

귀납적 사회과학연구를 위한 확장형 오피니언마이닝 시스템 설계

Design of expanded opinion mining system for supporting
inductive method in social science research

김 근 형* · 김 두 경**

(Kim, Keun Hyung · Kim, Doo Gyung)

목 차

- I. 서론
- II. 이론적 배경
- III. 확장형 오피니언마이닝시스템의 설계
- IV. 결론

I. 서론

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) 형태인 온라인 '고객리뷰들(Online Customer Reviews)'이 방대하게 존재하고 있으며 더욱 증가되고 있는 추세이다. 미국의 최대 온라인 관광 사이트인 TripAdvisor (www.tripadvisor.com)은 관광객들이 직접 기록

* 제주대학교 경상대학 경영정보학과 교수

** 제주대학교 경상대학 경영정보학과 교수

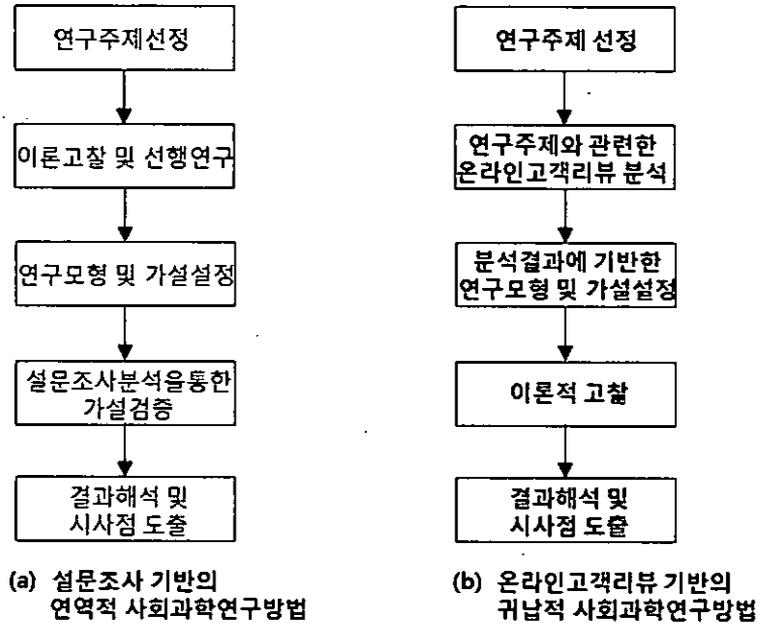
해 놓은 수십만 건의 관광정보와 경험·지식들이 비정형 데이터 형태로 게시되어 있고 많은 관광객들이 유용한 정보로 활용하며 확대 재생산하는 웹2.0의 대표적인 성공사례로 손꼽히고 있다.

웹2.0시대의 성숙화를 주도하고 있는 SNS(Social Network Services)는 우리의 일상생활에서부터 정치 환경에 이르기까지 커다란 영향을 미치고 있으며 빅데이터 생성의 중요한 원천이 되고 있다(김상락, 2012). SNS의 웹사이트에서는 수많은 사람들이 일상생활의 광범위한 분야에서 무엇을 좋아하고 싫어하는지, 삶의 많은 양상에 대한 다양한 의견(opinion)이나 리뷰(reviews)들을 게시하고 교환하면서 그 데이터량이 급속도로 증가하고 있다.

웹2.0과 SNS로 대표되고 있는 지식정보화시대에는 인터넷과 같은 사이버공간에서 이루어지는 인간행동들이 여론을 주도하며 사회변화에 커다란 영향을 미치고 있다. 기존의 사회과학연구방법의 주된 흐름은 설문조사(questionnaire survey) 등의 방식을 통한 인간행동의 관찰이 주를 이루었지만, 지식정보화시대에는 웹2.0과 SNS와 같은 사이버공간 상에서의 인간행동을 관찰하는 사회과학연구도 병행될 필요가 있다. 설문조사는 기존의 사회과학연구방법으로 인기가 있으며 현재도 많이 사용되고 있지만, 설문 응답자들의 무성의한 답변에 대한 염려 때문에 그 결과에 대한 신뢰성을 의심하는 학자들도 많이 있다. 그러나 웹사이트나 SNS상에 게시된 온라인고객리뷰는 작성자의 흥미와 의지가 전제되어야 만들어질 수 있는 텍스트데이터이기 때문에 설문조사에 기반한 데이터보다 내용면에서 더 충실한 자료라고 할 수 있다. 앞으로는 사회과학자들도 새로운 사회 현상을 규명하기 위하여 웹2.0 상에 존재하는 방대한 네티즌 의견들을 분석하고 반영하여야 그 학문적 가치를 인정받는 시대가 될 것이다.

온라인 고객리뷰를 분석하기 위해서는 기존의 사회과학연구방법에 대한 패러다임이 변화될 필요가 있다. <그림1>은 전통적 사회과학연구방법과 새로운 사회과학연구방법을 비교하여 나타내고 있다. <그림1>의 (a)는 설문조사기반의 전통적 사회과학연구방법을 나타내고 있으며, (b)는 본 논문에서 제안하고자 하는 온라인고객리뷰 기반의 새로운 사회과학연구방법을 나타내고 있다. 전통적 사회과학연구방법은 일반적으로 기존 이론모형을 기반으로 새로운 연구모형과 가설을 도출하고 설문조사를 기반으로 데이터를 수집하여 가설검증을 하는 연역적(deductive) 방법이다. 이에 반하여, 본 논문에서 제안하고자 하는 온라인고객리뷰 기반의 새로운 사회과학연구방법은 먼저 연구주제와 관련한 고객리뷰데이터를 수집하여 분석한 후 이를 기반으로 이론을 정립하는 형태의 귀납적(inductive) 방법이다. 정보기술이 발전하지 않고 온라인고객리뷰가 많지 않았던 시절에는 연역적 방법이 주로 사용되었으나 웹2.0시대에는 온라인고객리뷰기반의 귀납적 연구방법이 필요하며 그 중요

성이 커졌다.



<그림1> 사회과학연구방법의 비교

II. 이론적 배경

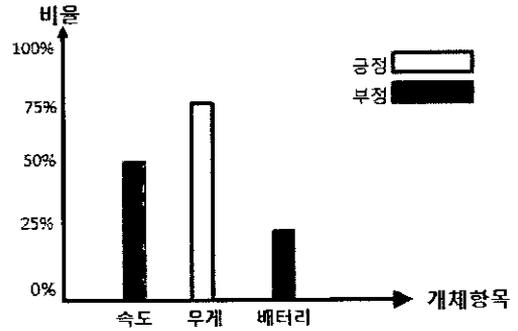
2.1 오피니언마이닝 개요

최근에 연구되고 있는 오피니언마이닝(opinion mining)은 웹사이트에 게시되어 있는 온라인 고객리뷰들을 분석 대상으로 하는 텍스트마이닝(Text Mining)의 한 분야로서 고객 의견에 대한 긍정(positive)과 부정(negative)의 분포 등을 분석할 수 있다. 온라인고객리뷰 기반의 귀납적 연구방법은 온라인고객리뷰의 분석을 전제로 한다. 오피니언마이닝 기술은 온라인고객리뷰라는 텍스트 데이터에서 빈번하게 출현하는 개체들을 먼저 추출하고 이들 개체들의 긍정/부정 비율을 판별하는 기술이다. <그림2>는 오피니언마이닝의 예를 나타내고 있다. <그림2>의 (a)는 최신 스마트폰에 대한 품평이 들어 있는 4개의 온라인고객리뷰 예들을 나타내고 있다. <그림2>의 (b)는 4명의 고객들이 작성한 온라인고객리뷰들 중에서 한번 이상 출현한 개체들에 대한 고객들의 긍정/부정 의견의 비율을 나타내고 있

다. 스마트폰의 ‘속도’ 개체에 대해서는 4명 중에서 2명(50%)의 고객이 부정적인 의견을 보이고 있다. 스마트폰의 ‘무게’ 개체에 대해서는 4명중에서 3명(75%)이 긍정적인 의견을 나타내고 있다.

리뷰1	속도는 느리지만, 무게가 가벼워요
리뷰2	속도가 너무 느려요
리뷰3	무게는 가볍지만 배터리 지속시간이 너무 짧아요
리뷰4	무게가 가벼워서 좋네요

(a) 4개의 온라인고객리뷰 예



(b) 오피니언마이닝 분석 결과 예

<그림2> 오피니언마이닝 예

2.2 오피니언마이닝 관련 연구 고찰

온라인고객리뷰 분석기술은 현재 오피니언마이닝(Opinion Mining) 분야에서 활발하게 연구되고 있다.

Liu(2005)는 기계학습 및 자연어처리기술을 활용하여, 온라인고객리뷰 데이터에 대한 감성분석과 분석결과의 요약기법을 제시하고 있으며, 오피니언옵저버(Opinion Observer)라는 시스템을 개발하였다. Christopher Scaffidi(2007)은 미국 카네기멜론 대학교에서는 레드오파(Redopal) 시스템을 개발한 사례가 있으며, 이는 고객리뷰 데이터와 사용자 평가 점수를 활용하여 요약보고서를 생성하는 시스템이었다. Xiaowen Ding(2007)은 문장구조와 문장 사이의 관계, 문장성분의 패턴정보 등과 같은 언어규칙을 이용하면서 통계학적 방법으로 오피니언마이닝에 접근하고 있다. 워드넷을 활용하여 어휘의 긍정이나 부정적 의미를 판단하고 이를 센티워드넷(SentiwordNet)으로 응용하여 감정의 폭을 정량화하는 방법을 제시하는 다수의 연구들도 있다(Xiaowen Ding et. al, 2007; Courses et. al, 2008; Mingqing Hu et. al, 2004) W.Y.Kim(2009)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. Alexander Pak(2010)은 이모티콘

을 활용하여 텍스트의 긍정적 또는 부정적 감정을 인식하도록 하였는데, 이모티콘 기반의 감정분류 성능은 70~80% 사이의 정확도를 보였다. SNS에는 매 순간 엄청난 수의 사용자가 이용하기 때문에 긍정/부정 오피니언의 변화가 지속적으로 일어날 수 있으며, 이와 관련한 최신의 데이터에 기반하여 효율적으로 분석결과를 업데이트하는 방법이 제안되었다(Ismael S. Silva et. al, 2011). J.Leskovec(2010)은 SNS의 모든 리뷰를 동일선상에서 고려하는 것이 아니라 신뢰할 만한 사용자가 작성한 리뷰에 가중치를 부여하여 전체 오피니언의 결정에 더 영향을 발휘하도록 하고 있다. Yang Yuehua(2010)은 SNS와 관계된 다양한 엔티티들(즉, 사용자, 사용자그룹, 응용서비스, 포스트 등) 사이의 본질적 관계성을 표현하는 그래프 생성모델을 제안하고 있으며, 이러한 엔티티들 사이의 관계성에 기반한 규칙들을 추출하는 방법을 제안하고 있다. Guang Giu(2009)는 제품속성단어와 제품속성을 수식하는 감성단어 사이의 의존관계를 통하여 핵심 감성단어를 자동 추출하는 방법을 제안하고 있으며, Lei Zhang(2010)은 제품속성 단어와 제품속성을 수식하는 감성단어 사이의 의존관계에 HITS(Hyperlink-induced topic search) 알고리즘을 적용하여 제품속성의 랭킹을 결정하는 기법을 제안하고 있다.

이러한 연구결과들을 종합하여 볼 때, 오피니언마이닝과 관련한 기존 연구들은 대부분 오피니언 게시자들이 동일한 속성을 가진 사람들인 것으로 간주하여, 표면적인 리뷰 분석에만 집중해왔다는 한계가 있었다. 즉, 리뷰의 대상이 되는 제품이나 제품속성에 대응하는 정확한 명사를 찾는 방법이라든가, 긍정/부정을 나타내는 감성단어를 정확하게 분류하는 방법 등과 관련된 것들이었다. 덕분에 오피니언마이닝의 분석 정확도는 많이 개선되었다. 그러나, 오피니언마이닝을 사회과학연구방법에 적용하기 위한 기술확장에 대한 연구는 전무한 실정이었다.

Ⅲ. 확장형 오피니언마이닝시스템의 설계

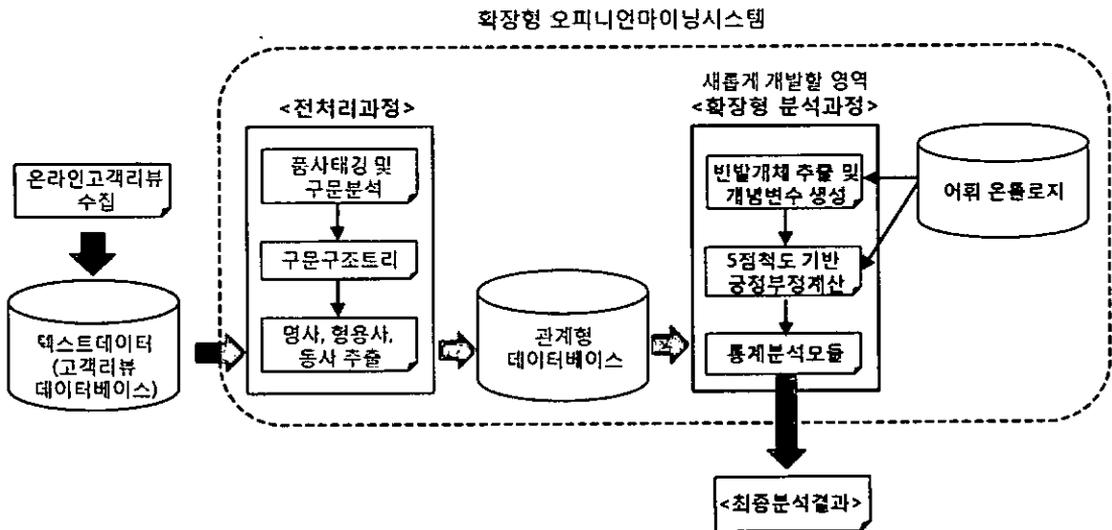
3.1 시스템 개요

<그림3>은 확장형 오피니언마이닝시스템이 제공해야 할 기능들을 포함하고 있다. 웹사이트 상에 게시된 온라인고객리뷰들은 텍스트문서 형태로 고객리뷰 데이터베이스에 저장된다. 확장형 오피니언마이닝시스템은 고객리뷰 데이터베이스 상의 텍스트 데이터를 입력으로 받아 <전처리과정>(preprocessing)과 <확장형 분석과정>을 거치면서 <최종 분석결

과>를 도출한다. 확장형 오피니언마이닝시스템에서 <전처리과정>은 기존 오피니언마이닝 연구에서도 다루었던 영역인 반면, <확장형 분석과정>은 기존 오피니언마이닝에서는 다루 어지지 않았으며 본 논문에서 새롭게 제안하는 영역이다.

텍스트문서는 비정형 데이터이므로 전처리과정을 통하여 정형 데이터인 관계형 파일로 변환된다. <전처리과정> 중에서 텍스트문서내의 각 문장들은 구문분석기에 의하여 각 단어에 품사가 부여된 형태의 구문구조트리로 변환된다. 구문구조트리 파일로부터 명사, 형용사, 동사 등에 해당하는 단어들 추출되어 테이블 형태의 관계형 파일에 저장된다. 관계형 데이터베이스는 관계형 파일들인 명사테이블, 형용사테이블, 동사테이블 등을 포함한다. 관계형 데이터베이스 상의 각 파일들은 <확장형 분석과정>에서 사용된다.

<확장형 분석과정>에서 빈발개체 추출 모듈은 일정 횟수 이상 출현하는 명사들을 추출 하는 기능을 갖는다. 빈발개체들은 어휘 온톨로지의 정보를 바탕으로 각각의 개념변수들 로 묶이게 된다. 어휘 온톨로지는 단어들 사이의 유사어나 반의어 관계, 포함관계 등의 정 보가 들어있는 데이터베이스이다. 긍정부정계산 모듈은 빈발개체들의 긍정/부정 정도를 5 점 척도로 평가한다. 긍정부정계산 모듈도 어휘 온톨로지의 정보를 참조하면서 그 기능을 수행한다. 개념변수에 포함된 빈발개체들의 긍정/부정 정도의 평균값은 개념변수의 긍정/ 부정 정도값이 된다. 통계분석모듈은 개념변수의 긍정/부정 정도값을 이용하여 개념변수들 사이의 상관분석, 회귀분석 등의 전통적 통계분석을 수행한다. <최종분석결과>에는 개념 변수들 사이의 연관관계와 통계적 유의성이 출력된다.



<그림3> 확장형 오피니언마이닝시스템 구조

3.2 시스템 구성요소

확장형 오피니언마이닝시스템의 주요 구성요소들은 <전처리과정>과 <확장형 분석과정>에 속하는 모듈들(modules)과 어휘 온톨로지(ontology) 부분이다.

3.2.1 <전처리과정>의 주요 모듈

(가) 구문분석기

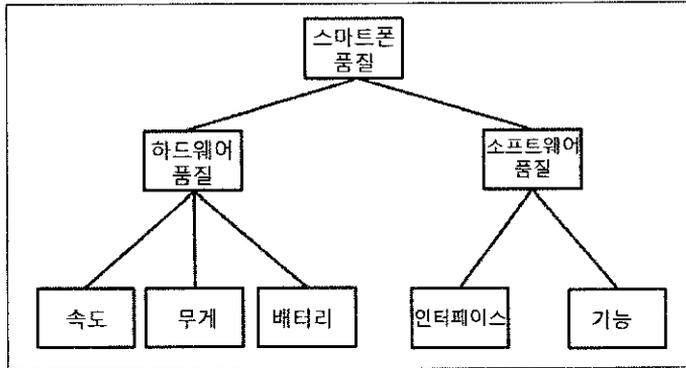
텍스트 데이터를 입력으로 하여 구문구조트리를 생성하는 구문분석기는 공개된 다양한 소프트웨어들이 존재한다. 대표적으로 미국 스탠포드대학에서 개발한 영어 구문분석기 (<http://nlp.stanford.edu/software/lex-parser.shtml>) 등을 사용할 수 있다. 한국어 구문분석기로는 KNLP(<http://nlp.kookmin.ac.kr/demo/parsedemo.html>) 등이 있다.

(나) 명사, 형용사, 동사 추출모듈

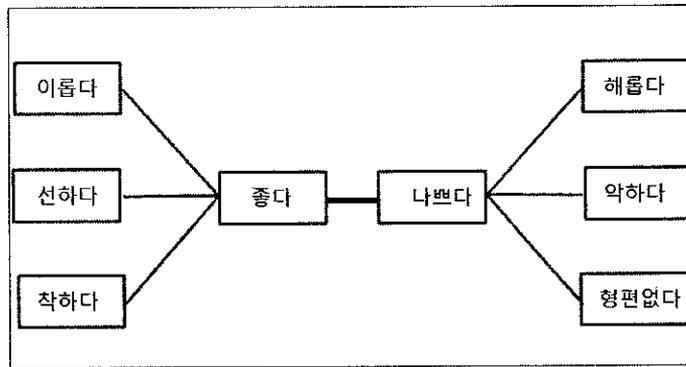
구문구조트리로부터 명사나 형용사, 동사 등을 추출하여 각각의 관계형 테이블에 입력하는 모듈로서, 프로그래밍언어와 DBMS를 이용하여 구현할 수 있다.

3.2.2 어휘 온톨로지 구축

온톨로지는 단어와 단어 사이의 관계들로 구성된 일종의 메타(meta) 사전으로서 생각할 수 있으며, 그 속에는 특정 도메인(domain)에 관련된 단어들이 계층적으로 표현되어 있고, 추가적으로 이를 확장할 수 있는 추론 규칙이 포함되어 있어, 웹 기반의 지식 처리나 응용 프로그램 사이의 지식 공유, 재사용 등이 가능토록 되어 있다(네이버 지식백과). <그림 4>는 온톨로지의 예를 나타내고 있다. <그림 4>의 왼쪽 그림은 단어들 사이의 <상하-포함 관계>를 계층적으로 표현한 온톨로지이고, 오른쪽 그림은 단어들 사이의 <반의어-유의어 관계>를 표현한 온톨로지이다. 본 과제에서는 확장형 오피니언마이닝시스템이 분석할 온라인고객리뷰에 대한 온톨로지를 구축할 것이다. 온톨로지를 구축하기 위한 언어들(RDF 또는 OWL 등)이 있지만, 구축의 수월성을 위하여 MySQL 등의 DBMS를 이용하여 구축할 수 있다. 것이다. 또한, 분석자가 직접 온톨로지를 구축할 수 있도록 지원하는 응용 프로그램을 개발할 수 있다.



<상하-포함관계>



<반의어-유사어 관계>

<그림4> 온톨로지의 예

3.2.3 <확장형 분석과정>의 주요 모듈

(가) 빈발개체 추출 모듈

빈발개체 추출모듈은 관계형데이터베이스 상의 명사 테이블에서 최소지지도 이상으로 빈번하게 출현하는 개체들을 추출하는 기능이다. 예를 들면, 새로 출시된 스마트폰에 대한 상품평이 있는 온라인고객리뷰에서 '속도', '무게', '배터리' 등과 같은 개체들이 최소지지도 이상의 빈발도로 출현하는 경우 빈발개체 추출모듈은 '속도', '무게', '배터리' 등과 같은 개체들을 빈발개체로 추출한다. 빈발개체 추출모듈은 프로그래밍언어와 DBMS를 이용하여 개발할 수 있다.

(나) 개념변수 생성모듈

개념변수 생성모듈은 빈발개체들 중에서 특정 개념에 의하여 서로 관련된 빈발개체들을

뭉는 기능을 갖는다. ‘속도’, ‘무게’, ‘배터리’ 등과 같은 개체들이 빈발개체인 경우, <그림 4>의 온톨로지에서 <상하-포함관계>에 의하여 ‘하드웨어품질’이라는 개념변수가 생성된다. 개념변수 생성모듈은 프로그래밍언어와 DBMS를 이용하여 개발할 수 있다.

(다) 긍정/부정 계산모듈

긍정/부정 계산모듈은 빈발개체들에 대하여 개체를 수식하는 형용사 및 동사의 긍정/부정 정도를 계산하는 기능을 갖는다. <표2>는 2점 척도와 5점 척도의 긍정/부정 정도값을 비교하여 나타내고 있다.

<표 2> 2점 척도와 5점 척도의 비교

유형	척도	정도	비고
2점 척도	1점	부정	기존 오피니언마이닝
	2점	긍정	
5점 척도	1점	매우 부정	확장형 오피니언마이닝
	2점	부정	
	3점	보통	
	4점	긍정	
	5점	매우긍정	

예를 들면, 긍정/부정 정도의 5점척도를 <매우부정(1), 부정(2), 보통(3), 긍정(4), 매우 긍정(5)>로 설정하였을 때, 앞에서 살펴보았던 <그림3> (a)의 온라인고객리뷰에서 리뷰 1(“속도는 느리지만, 무게가 가벼워요”)에 대한 5점척도 평가는 “부정(2)”로 평가되며 리뷰 2(“속도가 너무 느려요”)에 대한 5점척도 평가는 “매우부정(1)”으로 평가된다. <그림8>은 빈발개체가 속하는 개념변수와 함께 각 빈발개체들에 대한 긍정/부정 정도값이 나타나 있다. <그림3>(a)에 있는 4개의 온라인고객리뷰들에서 최소지지도가 1인 빈발개체들에 대하여 긍정/부정 계산모듈이 생성한 5점척도 긍정/부정 정도값은 <그림5>와 같다. <그림 5>에서 어떤 빈발개체의 긍정/부정 정도값은 결측치로 나타나고 있다.

리뷰번호	개념변수	빈발개체	매우부정	부정	보통	긍정	매우긍정
1	하드웨어 품질	속도		√			
		무게				√	
		배터리					

리뷰번호	개념변수	빈발개체	매우부정	부정	보통	긍정	매우긍정
2	하드웨어 품질	속도	√				
		무게					
		배터리					

리뷰번호	개념변수	빈발개체	매우부정	부정	보통	긍정	매우긍정
3	하드웨어 품질	속도					
		무게				√	
		배터리					√

리뷰번호	개념변수	빈발개체	매우부정	부정	보통	긍정	매우긍정
4	하드웨어 품질	속도					
		무게				√	
		배터리					

<그림5> 개념변수 및 5점척도 계산의 예

(라) 통계분석 모듈

통계분석모듈은 개념변수들에 대한 5점척도 데이터를 입력으로 하여 개념변수들 사이의 상관관계와 회귀관계를 분석하는 기능을 갖는다. 긍정/부정 계산모듈에 의한 5점 척도 데이터가 생성되면 기존의 통계패키지를 활용하여 수동적으로 통계분석을 할 수도 있다.

IV. 결론

기존의 사회과학연구방법은 가설을 먼저 설정하고 이를 지지할 수 있는 실증적 자료를 수집하는 연역적 연구방식을 따르지만, 빅데이터 시대에는 이미 존재하는 텍스트 데이터를 분석하여 특정 사회현상이 반영된 패턴을 도출하고 이 패턴을 기반으로 이론을 정립하는 귀납적 연구방식이 사회과학연구의 또 다른 한 축으로 정립될 필요가 있다.

본 논문에서는 텍스트 데이터를 분석함으로써 귀납적 사회과학연구를 지원할 수 있는

오피니언마이닝시스템의 개념을 소개하고 그 기능에 대한 설계도를 제안하였다.

본 논문에서 제안한 확장형 오피니언마이닝시스템과 그 관련 기술은 다음과 같이 네 가지 측면에서 그 활용 효과를 기대할 수 있다. 첫째, 확장형 오피니언마이닝시스템을 활용하여 새로운 차원의 귀납적 사회과학 연구 방법을 수행할 수 있다. 기존의 사회과학연구 방법론은 대부분 설문조사에 의하여 확보된 정형화된 데이터만을 통계분석 대상으로 하였지만, 확장형 오피니언마이닝시스템을 활용한다면 이미 생성되어 존재하고 있는 다양한 일반 텍스트 문서들도 사회현상 규명을 위한 통계분석의 대상이 될 수 있으며, 비정형 텍스트문서를 분석한 결과의 통계적 유의성 여부를 정량적으로(quantitatively) 판단할 수 있다. 특히, 텍스트 데이터는 작성자의 흥미와 의지가 있어야 만들어질 수 있는 것이기 때문에 설문조사에 의한 데이터보다 질적으로(qualitatively) 더 충실한 자료이며, 결국 본 연구의 결과는 정성적으로도 한층 신뢰성이 있는 사회현상 분석을 가능하게 할 수 있다. 둘째, 사회현상을 분석할 때 본 연구의 결과를 바탕으로 비정형 데이터 형태로 작성된 네티즌들의 방대한 의견들을 신속하게 반영할 수 있음으로써 보다 일반화된 사회현상을 규명할 수 있다. 웹2.0 시대에는 사회 문화적으로 네티즌들의 영향력이 매우 크기 때문에 새로운 사회현상을 설명하는 연구모델에는 네티즌들의 의견이 분석되고 반영되어야 그 신뢰성과 가치가 높아질 것이다.

참고문헌

- [1] 김상락 외, 2010, “빅데이터가 여는 미래세상”, 정보과학회지, 30권 6호, 2012.
- [2] Alexander Pak and Patrick Paroub, 2010, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” In Proceedings of the European Language Resources Association.
- [3] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, Chun Jin, 2007, " Red Opal: Product-Feature Scoring from Reviews", Proc. of the 8th ACM conference on Electronic commerce, pp.11-15.
- [4] Courses, E., and Surveys, T., 2008, "Using SentiWordNet for multilingual sentiment analysis", Data Engineering Workshop ICDEW, pp.102-110.
- [5] Guang Giu, Bing Liu, Jiajun Bu and Chun Chen, 2009, "Expanding DomainSentiment Lexicon through Double Propagation,", Proc. of 21th IJCAI-09, pp.1199-1204.
- [6] Ismael S. Silva, JanainaGomide, Adriano Veloso, Wagner Meira Jr., Renato Ferreira, 2011, "Effective Sentiment Stream Analysis with Self-Augmenting Training and Demand-Driven Projection," In SIGIR, pp.475-484.
- [7] J.Leskovec, D.P.Huttenlocher, and J.M.Kleinberg, 2010, "Predicting positive and negative links in online social networks." proc. of 19th WWW, pp.641-650.
- [8] Lei Zhang, Bing Liu, Suk Hwan Lim and Eamonn O'Brien-Strain, 2010, "Extracting and Ranking Product Features in Opinion Documents," Proceedings of the 23rd International Conference on Computational Linguistics, pp.1462-1470.
- [9] Liu, B., Hu, M., and Cheng, J., 2005, "Opinion observer: analyzing and comparing opinions on the Web", Proc. of the 14th international conference on WWW, pp.10-14.
- [10] Melvin R. Crask,1981, "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.
- [11] Minqing Hu and Bing Liu, 2004, "Mining and Summarizing Customer Reviews", KDD'04, pp.168-177.

- [12] Xiaowen Ding, and Bing Lui, 2007, "The Utility of Linguistic Rules in Opinion Mining", SIGR pp.811-812.
- [13] Xiaowen Ding et. al, 2008, "A Holistic Lexicon-Based Approach to Opinion Mining", Proc. of the international conference on web search and web mining, pp. 231-240.
- [14] Yang Yuehua, Du Dunoing, Jia Yingmin, Sun Zengq, 2010, "Study on SNS graph generation and prediction", ICCAS, pp.1188-1191.
- [15] W.Y.Kim, J.S. Ryu, K.I.Kim, U.M.Kim, 2009, "A Method for Opinion Mining of Product Reviews using Association Rules", ICIS, pp.270-274.

국문개요

SNS(Social Network Service)와 스마트폰의 대중화로 인하여 다양한 사회구성원들로부터 대량의 텍스트 데이터가 폭발적으로 생성되고 있다. 이러한 텍스트 데이터는 빅데이터의 유형 중의 하나로서, 사회여론을 파악하거나 새로운 사회현상을 규명하는데 있어 중요한 자료가 될 수 있다.

기존의 사회과학연구방법은 가설을 먼저 설정하고 이를 지지할 수 있는 실증적 자료를 수집하는 연역적 연구방식을 따르지만, 빅데이터 시대에는 이미 존재하는 텍스트 데이터를 분석하여 특정 사회현상이 반영된 패턴을 도출하고 이 패턴을 기반으로 이론을 정립하는 귀납적 연구방식이 사회과학연구의 또 다른 한 축으로 정립될 필요가 있다.

본 논문에서는 빅데이터 상의 텍스트 데이터를 분석함으로써 귀납적 사회과학연구를 지원할 수 있는 오피니언마이닝시스템의 개념을 소개하고 그 기능에 대한 설계도를 제안하고자 한다. 확장형 오피니언마이닝시스템의 주요 기능은 텍스트 데이터를 분석하여 개념 변수를 생성할 수 있어야 하며, 이러한 개념변수에 대한 긍정/부정 의견을 5점 척도로 계산할 수 있어야 한다.

확장형 오피니언마이닝시스템은 사회과학연구의 새로운 패러다임을 도입하는데 기여할 것이며, IT융합의 새로운 적용모델로서 그 가치가 있을 것이다.