SNS 사용자들의 긍정/부정 전파패턴 마이닝시스템 설계

System design for mining positive/negative propagation patterns among users in social network services

김 근 형*
(Kim, Keun-Hyung)

목 차

Ⅰ. 서론

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.

페이스북, 트위터, 네이버밴드 등과 같은 SNS는 기업활동이 해당 소비자에게 미치는 1

* 제주대학교 경상대학 경영정보학과 교수

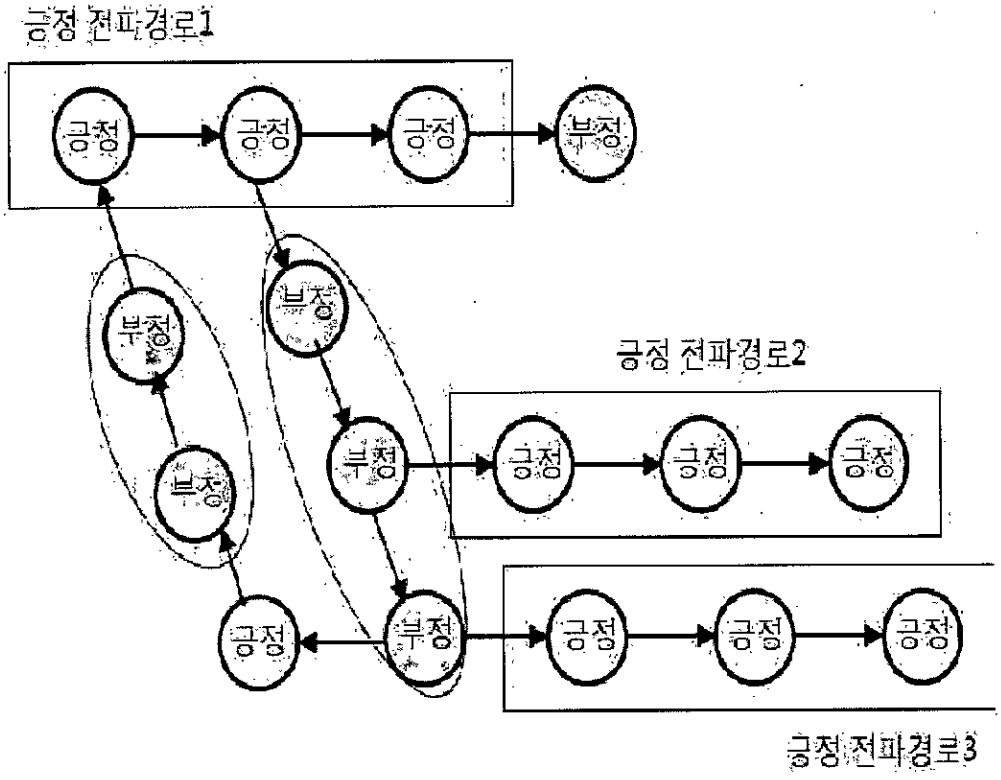
차적인 영향뿐만 아니라 이들 소비자들이 다른 소비자들에게 2차적으로 확산시키는 효과를 빠르고 폭넓게 파악할 수 있게 한다. 트위터가 인기를 끌고 있는 일본에서는 트위터에서의 반응을 성과지표로 활용하여 1일 단위로 광고 슬로건이나 매장진열을 변경하는 사례가 점차적으로 증가하고 있다. NEC 빅로브(Biglobe)가 조사한 자료에 의하면 트위터를 업무에 활용하는 기업 가운데 관련 데이터를 분석하여 효과를 측정한 기업은 전체의 0.9%에 불과 하지만 40%가 결과에 만족하고 80%가 재이용을 희망 하고 있다(채승병, 2011). 이러한 결과는 SNS상의 데이터 분석이 그 어느 때 보다도 필요하고 중요한 것임을 보여주는 사례라 할 수 있다.

SNS가 유용한 이유는 다양한 주제의 많은 텍스트 게시물들을 포함하고 있다는 것이지만, 또 다른 중요한 이유 중의 하나는 상이한 개인프로파일을 갖는 다양한 성향의 사람들이 의견을 게시한다는 점이다. 어떤 성향의 사람이 의견을 게시했는지 표면적으로 드러나지는 않지만, SNS 운영자는 회원가입 시의 사용자 등록정보를 이용하여 의견 게시자의 속성이나 특징 등을 파악할 수 있다. 이러한 상황에서, SNS 사용자네트워크 상에서 긍정적인 의견들의 전파경로가 어떻게 되며, 이러한 경로에 포함된 사람들의 속성들은 어떻게 되는지, 어떤 속성들이 긍정적 의견을 전파하는 패턴에 포함되는지 파악할 수 있다면 이를 통하여 입소문 마케팅 전략을 효과적으로 수립할 수 있다. 해당 기업의 제품에 대하여 단순히 긍정적인 고객보다 긍정 의견을 전파하는 고객이 기업제품의 홍보에 중요한 기여를 할 것이라는 측면에서 그 가치가 더 클 수 있다. 마찬가지로, 부정적 의견을 전파하는 고객 또한 기업입장에서는 중요한 고객정보가 될 수 있다.

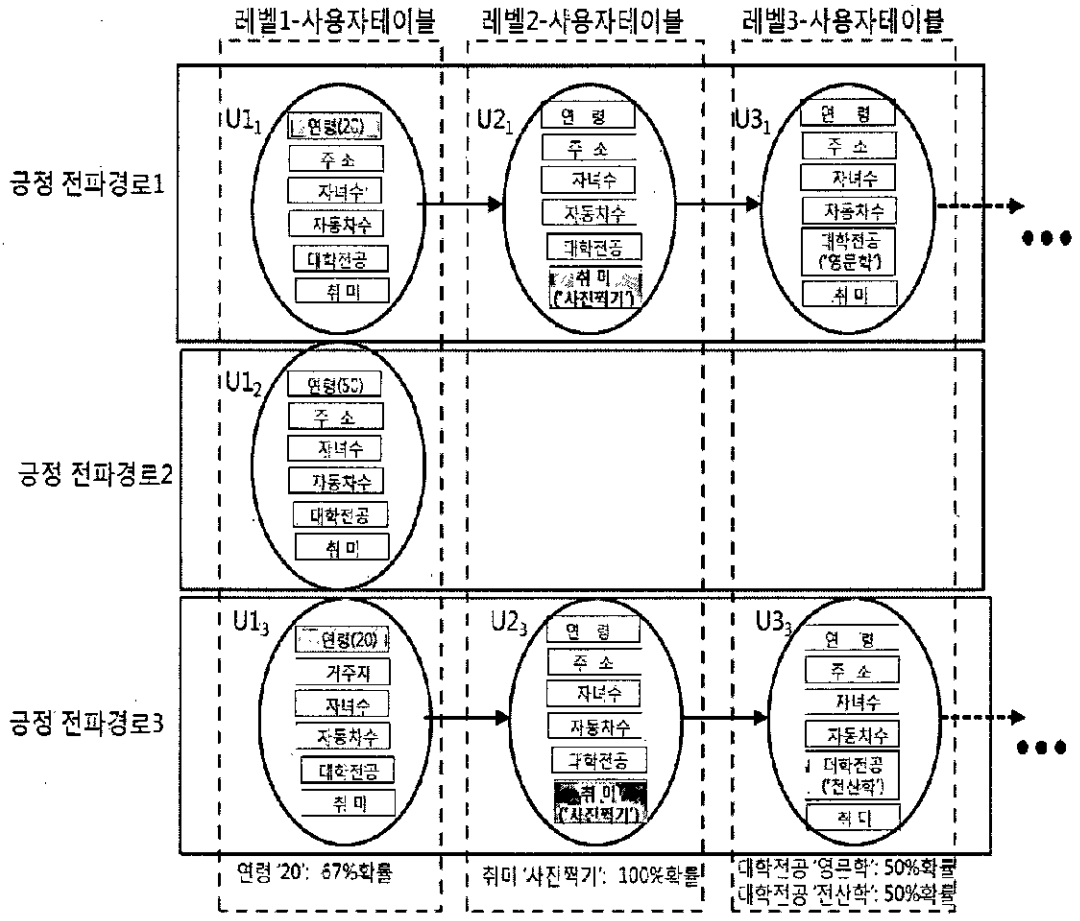
예를 들어, 카메라 기능이 개선되고 관련 앱이 탑재된 새로운 스마트폰이 출시되었을 경우, '20대'의 조기혁신수용자들이 반응을 보이면서 SNS상에서 긍정적인 의견을 게시할 수 있다. 이때 의견을 게시한 조기혁신수용자들과 연결된 다른 SNS사용자들 중에서 사진찍기가 취미인 사용자들이 긍정적인 의견을 게시할 수 있다. 비슷한 방식으로, 취미가 '사진찍기' 인 사용자들과 연결된 다른 SNS사용자들 중에서 전공이 '전산학' 인 사람들이 긍정적인 반응을 나타낼 수 있다. 이러한 과정들이 반복되면서 긍정의견들은 SNS를 통하여 전파될 수 있다. 이때, 긍정/부정 전파패턴에서의 연속출현 속성은 「연령('20대')->취미('사진찍기')->대학전공('전산학')」과 같이 나타나며, 이러한 속성에 해당하는 사람들을 대상으로 한 마케팅 전략을 수립할 수 있다.

<그림1>과 <그림2>는 이러한 내용을 그림으로 표현하여 나타내고 있다. <그림1>은 SNS의 사용자들 사이에서 특정제품에 대하여 긍정적인 반응을 보인 사람과 부정적인 반응을 보인 사람들 사이의 연결관계를 나타내고 있다. 원은 SNS사용자(리뷰작성자)를 의미하고, 직사각형으로 둘러싸인 부분은 긍정의견의 전파경로이며, 타원형으로 둘러싸인 부분

은 부정의견의 전파경로이다. <그림2>는 최소지지도가 50%일 때, 긍정 전파경로에서 연속적으로 연결된 사용자속성그룹들을 마이닝하여 도출한 전파패턴의 예를 나타내고 있다. 새롭게 출시된 스마트폰에 대해서 연령이 20대인 사람들은 67%의 확률로 최초의 긍정적인 반응을 보였다. 긍정적인 반응을 보인 20대 사용자들과 연결된 사람들 중에서는 취미가 '사진찍기'인 경우 100%의 긍정적인 반응을 나타냈다. 취미가 '사진찍기'인 긍정적 반응을 나타낸 사람들과 연결된 사람들 중에서는 대학전공이 '전산학'이거나 '영문학'인 사람들이 50%의 확률로 긍정적인 반응을 나타내고 있다. 레벨1-사용자테이블에서 사용자 U12와 같은 사람의 출현빈도(약 33%)는 최소지지도(50%)에 못 미치고 있기 때문에 사용자 U12와 연결된 사용자는 분석대상에서 제외되고 있다.



<그림1> 긍정/부정 전파경로 예



긍정 전파경로에서 연속적으로 연결된 속성그룹들을 마이닝하여
도출된 전파패턴 규칙

연령이 '20대'인 사람들은 약 67%의 확률로 긍정적인 반응을 보였으며,
이 사람들과 연결된 사람들 중에서 취미가 '사진찍기'인 사람들은
100%의 확률로 긍정적인 의견을 전파했으며,
이 사람들과 연결된 사람들 중에서 대학전공이 '전산학' 인 사람들은
50%의 확률로 긍정적인 반응을 보였다.

<그림2> 긍정 전파경로를 마이닝한 전파패턴의 예(최소지지도 50% 가정)

본 논문에서는 SNS 빅데이터에서 긍정/부정 의견이 전파되는 패턴을 파악할 수 있는 새로운 마이닝 방법을 제안하고자 한다. 이를 위하여 첫째, 데이터웨어하우징 방법 중의 하나인 스타큐빙 알고리즘의 문제점을 분석하고 이를 개선한 새로운 S-스타큐빙 알고리즘을 제안하며, 둘째 S-스타큐빙 알고리즘을 기반으로 한 긍정/부정 전파패턴 마이닝 시스템을 설계하고자 한다.

Ⅱ. 선행연구

SNS 웹사이트 상에 게시된 리뷰들에 대한 내용분석은 오피니언마이닝과 관련되며, SNS사용자속성들에 대한 분석은 데이터마이닝기술이나 데이터웨어하우징 기술과 관련된다.

2.1 오피니언마이닝 관련

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. Liu(2005)는 기계학습 및 자연어처리기술을 활용하여, 온라인고객리뷰 데이터에 대한 감성분석과 분석결과의 요약기법을 제시하고 있으며, Opinion Observer라는 시스템을 개발하였다. 미국 카네기멜론 대학교에서는 Redopal 시스템을 개발한 사례가 있으며(Christopher Scaffidi, et. al, 2007), 이는 고객리뷰 데이터와 사용자 평가점수를 활용하여 요약보고서를 생성하는 시스템이었다. Xiaowen Ding(2007)은 문장구조와 문장 사이의 관계, 문장성분의 패턴정보 등과 같은 언어규칙을 이용하면서 통계학적 방법으로 오피니언마이닝에 접근하고 있다. Courses(2008)은 워드넷을 활용하여 어휘의 긍정이나 부정적 의미를 판단하고 이를 센티워드넷(SentiwordNet)으로 응용하여 감정의 폭을 정량화하는 방법을 제시하고 있다. W.Y.Kim(2009)는 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 이모티콘을 활용하여 텍스트의 긍정적 또는 부정적 감정을 인식하도록 하였는데, 이모티콘 기반의 감정분류 성능은 70~80% 사이의 정확도를 보였다.

SNS는 매 순간 엄청난 수의 사용자가 이용하기 때문에 긍정/부정 오피니언의 변화가 지속적으로 일어날 수 있다. 이와 관련하여 최신의 데이터를 반영하면서 효율적으로 분석결과를 업데이트하는 방법이 제안되었다(Ismael S. Sliva, 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 큰 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 기반하여 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 제품속성단어와 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 제품속성과 대응하는 명사를 정확하게 찾는 방법이라든가, 긍정/부정을 나타내는 감성단어를 정확하게 분류하는 방법 등과 관련된 것들이었다. 덕분에 오피니언마이닝의 분석 정확도는 많이 개선되었다.

그러나, SNS환경에서 오피니언 게시자의 다양한 성향을 고려한 연구들은 거의 없음을 알 수 있다. 특히, SNS환경에서 긍정/부정 의견의 전파경로를 추출하거나 긍정/부정 전파경로에 포함된 사람들의 연속적(consecutive) 속성출현 규칙들 ,즉 전파패턴을 마이닝하는 연구는 전무한 실정이다.

2.2 데이터마이닝 관련

SNS 빅데이터에서 긍정/부정 전파경로 상의 사용자들에 대한 데이터는 정형적 구조를 갖지만 전통적인 데이터마이닝 기법들인 연관규칙탐사기법(박종수, 1998; R.Agrawal, et. al, 1994)이나 의사결정나무기법(Rajeev Rastigi, et. al, 1998), 순차패턴탐색기법(R.Srikant, 1996; J.Pei, 2001) 등을 적용하기에는 한계가 있다. 그 이유를 다섯 가지 관점에서 설명할 수 있다. 첫째, 전통적인 데이터마이닝기법은 하나의 정형 테이블에만 적용

되는 것을 가정하고 있지만, 긍정/부정 전파경로 상의 사용자 데이터는 여러 개의 테이블이 연결되어 있는 구조를 갖는다. 둘째, 연관규칙탐사기법에서 도출한 규칙모형은 긍정/부정 전파패턴과 상이한 구조를 갖는다. 셋째, 의사결정나무기법은 분류속성을 필요로 하지만 긍정/부정 전파경로 상의 사용자들에 대한 데이터에는 분류속성이 없다. 즉, 긍정/부정 전파경로 상의 분석대상 테이블은 모두 긍정이거나 부정인 데이터에 속한다. 넷째, 순차패턴탐색기법에서 시퀀스(sequence)항목구조는 불규칙적이지만 긍정/부정 전파경로 상에 있는 사용자데이터의 시퀀스 구조는 규칙적이다. 다섯째, 순차패턴탐색기법의 시퀀스에서 탐색대상 항목들은 바로 이웃하여 출현하지 않아도 되지만, 긍정/부정 전파패턴에 포함되는 항목들은 바로 이웃하여 연속적으로 나타나야 한다. 본 연구에서 '연속적'이라는 말은 '이웃하여 연속적으로 이어진다는 의미'로서, 순서를 지켜서 나타난다는 의미의 '순차적'이라는 말과 구분하여 사용하기로 한다. <표 1>은 전통적 데이터마이닝기법의 이러한 한계점을 정리하여 나타내고 있다.

이러한 관점에서 긍정/부정 전파패턴을 마이닝 하기 위한 새로운 알고리즘이 개발될 필요성이 있다.

<표 1> 전통적 데이터마이닝 기법과 전파패턴마이닝기법의 차별성

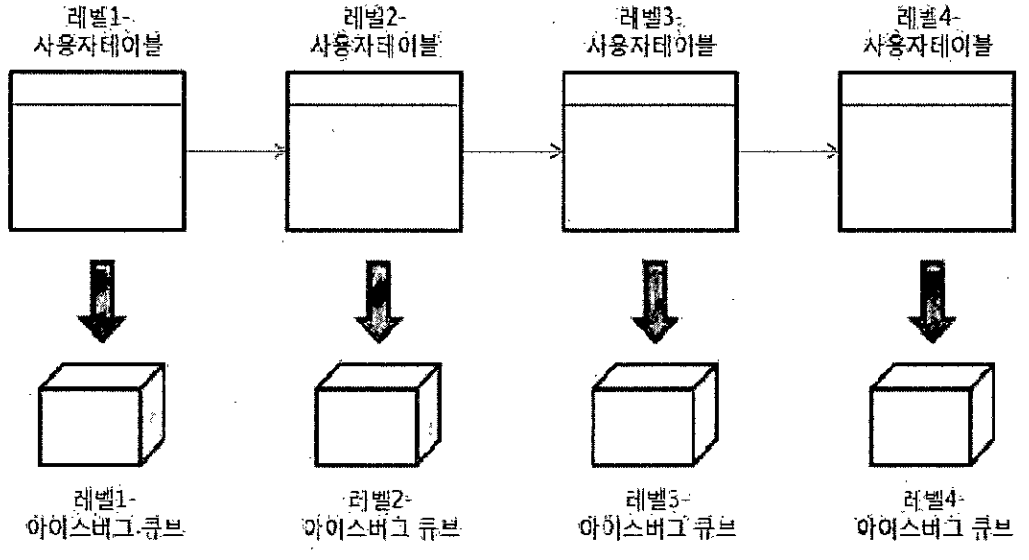
기법	생성모형	대상 테이블 수	분류속성	비고
연관규칙탐사	연관규칙(A→B)	단일	무	전통적 데이터마이닝 기법
의사결정나무	분류모형	단일	유	
순차패턴탐사	순차적 출현 규칙	단일	무	
전파패턴 마이닝	연속적 출현 규칙	다중	무	새롭게 제안할 마이닝 방법

2.3 데이터웨어하우징 관련

긍정/부정 전파경로 상의 전파패턴을 마이닝하기 위하여 데이터웨어하우스 구축기술 중의 하나인 큐브생성기법을 사용할 수 있다(Y. Zhao, 1997; K.Beyer, 1999; Dong Xin, 2003). 긍정/부정 전파패턴은 긍정/부정 경로 상에서 연속적으로 연결된 사용자데이터의 속성관계에 의하여 표현된다. 전파패턴에 포함되는 속성값들은 해당 속성값의 출현횟수(지지도)에 의하여 결정되므로 MOLAP (Multi- dimensional OLAP) 큐브생성 시

사용되는 지지도 계산방법을 이용할 수 있다. 생성된 큐브를 통하여 다차원 관점의 다양한 분석을 수행할 수 있다.

큐브를 효율적으로 생성하기 위하여 불필요한 계산과 중복계산을 피하는 것이 필요하다. 큐브생성 기법 중에서 Multiway방법은 Top-Down방식을 통하여 중복계산을 피하기 위한 동시집계과정(simultaneous aggregation)과정을 제안하고 있지만 불필요한 계산과정을 생략하지는 못한다(Y. Zhao, 1997). BUC 방법은 Buttom-Up방식을 통하여 불필요한 계산과정을 피하기 위한 가지치기(pruning)가 가능하지만 중복계산을 하는 비효율이 존재한다(K.Beyer, 1999). 스타큐빙(Star-Cubing) 방법은 Multiway와 BUC의 장점을 수용하는 기법을 제안하고 있다(Dong Xin, 2003). 즉, 공유차원(shared dimensions)과 스타노드(star node)의 개념을 사용하여 불필요한 계산과 중복계산을 피할 수 있는 방안을 제시하고 있다. 그러나, 스타큐빙기법 역시 단일 테이블에 적용하는 것을 가정하고 있으며 연속적으로 이웃하여 연결된 다중테이블에 적용될 때는 불필요한 계산을 해야 하는 비효율이 발생한다.



<그림3> 스타큐빙기법을 연속적으로 연결된 다중 테이블에 적용할 때의 문제점

<그림3>은 이러한 개념을 나타내고 있다. <그림3>에서 각 사용자테이블은 연속적으로 이웃하여 레벨 별로 연결되어 있다. 이때 스타큐빙기법이 각 사용자테이블에 독립적으로 적용되어 아이스버그큐브(Iceberg Cube)를 생성하게 된다. 아이스버그큐브는 특정

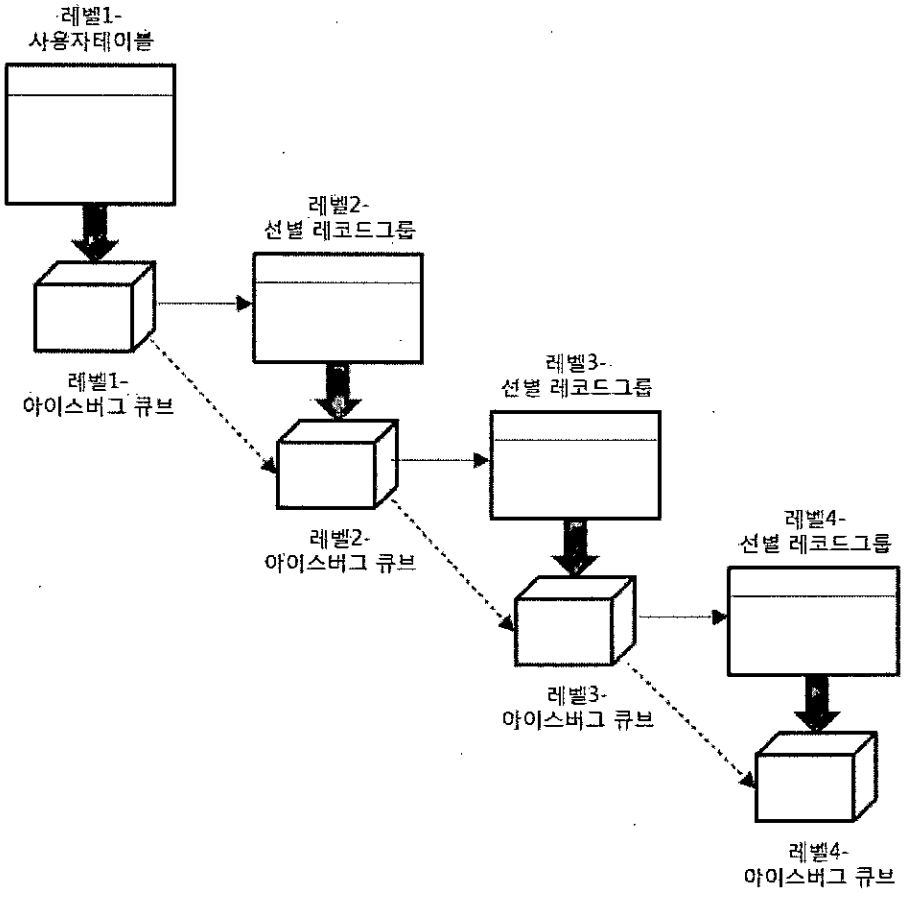
조건(최소지지도 또는 출현횟수)을 만족하는 속성값들을 모아놓은 다차원 테이블이다. 레벨1 아이스버스큐브를 생성한 후, 레벨2 사용자테이블의 모든 레코드들을 대상으로 스타큐빙기법을 적용하여 레벨2 아이스버그큐브를 생성하게 된다. 이때 불필요한 계산 과정이 발생하게 된다. 왜냐하면 레벨2 사용자테이블의 모든 레코드들을 처리 대상으로 할 필요가 없기 때문이다. 레벨2 사용자테이블의 레코드들 중에서 긍정/부정 전파패턴에 포함될 속성값들은 레벨1 아이스버그큐브의 속성값과 연관된 레코드들, 즉 특정조건을 만족하는 레코드들 에만 존재한다. 레벨3, 레벨4 등의 처리에서도 마찬가지이다.

따라서, 긍정/부정 전파패턴을 마이닝 하기 위하여 스타큐빙기법을 적용하고자 할 때 연속적으로 연결된 다중테이블에 적용할 수 있도록 자료구조와 알고리즘이 수정되어야 한다.

Ⅲ. S-스타큐빙 알고리즘의 제안

3.1 S-스타큐빙 알고리즘 개념

S-스타큐빙 알고리즘은 연속적으로 이웃하여 연결된 다중테이블에 적용될 수 있도록 기존 스타큐빙 알고리즘을 수정한 것이다. 긍정/부정 전파패턴은 연속적으로 연결된 아이스버그큐브로부터 추출된다. 아이스버그큐브는 아이스버그조건을 만족하는 속성값과 그 출현횟수 등의 데이터를 포함한다. 연속적으로 연결된 아이스버그큐브는 서로 연관되어 있다. 첫 번째레벨1-아이스버그 큐브는 레벨1-사용자테이블로부터 생성된다. 레벨2-아이스버그는 레벨1-아이스버그큐브와 연관된 레코드들(레벨2-사용자테이블로부터 선별된 레코드들)로부터 생성된다. 기존 스타큐빙기법이 레벨2-사용자테이블에 포함된 모든 레코드들을 처리 대상으로 삼는 반면, S-스타큐빙기법에서는 선별된 레코드들만을 처리대상으로 하기 때문에 불필요한 계산을 피할 수 있다. <그림4>는 S-스타큐빙 알고리즘의 개념을 나타내고 있다. 아이스버그큐브와 연관된 선별레코드를 추출할 수 있도록 기존 스타큐빙 알고리즘의 자료구조와 처리절차 등이 수정되어야 한다.



<그림4> S-스타큐빙 알고리즘의 개념

3.2 S-스타큐빙 알고리즘의 설계

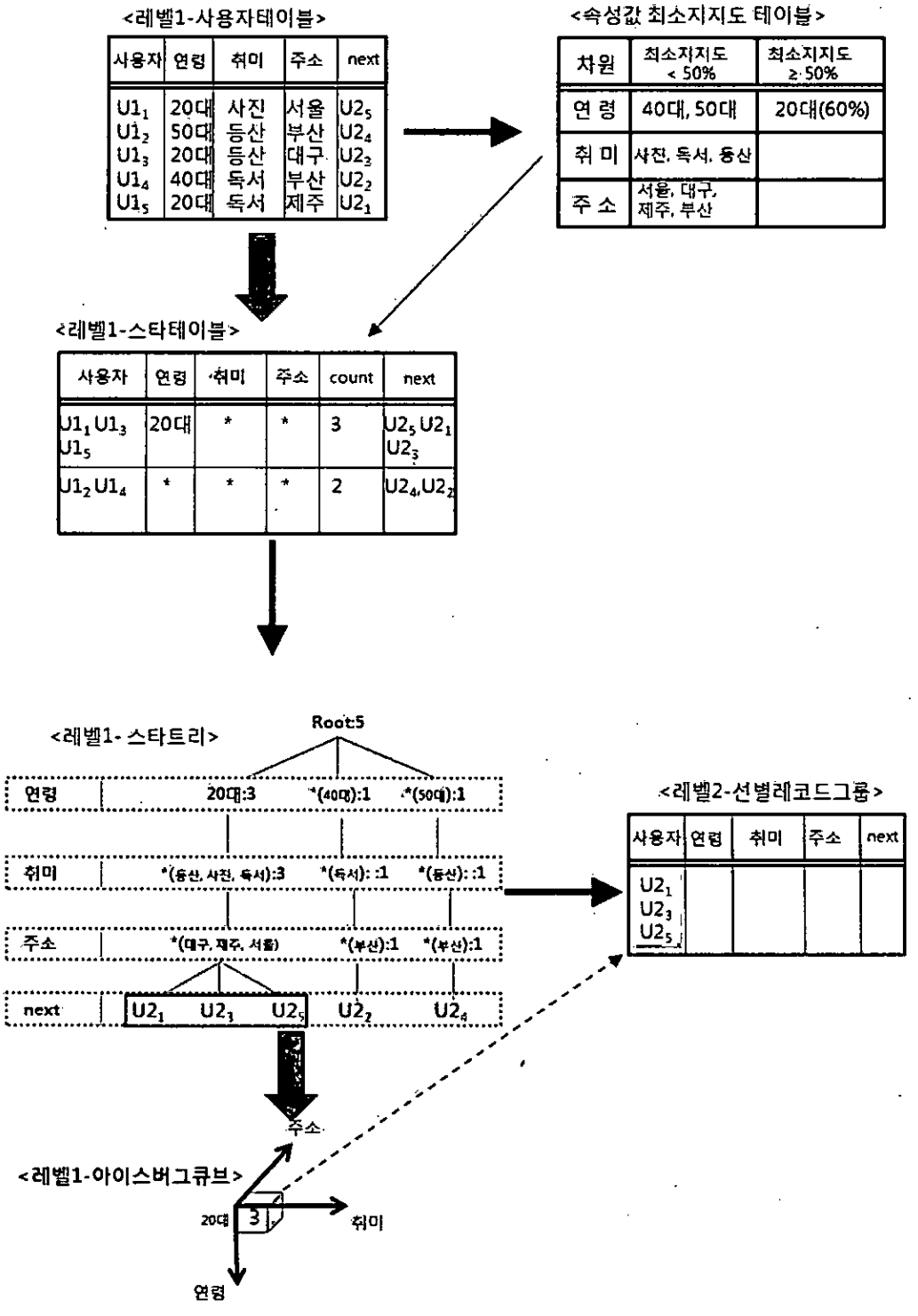
<그림5>는 S-스타큐빙 알고리즘의 개략적인 설계내용을 나타내고 있다. <그림5>에서 <레벨1-사용자테이블>은 사용자 테이블의 처음 상태를 의미한다. <레벨1-사용자테이블>에 있는 각 사용자 레코드들은 긍정/부정 의견이 전파된 다음 사용자 레코드와 연결되어 있다. 이때, <선별레코드그룹>은 긍정/부정 의견을 받은 사용자들로 이루어진 테이블을 의미한다. <선별레코드그룹>은 <레벨1-사용자테이블>의 부분집합이 되고, 알고리즘의 반복명령이 수행될수록 <선별레코드그룹>의 크기는 작아지며 결국 어느 순간에 <선별레코드그룹>은 존재하지 않게 되어 반복명령을 종료된다. 반복명령이 수행되는 동안에 각 <선별

레코드그룹>과 대응되는 아이스버그 큐브가 생성되며 각 아이스버그 큐브들은 연속적으로 연결되어 다양한 관계식을 생성할 수 있는 기반이 된다.

```
〈선별레코드그룹〉 ← 〈레벨1-사용자테이블〉
while (〈선별레코드 그룹〉이 존재)
 begin
   단계1: 〈선별레코드그룹〉들을 읽는다.
   단계2: 〈선별레코드그룹〉으로부터 〈스타테이블〉을 생성한다.
   단계3: 〈스타테이블〉로부터 〈스타트리〉를 생성한다.
   단계4: 〈스타트리〉를 DFS(Depth First Search) 방법으로 탐색하면서 〈아이스버그조건〉
을 만족하는 〈아이스버그큐브〉와 〈next선별레코드그룹〉을 생성한다.
 end
```

<그림5> S-스타큐빙 알고리즘의 설계 내용

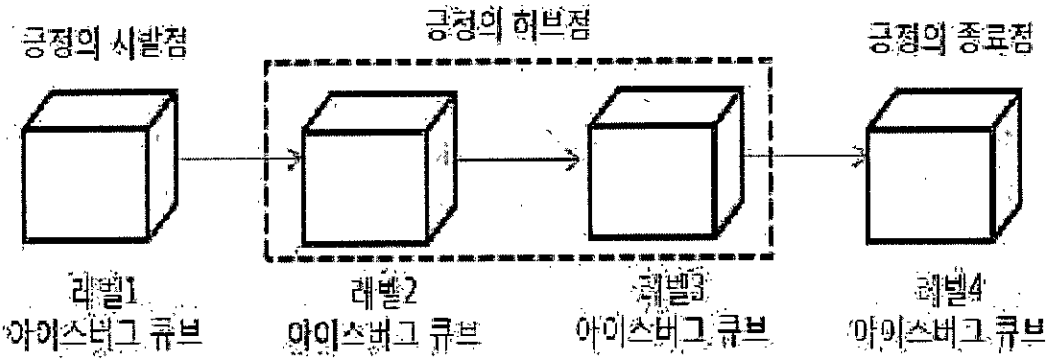
<그림6>은 S-스타큐빙 알고리즘의 작동과정을 실제적인 예를 통하여 나타내고 있다. 즉, 5명의 사용자들에 대한 <레벨1-사용자테이블>이 주어졌고, 아이스버그조건이 최소지지도 50% 이상으로 설정되어 있을 때, S-스타큐빙 알고리즘의 첫 번째 반복과정이 어떻게 작동하는지를 나타내고 있다. 먼저 <레벨1-사용자테이블>이 스캔되어 <속성값 출현횟수 테이블>이 생성되고 이로부터 <레벨1-스타테이블>이 만들어진다. <레벨1-사용자테이블>에서 아이스버그조건에 만족하지 않는 속성값들은 <레벨1-스타테이블>에서 스타(*)로 표시된다. 아이스버그조건을 만족하지 않는 속성값들은 아이스버그큐브의 차원 값으로 존재할 필요가 없기 때문이다. <레벨1-스타테이블>로부터 <레벨1-스타트리>가 생성되며 이 스타트리를 DFS에 의하여 탐색하면서 <레벨1-아이스버그큐브>가 만들어진다. <레벨1-아이스버그큐브>에는 <아이스버그조건>이 만족되는 속성값들에 대해서만 계산된 집계값이 포함된다. 즉, '20대' 속성값에 대한 집계값만 포함한다. 특히, <레벨2-선별레코드그룹> 테이블에는 '20대' 속성값과 연관된 사용자들인 U21, U23, U25 만이 포함되어 다음 반복처리과정에서 처리된다.



<그림6> S-스타큐빙 알고리즘의 처리과정 예

3.3 전파패턴 마이닝을 위한 관계식 도출 방법

S-스타큐빙 알고리즘으로부터 생성된 연속연결 아이스버그큐브들로부터 의미있는 전파패턴을 마이닝하기 위한 관계식들을 도출할 수 있다. MOLAP에서의 관계식은 자주 사용되거나 유의미한 정보를 생성하는 큐브기반 질의이다. 예를 들어, <그림7>의 연속연결 아이스버그 큐브에서 첫 번째 큐브는 긍정적 의견을 시작한 사용자들의 특성을 담고 있는 긍정시발점큐브이고, 마지막 큐브는 긍정적 의견을 종료한 사용자들의 특성을 담고 있는 긍정종료점 큐브가 될 것이다. 또한, 긍정적 의견을 퍼뜨리는 긍정허브점에 해당하는 큐브도 있다. 이런 연속연결 아이스큐브로부터 <표2>와 같은 관계식들이 적용될 필요가 있다.



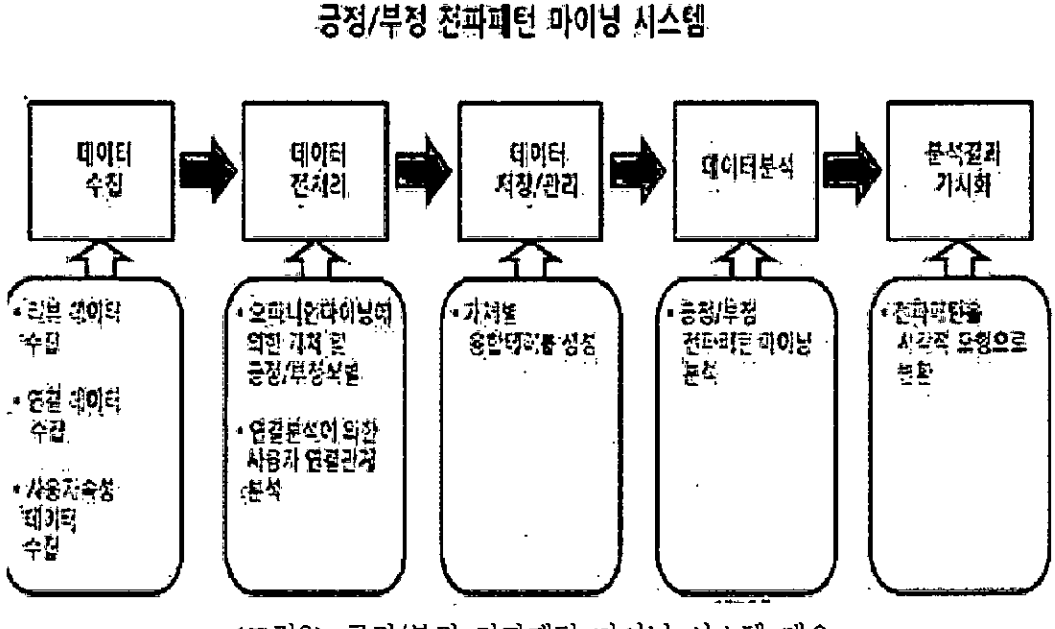
<그림7> 연속연결 아이스버그큐브

<표2> 전파패턴 관계식의 예

번호	관계식 내용	관계식	대상큐브	대상차원
1	긍정의 시발점이 되는 사람과 종료점이 되는 사람의 나이분포 비교	대상큐브와 대상차원, 속성값등을 포함하는 수학계산식	레벨1 아이스큐브, 레벨4 아이스큐브	나이
2	긍정의 시발점이 되는 사람과 종료점이 되는 사람의 직업 비교	...	...	...
3 4	...			

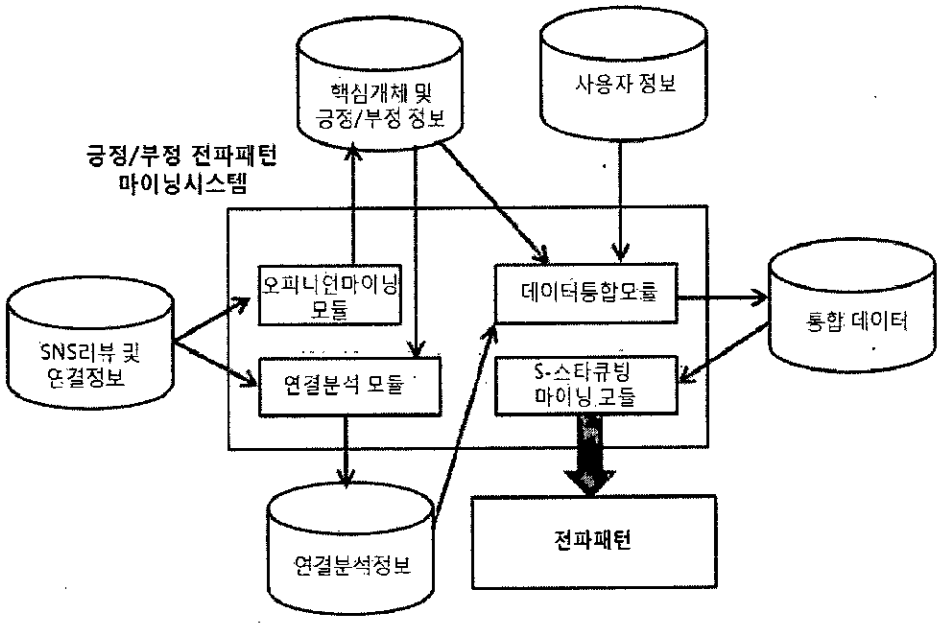
Ⅳ. 긍정/부정 전파패턴 마이닝시스템의 설계

<그림8>은 긍정/부정 전파패턴 마이닝 시스템의 전체적인 기능을 나타내고 있다. 데이터수집 단계에서는 SNS상에 게시된 특정 분야의 리뷰들이 수집되며, 리뷰작성자들 사이의 연결관계 및 사용자속성 데이터가 획득되어야 한다. 데이터전처리 단계에서는 수집된 리뷰내용들을 오피니언마이닝에 의하여 분석하며 핵심개체들과 이들 각 개체에 대한 긍정/부정 여부가 판별된다. 또한, 리뷰작성자들 사이의 연결관계도 분석되어야 한다. 데이터저장/관리 단계에서는 데이터 전처리 단계에서 파악된 개체들과 긍정/부정 여부, 사용자 연결관계 등에 대한 데이터가 관계형 테이블 구조로 생성된다. 데이터분석 단계에서는 해당 관계형 테이블에 대하여 S-스타큐빙기법을 적용하여 긍정/부정 전파패턴 마이닝 분석이 이루어진다. 분석결과 가시화 단계에서는 전파패턴 마이닝 결과를 시각적인 모형으로 변환하여 분석자에게 제공한다.



<그림8> 긍정/부정 전파패턴 마이닝 시스템 개요

<그림9>는 긍정/부정 전파패턴 마이닝시스템의 전체적인 구조를 나타내고 있다. 각 구성요소들은 <그림11>에 나타난 바와 같이, 오피니언마이닝모듈, 연결분석모듈, 데이터통합모듈, S-스타큐빙마이닝모듈이다.



<그림11> 긍정/부정 전파패턴 마이닝시스템 구조

각 모듈 별 세부 기능은 다음과 같다.

4.1 오피니언마이닝 모듈

SNS리뷰에 나타나는 각 문장들은 구문분석기에 의하여 각 단어들에 품사가 부여된 형태의 구문구조트리로 변환된다. 구문구조트리 파일로부터 명사와 형용사에 해당하는 단어들이 추출된다. 출현빈도가 높은 명사에 대하여 $\sum_i$형용사$_i$ • 명암도$_i$를 적용함으로써 긍정/부정 여부를 계산한다. 이를 위하여 형용사 및 각 형용사의 명암도를 나타내는 온톨로지(ontology)도 구축되어야 한다.

4.2 연결분석 모듈

SNS리뷰들의 연결관계 정보와 오피니언마이닝 결과정보를 기반으로 리뷰작성자의 유형을 4가지로 구분하면서 연결분석정보를 생성한다. 리뷰작성자들은 <표3>과 같이 4가지 유형으로 구분된다.

<표3> 리뷰작성자의 4유형

유형구분	의미
긍정(in)–>긍정(out)	다른 사람의 긍정의견을 받아서 긍정적으로 전파하는 유형
긍정(in)–>부정(out)	다른 사람의 긍정의견을 받아서 부정적으로 전파하는 유형
부정(in)–>긍정(out)	다른 사람의 부정의견을 받아서 긍정적으로 전파하는 유형
부정(in)–>부정(out)	다른 사람의 부정의견을 받아서 부정적으로 전파하는 유형

4.3 데이터통합 모듈

오피니언마이닝 결과와 연결분석정보, 사용자(속성)정보를 입력으로 받아 통합한다. 통합테이블 구조는 <표4>와 같다. <표4>에서 음영색으로 표시된 첫 번째 레코드의 의미는 「'가' 지역에 거주하면서 연령은 '25세', 취미가 '독서'인 사람은 SNS사이트에서 'S사제품'에 대한 긍정의견에 댓글을 달았으며, 이 댓글에 긍정의견을 표명한 사람의 번호는 12이다』로 해석할 수 있다.

<표4> 핵심개체 'S사제품' 에 대한 통합테이블 형식

사용자번호	사용자특성			전파유형	link
	연령	지역	취미		
1	25	가	독서	긍정(in)–>긍정(out)	12
2	•••	•••	•••	긍정(in)–>부정(out)	15
3	•••	•••	•••	부정(in)–>긍정(out)	26
4	•••	•••	•••	부정(in)–>부정(out)	•••
•••					

4.4 S-스타큐빙마이닝 모듈

S-스타큐빙마이닝 알고리즘을 <표4>의 테이블구조에 맞게 수정한다.

V. 결론

본 논문에서 제안한 긍정/부정 전파패턴 마이닝시스템은 SNS분석 및 솔루션개발, 빅데이터 처리기술, 데이터웨어하우스 구축, 사회과학연구의 새로운 방법론 등에서 활용될 수 있다.

첫째, 긍정/부정 전파패턴 마이닝시스템은 SNS상의 데이터를 분석하는데 활용될 수 있다. SNS상의 데이터는 사용자 리뷰 등의 비정형데이터와 사용자정보, 연결정보 등으로 구성된 복잡한 구조를 갖기 때문에 기존의 개별 기술로는 종합적인 분석이 불가능하다. 긍정/부정 전파패턴 마이닝시스템은 기존의 개별 기술들을 융복합함으로써 복잡한 구조의 데이터를 분석할 수 있으며 향후, SNS 데이터분석을 위한 상용화제품을 개발하는데 기여할 수 있다.

둘째, 본 연구의 결과는 빅데이터 처리기술을 선도하는데 기여할 수 있다. 빅데이터는 대용량 데이터일 뿐만 아니라 정형/비정형 데이터가 혼재된 형태이기 때문에 기존의 데이터 처리기술들을 융합하여야 분석이 가능하다. 긍정/부정 전파패턴 마이닝시스템은 정형/비정형 데이터를 통합할 뿐 아니라 개별데이터 처리기술을 융합·적용하는 기술이 포함되기 때문에 빅데이터 처리기술이 지향하는 바를 내재하고 있다. 따라서, 긍정/부정 전파패턴 마이닝시스템은 빅데이터 처리기술의 모델로 활용될 수 있다.

셋째, S-스타큐빙 알고리즘은 새로운 유형의 데이터웨어하우스 구축기술을 선도하는데 활용될 수 있다. 본 연구에서 생성할 연속연결 데이터큐브는 기존 데이터웨어하우스에서는 취급하지 않았던 데이터구조로서 빅데이터를 분석하기 위한 차세대 데이터웨어하우스 구조라 할 수 있다. 특히, 데이터큐브 내에서 뿐만 아니라 데이터큐브들 사이에서의 관계식은 새로운 차원의 데이터웨어하우징 기술이라 할 수 있다. 연속연결 데이터큐브나 데이터큐브들 사이의 관계식 등은 데이터웨어하우스 연구영역을 확장하는 시발점이 될 수 있다.

넷째, 리뷰내용과 리뷰작성자를 연계하여 분석하는 기술은 새로운 사회과학 연구방법론으로 사용할 수 있다. 설문조사는 기존의 사회과학연구방법으로 인기가 있었지만, 설문응답자들의 무성의한 답변에 대한 염려 때문에 그 신뢰성에 대한 의구심을 갖는 학자들도 많이 있었다. 그러나 SNS리뷰는 작성자의 흥미와 의지가 있어야 생성될 수 있는 것이기 때문에 설문조사에 의한 데이터보다 더 객관적인 자료가 될 수 있다. SNS 리뷰를 작성한 사람들의 특징을 고려하면서 텍스트 데이터를 분석함으로써 새로운 차원의 사회 현상을 규명할 수 있다.

다섯째, 본 연구에서 개발된 기술에 의하여 대량의 텍스트 데이터를 신속하고 세밀하게 분석할 수 있다. 특히, 네티즌들이 게시한 온라인고객리뷰들에 대하여 기존의 오피년마이닝 기술이 지원할 수 없는 분석기능을 지원함으로써 네티즌들의 성향 및 동향을 세부적으로 분석할 수 있다.

여섯째, 본 연구의 기술에 의하여 웹사이트의 기능을 향상시킬 수 있다. 긍정/부정 전파패턴 마이닝시스템의 오피니언마이닝모듈이 웹사이트 구축 시 탑재된다면, 대부분의 웹사이트들이 운영하고 있는 고객게시판의 고객의견을 실시간으로 분석할 수 있다. 고객의견의 실시간 분석결과는 기업경영자 뿐만 아니라 새로운 고객들에 의해서도 유용하게 활용될 수 있다.

일곱째, S-스타큐빙 알고리즘의 성능평가를 위하여 생성한 실험용 데이터를 공개함으로써 데이터마이닝 기술발전에 일조할 수 있다.

본 논문에서 제안한 시스템은 설계단계에 국한되었으며 구체적인 구현과정이 있어야 실제 활용될 수 있다는 한계가 있다. 본 시스템의 가치를 인정하는 기업이 본 논문에서 제안한 설계도를 바탕으로 산학협력을 통하여 구체적인 개발과정이 진행될 수 있기를 기대해 본다.

참고문헌

강만모 외(2012), "빅 데이터의 분석과 활용", 정보과학회지, 30권 6호.

김상락 외(2012), "빅데이터가 여는 미래세상", 정보과학회지, 30권 6호.

박종수 외(1998), "연관규칙탐사와 그 응용", 정보과학회논문지, 16권, 9호, pp.37-46.

안창원 외(2012), "빅 데이터 기술과 쥬요 이슈", 정보과학회지, 30권 6호.

채승병(2011), "정보홍수 속에서 금맥 찾기: '빅데이터(BigData)' 분석과 활용" 삼성경제연구소 SERI 경영 노트, 제91호, pp.1-12.

Alexander Pak and Patrick Paroub(2010), "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," In Proceedings of the European Language Resources Association.

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.

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

Dong Xin, Jiawei Han, Xiaolei Li and Benjamin W. Wah(2003), "Star-Cubing: Computing Iceberg Cubes by Top-Down and Bottom-Up Integration", VLDB.

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

Hao Ma, Irwin King, and Michael R.LYU(2011), "Learning to Recommand with Explicit and Implicit Social Relations," ACM Trans. on Intelligent Systems and Technology, Vol.2, No.3, pp.29-48.

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.

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, ACM.

J.Pei, J.Han, B.Mortazavi-A, H.Pinto, Q.Chen, U.Dayal and M-C(2001), "Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, ICDE'01, pp.215-224.

K.Beyer and R.Ramarkrishnan(1999), "Bottom-up computation of sparse and iceberg cubes", SIGMOD, pp.359-370.

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.

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.

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

R.Agrawal and R.Srikant(1994), "Fast Algorithm forMining Association Rules", 20th VLDB, pp.487-499.

Rajeev Rastigi, Kyuseok Shim(1998), "PUBLIC:A Decision Tree Classifier that Integrates Building and Pruning", 24th VLDB, pp.404-415.

R.Srikant and R.Agrawal(1996), "Mining sequential patterns: Generations and performance improvement EDBT'96, pp.3-17.

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.

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

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.

Yang Yuehua, Du Dunoing, Jia Yingmin, Sun Zengq(2010), "Study on SNS graph generation and prediction", ICCAS, pp.1188-1191.

Y. Zhao, P.Deshpande, J.F.Naughton(1997), "An Array Based Algorithm form Simultaneous Multidimensional Aggragates", SIGMOD, pp.159-170.