 사전세분화 및 사후세분화의 교차분석결과는 [표 4]와 같다.

표 4. 교차분석결과

분석방법 \ 통계량	Pearson카이제곱값	자유도	유의확률(양측)
사전세분화방법	0.356	1	0.551
사후세분화방법	2.868	1	0.090*

*<.1

세분화기준을 휴가유형으로 하여 어트랙션형 휴가와 휴양형 휴가로 분류하였던 사전세분화의 경우, 휴가유형과 비용관심도 사이에는 상호관련성이 없는 것으로 나타났다. 따라서 사전세분화의 경우에는 단계1에서 설정하였던 가설이 기각되었다. 반면, 세분화기준이 결정되지 않은 채 두 범주유형인 클러스터1과 클러스터2로 분류하였던 사후세분화의 경우, 범주유형과 비용관심도 사이에는 90%의 유의확률로 상호관련성이 있는 것으로 나타났다. 각 범주를 대표하는 키워드가 추출되지 못하였기 때문에 각 범주가 어떤 세분시장인지를 유추하는 것은 현재로서는 어렵다. 이것은 분류된 클러스터들을 대표하는 주제 또는 핵심키워드를 추출할 수 있는 기능이 중요할 것임을 시사한다.

Ⅵ. 결 론

고객리뷰는 구전의 한 종류로서 소비자들의 구매의사결정에 중요한 역할을 하여왔다. 인터넷의 발달과 함께 고객리뷰는 온라인고객리뷰의 형태로 변화되어 더욱 많은 사람들이 손쉽게 생성하고 활용할 수 있기 때문에 그 양이 급속도로 증가하고 있고 따라서 유용성 및 중요성은 더욱 커지고 있다.

본 논문에서는 텍스트마이닝 기술을 이용하여 온라인고객리뷰를 분석할 수 있는 방법론을 제안하였다. 소비자들의 경험과 의견을 웹상에서 표현한 온라인고객리뷰를 분석할 때 동일한 유형의 소비자의견을 묶어서 분석하면 그 효율성과 효과성이 클 수 있을 것이기 때문에 시장세분화의 개념을 도입하였다. 즉, 시장세분화의 개념과 부합하는 텍스트마이닝 기술들인 범주화 기능과 정보추출기능을 이용하여 온라인고객리뷰를 분석할 수 있는 방법론을 제안하였다. 또한, 통계적으로 견고한 분석결과를 도출하기 위하여 전통적인 통계분석방법 중의 하나인 교차분석을 방법론에 포함시켰다. 제안한 방법론의 타당성 여부, 그 응용방법 또는 기술적인 미비점을 확인하기 위하여 관광관련 온라인고객리뷰들을 실제로 분석하여 보았다. 텍스트마이닝 관점에서 기술적으로 더 발전되어야 할 부분은 첫째, 정보추출기능을 보다 더 확장시키고 지능화시킬 필요가 있다. 분석자가 원하는 개체타입은 대부분 지원할 수 있어야 하고 더 나아가 분석자가 분석과정 중에서 개체타입을 생성할 수 있으면 더욱 바람직할 것이다. 또한, 문맥(context)에 따라 지능적으로 개체인스턴스들을 분류할 수 있어야 보다 정확한 세분시장 특성을 파악할 수 있다. 둘째, 비감독형 범주화 기능 중에서 클러스터를 대표하는 핵심주제 또는 핵심키워드를 추출하는 기능이 필요하다. 일반적으로 비감독형 범주화 분석이 감독형 범주화보다 유용한 분석결과를 생성할 수 있다고 알려져 있지만 이는 클러스터를 대표하는 핵심주제를 추출할 수 있을 때 그 효과성이 있다고 할 수 있다. 본 논문의 의의는 첫째, 텍스트 마이닝 관련 기존의 연구들이 기술 중심으로 이루어졌음에 비하여 본 연구에서는 기존의 다양한 텍스트마이닝 기술들의 적절한 조합을 통한 그 응용방법을 제안하였다. 특히, 전통적 통계분석방법과의 접목을 통하여 통계적으로 보다 견고한 분석방법을 제안하였다. 둘째, 제안한 방법론은 설문지 작성을 위한 사전조사 및 분석에 활용될 수 있다. 성의 있게 작성할 것으로 판단되는 인터넷 고객리뷰를 분석한 결과는 무성의한 응답의 문제가 있는 설문조사의 결과보다 객관적일 수 있다. 특히, 설문지의 내용을 결정하기 위한 1차 자료로써 텍스트마이닝 결과를 활용함으로써 보다 현실적이고 충실한 설문지를 만들 수 있고 이러한 설문지를 바탕으로 한 설문조사 결과는 보다 객관적인 자료가 될 수 있다. 셋째 공학적 관점과 사회과학적 관점을 융합하여 새로운 분석방법론을 제안함으로써 시너지 효과를 창출하였다.

본 연구의 한계는 방법론응용 부분에서 오픈소스 기반의 한글 텍스트마이닝 도구가 부재하고 양질의 한글판 온라인 고객리뷰를 구하는 것이 어려워서 영문 텍스트마이닝 도구와 영문 온라인 고객리뷰를 이용하였다는 점이다. 한국에서도 오픈소스 기반의 다양한 한글 텍스트마이닝 도구들이 개발되어야 할 것이며 특히, 미국의 TripAdvisor 사이트 같은 양질의 웹사이트들이 활성화된다면 본 논문에서 제안했던 방법론의 기여도는 더욱 커질 것이다.

참 고 문 헌

[1] A. W. Joshi and S. Sharma, "Customer Knowledge Development: Antecedents and Impact on New Product Performance," Journal of Marketing, 68, pp.47-59, 2004.

[2] 양소영, *온라인 고객리뷰의 내용분석 : 채널, 제품 속성, 가격을 중심으로*, 한국과학기술원, 석사학위논문, 2006.

[3] Alexander Maedche and Steffen Staab, "Discovering Conceptual Relations from Text", European Conference on AI, pp.1-6, 2000.

[4] Paola Velardi, Michele Missikoff and Roberto Basili, "Identification of relevant terms to support the construction of domain ontologies," Proc. of the workshop on Human Language Technology and Knowledge Management Vol.2001, pp.1-11, 2001.

[5] S. Fabrizio, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol.34, No.1, pp.1-47, 2002(3).

[6] C. Jamie and M. Teruko M, "Knowledge-Based Extraction of Named Entity," CIKM'02, pp.4-9, 2002(11).

[7] H. Minqing and L. Bing, "Mining and Summarizing Customer Reviews," KDD'04, August 22-25, 2004.

[8] L. Stanley, L. Fabiana, S. Ramiro, and L. Daniel, "A Tourism Recommender System Based on Collaboration and Text Analysis," Information Technology & Tourism, Vol.6, pp.157-165, 2004.

[9] Massimiliano Ciaramita, Aldo Gangemi, Esther Ratsch, Jasmin Saric and Isabel Rojas, "Unsupervised Learning of Sementic Relations between Concepts of a Molecular Biology Ontology," IJCAI-2005, pp.1-6, 2005(8).

[10] J. M. Raymond and B. Razvan, "Mining Knowledge from Text Using Information Extraction," SIGKDD Explorations, Vol.7, issue1, pp.3-10, 2005.

[11] K. Manu, *Text Mining Application Programming*, Thomson Charles River Media, 2006.

[12] Nadzeya Kiyavitskaya, Nicola Zeni, Luisa Mich, James R. Cordy, and Mylopoulos, "Text Mining through Semi Automatic Sementic Annotation," Lecture Notes in Computer Science, Vol.4333, 2006.

[13] Lipika Dey, Muhammad Abulaish, Jahiruddin and Gaurav Sharma, "Text Mining through Entity-Relationship Based Information Extraction," 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp.177-180, 2007.

[14] Michele Banko, Michele J Cagarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, "Open Information Extraction from the web," IJCAI(International Joint Conference on Artificial Intelligence)-07, pp.2670-2676, 2007.

[15] Venkatesh Ganti, Arnd Christian Konig and Rares Vernica, "Entity Categorization Over Large Document Collections", KDD'08, August, pp.24-27, 2008.

[16] R. L. Kai and E. M. David, "A Mathematical Approach to Categorization and Labeling of Qualitative Data: The Latent Categorization Method," Sociological Methodology, Vol.34, No.1, pp.349-392, 2004.

[17] Tomasz Miaskiewicz, Tamara Sumner and Kenneth A. Kozar, "A Latent Sementic Analysis Methodology for the Identification and Creation of Persons," CHI 2008 Proceedings, April, pp.5-10, 2008.

[18] J. E. Swan and R. L. Oliver, "Postpurchase communication by consumer," Journal of

Retailing, 65(Winter), pp.516-533, 1989.

[19] W. Smith, Product Differentiation and Market Segmentation as Alternative Marketing Strategies, Journal of Marketing, 21, pp.3-8, 1956.

[20] S. Dolnicar, A. Woodside, and D. Martin, Market Segmentation in Tourism, Tourism Management, Analysis, Behavior and Strategy, Cambridge, CABI, 2007.

[21] J. A. Mazanec, Market Segmentation. In: J. Jafari(Ed), Encyclopedia of Tourism, London:Routledge, 2000.

[22] Weiguo Fan and Linda Wallage, Stephanie Rich and Zhongju Zhang, "Tapping The Power of Text Mining," Communications of ACM, Vol.49, No.9, 2006.

[23] M. W. Berry, S.T.Dumais, and T.A.Letsche, "Computational methods for intelligent information access," Proceedings of Supercomputing'95, San Diego, CA, 1995.

[24] Anna Stavrianou, Periklis Andritsos and Nicolas Nicoloyannis, "Overview and Semantic Issues of Text Mining," SIGMOD Record, September, Vol.36, No.3, 2007.

[25] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy. In Proc. of the 14th IJCAI-95, Montreal, QC, Canada, pp.448-458, 1995.

[26] A. Hoskinson, Creating the ultimate research assistant. IEEE Computer, 38(11), pp.97-99, 2005.

[27] 채서일, 김선철, 최수호, *SPSS WIN을 이용한 통계분석*, 학현사, 2002.

[28] Kenneth J. Meier, Jeffrey L. Brudney, John Bohte, *Applied Statistics*, Michael Rosenberg, 2006.

[29] Prophis Research and Consulting Inc., *Travel 2.0 US Report*, 2008)9).

[30] R. C. Melvin, "Segmenting the Vacationer Market: Identifying the Vacation Preferences, Demographics, and Magazine Readership of Each Group," Journal of Travel Research, Vol.20, No.2, 29-34, 1981.

저 자 소 개

김 근 형(Keun-Hyung Kim) 정회원

- 1990년 2월 : 서강대학교 컴퓨터학과(공학사)
- 1992년 2월 : 서강대학교 컴퓨터학과(공학석사)
- 2001년 2월 : 서강대학교 컴퓨터학과(공학박사)
- 2001년 9월 ~ 현재 : 제주대학교 경영정보학과 부교수

<관심분야> : 텍스트마이닝, 데이터마이닝, MIS

오 성 열(Sung-Ryoel Oh) 정회원

- 1990년 2월 : 제주대학교 회계학과(경영학사)
- 1992년 2월 : 한양대학교 회계학과(경영학석사)
- 1996년 2월 : 한양대학교 회계학과(경영학박사)
- 1997년 3월 ~ 현재 : 제주산업정보대학 세무회계과 조교수

<관심분야> : 세무회계, 관리회계, 마케팅