ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 June: 20(2): 93~107 http://dx.doi.org/10.13088/jiis.2014.20.2.093

다계층 이원 네트워크를 활용한 사용자 관점의 이슈 클러스터링

김지은
국민대학교 Business IT 전문대학원
(*skan5@kookmin.ac.kr*)

김남규
국민대학교 Business IT 전문대학원
(*ngkim@kookin.ac.kr*)

조윤호
국민대학교 경영학부
(*www4u@kookmin.ac.kr*)

대부분의 인터넷 쇼핑몰은 자사 고객의 관심 분야를 파악하고 이를 상품 추천에 효과적으로 활용하기 위해 많은 노력을 기울이고 있다. 하지만 고객이 회원 가입 시 직접 입력한 개인 정보는 신뢰하기가 어렵고, 고객의 구매 패턴을 통해 파악한 관심 분야 정보는 자사 사이트 내에 진입한 이후에만 보인 한정된 패턴이라는 측면에서 해당 고객의 다양한 관심 분야를 제대로 나타낸다고 보기 어렵다. 이러한 한계를 극복하기 위해 본 연구에서는 고객의 평소 인터넷 사용 기록을 통해 최근 방문 사이트들의 주제를 분석함으로써, 고객의 실제 관심 분야를 파악할 수 있는 방안을 제시하였다. 또한 토픽 분석을 통해 각 사이트의 주제를 도출하고 도출된 주제를 다시 동시 방문자 관점에서 군집화 함으로써, 고객 관점에서 의미가 있는 상위 수준의 새로운 테마를 발굴하기 위한 방법론을 제안하였다. 연구의 특징은 유사주제 중심의 군집화라는 기존 연구와는 달리 사용자 관점의 관심주제 중심 군집화라 할 수 있다. 향후 사용자 중심의 카테고리 설계를 비롯한 새로운 관점의 고객군 정의 등 보다 높은 차원의 마케팅 전략 수립에 활용이 가능할 것으로 기대된다. 사용자 관점의 이슈 군집화 과정은 크롤링, 토픽 분석, 액세스 패턴 분석, 네트워크 병합, 네트워크 변환 및 군집화와 같은 여섯 가지 주요 단계로 구성되어있다. 이를 위해 텍스트 마이닝과 소셜 네트워크 분석 기법을 활용한 비정형 텍스트를 기반으로한 빅데이터의 활용 방법을 모색하였다. 제안 방법론의 실무 적용 가능성을 평가하기 위해, 국내 최대 포털 뉴스 사이트의 방문자 2,177명의 1년간 방문 기록과 뉴스기사 대한 분석을 수행하고 그 결과를 요약하여 제시하였다.

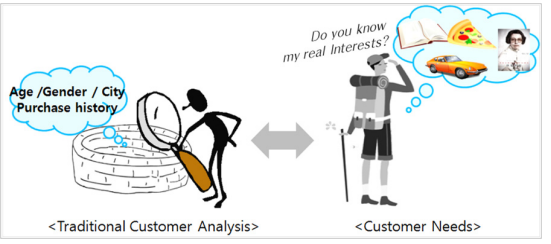
주제어 : 데이터 마이닝, 소셜네트워크 분석, 이슈 클러스터링, 토픽 분석

논문접수일 : 2014년 6월 15일 논문수정일 : 2014년 6월 20일 게재확정일 : 2014년 6월 23일
투고유형 : 학술대회우수논문 교신저자 : 김남규

1. 개요

웹 3.0과 스마트 기기의 일반화라는 흐름에 따라 고객은 자신이 원하는 제품이나 서비스에 대한 정보를 더욱 적극적으로 검색할 수 있는 환경을 갖추게 되었다. 수요자의 제품 정보력 확대는 판매자 관점에서는 기존 고객을 유지하고 새로운 고객을 유치하기 위해 고객의 니즈에 맞는 상품 및 서비스를 개발에 더욱 적극적인 노력을 기울여야 함을 의미한다. 이러한 노력의 일환으로 많은 쇼핑몰들은 각 고객의 특성을 고려하여 상품을 추천하기 위한 추천 시스템을 개발하여 운영하고 있지만, 이들 시스템은 고객이 최초 가입 시 입력한 정보와 자사의 사이트 내에 진입한 이후에 보인 한정된 구매 패턴만을 이용한다는 측면에서 한계를 갖는다. 즉, 온라인 쇼핑몰 운영자가 고객이 자사 사이트 내에서 보인 제한된 패턴만을 토대로 해당 고객의 실제 관심 분야를 제

대로 도출하기란 매우 어려울 것으로 판단된다 <Figure1>.



〈Figure 1〉 Limitations of Identifying Customer Interests in Specific Site

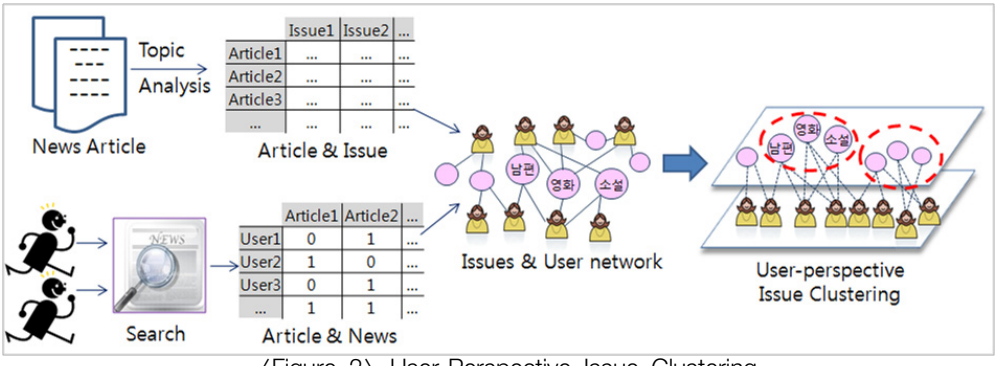
<Figure 1>은 특정 사이트 내의 정보만을 분석하여 고객의 실제 관심 분야를 제대로 파악하는 것은 한계가 있음을 보이고 있다. 이러한 한계를 극복하기 위해 현재 분석이 이루어지고 있는 고객 정보 외에 추가 정보에 대한 분석이 수행되어야 함은 매우 자명하다. 하지만 최근 계속되는 개인정보 유출 사고로 인해 정보 수집에 대한 규제는 점차 강화되는 추세이므로, 고객으로부터 직접적인 정보를 추가로 수집하기란 매우 어려울 것으로 판단된다. 이에 대한 대안으로 고객의 일상 생활 패턴, 특히 인터넷 사용 패턴으로부터 고객에 대한 추가 정보를 획득하는 방안이 모색될 수 있으며, 최근 텍스트 분석을 포함한 빅데이터 분석 기술은 이러한 필요를 충족시킬 수 있을 정도의 성과를 이루어왔다.

빅데이터 분석 기술 중 특히 토픽 분석(Topic Analysis)은 구조화된 정형 데이터를 활용하여 고객 정보를 얻기 위한 기존의 노력과 달리, 웹과 소셜 미디어 등을 통해 급증하고 있는 텍스트 형태의 비구조화된 비정형 데이터를 분석하여 새롭고 유용한 정보를 얻기 위해 확산되고 있다 (Liu, 2012). 본 연구에서는 고객의 최근 인터넷 사용 기록과 함께 방문 사이트들의 주제를 분석함으로써, 고객의 실제 관심 분야를 파악할 수 있는 방안을 제시하고자 한다. 분석 대상으로는 다양한 최신 주제를 다루고 있고 비교적 정제된 표현을 사용하고 있는 것으로 알려진 뉴스 기사를 사용하고자 하며, 수집한 뉴스 기사 전체에 대한 토픽 분석을 통해 기사 별 대응 토픽을 도출하고자 한다.

하지만 토픽 분석은 전적으로 각 뉴스 기사가 포함하고 있는 어휘의 종류 및 빈도수에 기반하여 수행되기 때문에, 실제로 고객 관점에서는 매우 밀접한 연관성을 갖지만 동일한 어휘를 서로 다수 공유하고 있지 않는 기사들을 아우르는 토픽은 발견되기 어렵다는 한계를 갖는다. 예를 들어 세 개의 기사가 각각 “어린이 교육”, “웰빙 먹거리”, 그리고 “거실 인테리어”의 이슈를 다루고 있다고 했을 때, 이들 세 이슈는 서로 많은 어휘를 공유하고 있지 않을 것으로 예상된다. 하지만 위의 세 기사에 모두 접속하는 고객이 다수 발견된다면, 이들 이슈를 아우르는 보다 상위 수준의 테마를 발굴하고 이를 구체화함으로써 향후 마케팅 전략 수립에 활용하는 것이 바람직할 것이다.

이를 위해 본 논문에서는 사용자 정보와 뉴스, 그리고 이슈 사이의 상호 관계를 활용하여 보다 상위 수준의 신규 테마(Theme)를 발굴하기 위한 방법론을 제안하고자 한다<Figure 2>. 구체적으로는 (i) 사용자와 기사간 네트워크 및 기사와 이슈간 네트워크를 각각 구축하고, (ii) 이 두 네트워크를 사용자와 이슈간 네트워크로 병합한 뒤, (iii) 사용자 관점에서의 이슈 클러스터링을 통해 상위 수준의 테마를 발굴하고자 한다.

본 논문의 이후 부분은 다음과 같이 구성된다. 2장에서는 본 연구와 수행을 위한 핵심 기법인



〈Figure 2〉 User-Perspective Issue Clustering

텍스트 마이닝 및 토픽 분석, 그리고 소셜 네트워크 분석에 대한 선행 연구들을 간략히 요약하고, 3장에서는 본 연구의 핵심인 사용자 관점의 이슈 클러스터링 방법론을 소개하고, 제안 방법론을 실제 데이터에 적용한 실험의 결과를 분석한다. 마지막 5장에서는 본 연구의 기여 및 한계 그리고 향후 발전 방향을 제시한다.

2. 관련연구

2.1 텍스트 마이닝 및 토픽 분석

텍스트는 사용자들이 인터넷이나 스마트 기기를 통해 정보를 표현하고 획득하는 가장 일반적인 방식이다(Witten, 2004). 이렇게 생성된 방대한 양의 텍스트에 대한 분석을 통해 의미 있는 정보를 추출하기 위한 연구가 다양한 관점에서 활발하게 수행되고 있으며, 특히 많은 문서로부터 여러 토픽을 추출하고 어떤 문서가 어떤 토픽에 해당되는지 식별하는 기법인 토픽 분석에 대한 관심이 점차 높아지고 있다. 더구나 스마트 기기가 보편화되면서 일반 사용자들이 소셜 미디어를 통해 각자의 성향을 텍스트 형태로 표현하고 이를 공유함에 따라, 소셜 미디어에 나타난 텍스트 분석을 통해 비즈니스 전략을 수립하기 위한 다양한 시도들(Choi, 2012; Kim, 2012)도 활발하게 이루어지고 있다. 토픽 분석은 텍스트 데이터를 다루는 대부분의 분야에 적용될 수 있으며, 온라인 쇼핑몰의 상품평 분석(Liu, 2012; Myung et al., 2008), 범죄 예측(Fan et al., 2006), 텍스트 범주화를 통한 리파지토리 구축(Sebastiani, 2006) 등이 그 예가 될 수 있다.

토픽 분석은 데이터 마이닝, 자연어 처리, 정보 검색, 전산 언어학, 토픽 추적 등 여러 분야의 기술을 종합적으로 활용한다. 특히 자연어 처리 기술은 토픽 분석에서 중요한 역할을 하는 핵심이라고 할 수 있으며, 자연어 처리의 대상이 되는 텍스트는 분석 목적에 따라 행렬, 계층, 벡터 등의 다양한 형태로 표현될 수 있다(Stanvrianou et al., 2007). 분석의 최소 단위는 각 문서가 되는데, 문서는 제목, 요약, 본문, 등 텍스트로 기술된 모든 데이터를 일컫는 폭넓은 개념으로 사용된다. 기본적으로 각 문서는 벡터공간모델(Vector Space Model)(Albright, 2006)을 이용하여 표현되며, 각 문서에 사용된 용어(Term)의 빈도에 따라 해당 문서의 주제 및 특성이 요약된다. 대부분의

경우 용어의 단순 빈도수보다는 TF-IDF(Term Frequency-Inverse Document Frequency)에 근거한 분석이 널리 활용된다. 이 개념은 어떤 문서 X에서 용어 A와 B가 동일한 빈도수로 발생하였을 때, A가 다른 문서들에서도 일반적으로 자주 발생하는 용어라면 문서 X에서 더 중요하게 사용되는 용어는 A가 아니라 B라는 인식을 반영한다. 빈도수에 기반한 분석에서 각 문서는 용어 수만큼의 차원을 갖게 되며, "(문서 수) × (용어 수)"로 표현된 행렬의 각 셀에 각 문서에서 해당 용어가 나타난 빈도수를 기재함으로써 모든 문서를 행렬로 구조화할 수 있다. 하지만 문서에 포함된 용어의 수는 일반적으로 매우 많기 때문에, 문서간 유사성 측정을 위해 각 문서는 SVD(Singular Value Decomposition) 등의 차원 축소기법을 통해 저장된다(Albright, 2006).

이러한 과정을 통해 비정형 텍스트 문서의 구조화가 완료되면, 이후 분석 과정은 전통적인 데이터 마이닝에서 사용되어 온 주요 기법들을 활용하여 수행할 수 있다. 특히 각 문서 벡터에 대한 클러스터링(Clustering)을 통해 유사 문서를 그룹화하는 기법은 이미 여러 영역에서 사용되고 있다. 하지만 전통적인 문서 클러스터링의 경우 각 문서가 하나의 클러스터에만 속하는 경우를 가정하고 있기 때문에, 하나의 문서가 다양한 주제로 분류되는 실제 상황과 부합하지 않는다는 한계를 갖는다. 이와 달리 토픽 분석은 문서와 토픽간의 다대다 관계를 허용함으로써, 복합 주제를 갖는 하나의 문서가 여러 토픽에 대응되는 현상을 잘 반영할 수 있다.

2.2 소셜 네트워크 분석과 구조적 지표

소셜 네트워크 분석은 개인이나 집단을 하나의 노드(Node)로, 그리고 각 노드들간의 관계를 링크(Link)로 표현하여, 개체간 연결 형태나 구조를 다양한 계량 지표로 분석하거나 도식화하여 시각화하는 분석기법(Cho and Kim, 2011; Kim, 2007; Kwak, 2014)을 의미한다. 초기에는 주로 사람들간의 관계 분석을 위해 사용되었으나, 점차 유전 네트워크(Kauffiman, 1993), 교통 네트워크(Yoon, 2005), 조직 네트워크(Choi, 2006) 등 다양한 분야의 구조 분석에 매우 활발하게 사용되고 있다. 특히 최근에는 지식 이전 요인 분석(Kang and Hau, 2012), 키워드 관계 분석(Cho and Kim, 2011), R&D 정보 패키징(Hyun et al., 2013) 등 다른 기법과의 접목을 통해 새로운 지식을 창출하는 영역으로 그 응용 범위가 확대되고 있다.

소셜 네트워크 상 연결 구조의 특성을 파악하기 위해 다양한 지표가 고안되었으며, 밀도(Density), 중심성(Centrality), 집중도(Centralization) 등이 널리 활용되고 있다(Kwak, 2014). 밀도는 네트워크 노드들 사이의 연결된 정도를 의미하며, 네트워크의 밀도가 높다는 것은 정보의 교류가 활발하여 정보의 확산이 빠름을 의미한다. 중심성은 특정 행위자가 전체 네트워크에서 중심에 가까이 위치하는 정도를 나타내는 지표로, 연결 정도 중심성, 근접 중심성, 매개 중심성 등으로 세분화된다. 또한 집중도는 네트워크 전체가 한 노드로 집중되는 정도를 표현하는 지표로써, 연결 정도 집중도, 근접 집중도, 매개 집중도로 구분된다. 연결 정도 집중도는 각 노드간 연결 정도에 따라 전체 네트워크의 집중화 정도를 측정하는 것이며, 근접 집중도는 각 노드간의 거리에 기반하여 전체 네트워크의 집중화 정도를 측정하는 지표이다. 그리고 매개 집중도는 각 노드의 매개성을 기반하여 전체 네트워크의 집중화 정도를 측정하는 지표이다(Cho and Kim, 2011; Kim, 2007;

Kwak, 2014).

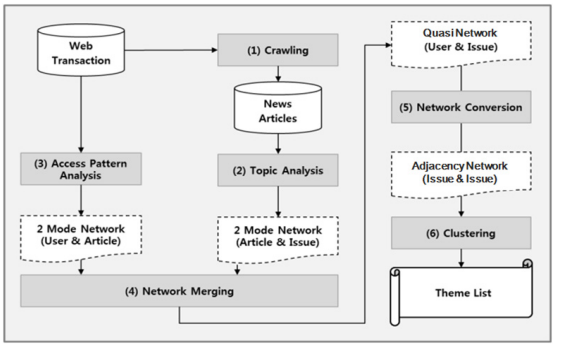
소셜 네트워크 분석을 위해 UCINET, NetMiner, R, 그리고 NodeXL 등의 도구가 자주 사용된다. 이들 도구는 데이터를 행렬로 표현하며, 행과 열의 개체간 관계가 존재하면 1, 그렇지 않은 경우에는 0으로 입력하거나 계량 값으로 표현한다(Kwak, 2014). 또한 개체간 직접적인 관계를 바탕으로 간접적인 관계를 도출하여 준 네트워크(Quasi Network) (Hong, 2013; Kwak, 2014)을 생성할 수 있다. 예를 들어 고객과 상품의 구매 관계를 구매(1), 비구매(0)로 표현한 고객/상품 행렬의 경우, 동일한 제품을 1개 이상 구매한 고객들을 연결하는 방식으로 고객간 준 네트워크를 구성할 수 있다. 본 연구에서는 사용자와 뉴스간의 이원 네트워크(2 Mode Network)와 뉴스와 이슈간의 이원 네트워크를 구축하고, 이를 병합하여 사용자와 이슈간의 준 네트워크를 구축한 뒤, 이를 이슈간의 인접 네트워크(Adjacency Network)로 변환하여 분석에 활용한다.

3. 사용자 관점의 이슈 클러스터링

3.1 제안모형 및 연구범위

본 절에서는 사용자 관점의 이슈 클러스터링을 통해 상위 수준의 신규 테마를 발굴하기 위한 방법론을 제시한다. 본 연구의 전체 개요는 <Figure 3>과 같다.

본 연구는 웹을 통해 확보한 사용자별 웹 이용 기록(Web Transaction)을 기본 자료로 활용한다. (1) Crawling 단계에서는 방문 기록 중 뉴스 사이트에 해당되는 페이지의 내용을 수집하고, (2) Topic Analysis 단계에서는 수집된 페이지에 대



〈Figure 3〉 Research Overview

한 토픽 분석을 통해 이슈와 각 기사간의 대응 관계를 2 Mode Network으로 정의한다. (3) Access Pattern Analysis 단계에서는 사용자와 기사간의 방문 관계를 2 Mode Network으로 나타내며, (4) Network Merging 단계에서는 (2)번과 (3)번 단계의 결과물로 나타난 두 네트워크를 병합하여 사용자와 이슈간의 준 네트워크(2 Mode Network)을 구성한다. 병합된 네트워크는 (5) Network Conversion을 통해 이슈들의 인접 네트워크로 변환되며, (6) Clustering 단계에서는 이에 대한 클러스터링을 통해 상위 테마를 정의한다.

본 장의 이후 절에서는 위 과정에 대한 자세한 내용을 실제 데이터에 대한 실험과 함께 소개한다.

3.2 실험 개요

본 연구의 제안 방법론의 실무 적용 가능성을 검증하기 위한 실험에서는 국내 한 인터넷 사이트 순위 분석 전문 업체인 ‘R’사에서 수집한 패널 5,000명의 2012년 7월부터 2013년 6월까지의 웹 사용 기록 약 1억 5천만건을 사용하였다. 전체 사용 기록 중 대형 인터넷 뉴스 포털 사이트인 'N'사의 방문 기록 234,776건을 추출하였으며, 이들 중 “생활문화” 카테고리에 해당되는 13,652

건을 다시 추출하여 이들 기사에 대한 크롤링을 수행하였다.

DOC_ID	DESCRIPTION
118633	임형섭 기자 = 한국에서 크루즈 관광을 즐기는 중국인 여행객들이 하루에 평균
118659	'수지 블루베리 총각 짝사랑'.수지가 블루베리 총각을 짝사랑했었다고 깜짝 고
118655	[오마이뉴스 김두나 기자].한국성폭력상담소는 성(性)에 대한 나와 우리사회의
50058	[한겨레21] [문화] 신간 할인을 제한폭 10%로 좁히고 규제 범위 구간까지 확대
50051	오는 28일 합격 축하와 친목의 자리 마련.교육기업 에듀윌이 '제15회 주택관리
50053	대놓고 금빛보단 은은한 핑크·로즈골드가 대세.낡은 듯 멋스러운 빈티지 스타일
50003	결혼식(자료사진).(서울=연합뉴스) 과거 여성 룸으로 인식돼온 사소한 집안일이

〈Figure 4〉 A Snapshot of "Life-Culture" Articles

크롤링 결과는 <Figure 4>와 같이 문서 번호와 기사 전문의 두 컬럼으로 저장되었으며, 저장된 기사 13,552건 중 한 건이라도 방문한 패널의 수는 2,177명으로 나타났다.

3.3 토픽 분석

본 단계에서는 데이터 마이닝 상용 도구 중 하나인 SAS Enterprise Miner 12.1의 Text Miner 모듈을 사용하여 기사 13,552건에 대한 토픽 분석을 수행하였다. 전체 토픽의 수는 100개, 토픽 별 키워드의 수는 5개로 한정하였으며, 그 결과는 <Figure 5>와 같은 형태로 나타난다. 토픽 분석의 과정은 이미 많은 연구에서 소개되었으므로 본 논문에서는 자세히 다루지 않기로 한다.

Topicid	DocCutoff	TermCutoff	Name
1	0.602	0.024	기온,내일,지방,날씨,전국
2	0.555	0.023	엔진,모델,연비,출력,주행
3	0.535	0.022	태풍,볼라,볼라벤,초속,기상청
4	0.405	0.021	컬러,스타일,패션,아이템,원피스
5	0.489	0.021	수입,모델,벤츠,시장,브랜드
6	0.470	0.020	연구팀,연구,건강,결과,정직한 지식
7	0.455	0.021	마을,숲,여행,풍경,코스
8	0.502	0.022	사람,남편,친구,생각,마음
9	0.368	0.019	택시,정부,대중교통,지원,택시업계

〈Figure 5〉 A Snapshot of Topic Analysis

토픽 분석의 결과 기사와 이슈간의 다대다 관계가 생성되며, 각 기사가 각 이슈에 대해 나타내는 대응도가 파악된다. 초기 대응도는 <Figure 6(a)>와 같은 형태로 나타나며, 특정 임계값 이상의 대응도를 갖는 경우 '1', 그렇지 않은 경우를 '0'으로 변환하여 각 기사와 각 이슈간 대응 여부를 <Figure 6(b)>와 같이 나타낼 수 있다.

Doc.	Topic_1	Topic_2	Topic_3
1	0.045	0.790	0.011
2	0.105	0.650	0.048
3	0.779	0.039	0.063
4	1.223	0.068	0.270
5	0.040	0.027	0.042
6	0.277	0.153	0.244
7	0.632	0.109	0.172
8	0.090	0.077	0.073
9	0.050	0.026	0.041
10	0.046	0.009	0.074

(a)Article–Issue Initial Matrix

Doc.	Topic_1	Topic_2	Topic_3
1	0	0	1
2	1	0	1
3	0	1	0
4	0	1	0
5	0	0	0
6	0	0	0
7	0	1	0
8	0	0	0
9	0	0	0
10	0	0	0

(b)Article–Issue Refined Matrix

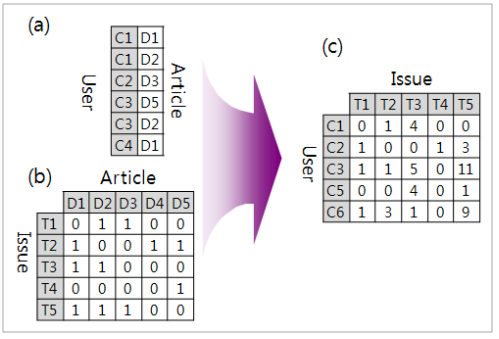
〈Figure 6〉 A Partial Matrix of Articles-Issues

3.4 방문 패턴 분석 및 네트워크 병합

본 절에서는 방문 패턴 분석을 통해 사용자와 기사간 이원 네트워크를 생성하고, 이를 병합하여 사용자와 이슈간 이원 네트워크를 생성하는 과정을 소개한다. 본 연구에서 매트릭스는 이원 네트워크와 일대일로 대응되는 동일한 역할을 수행한다. 다만 이원 네트워크의 경우 매우 복잡하게 나타날 뿐 아니라 일부를 추출하여 설명하기 어려운 측면이 있으므로, 동일한 개념인 매트릭스를 사용하여 분석 과정을 설명한다.

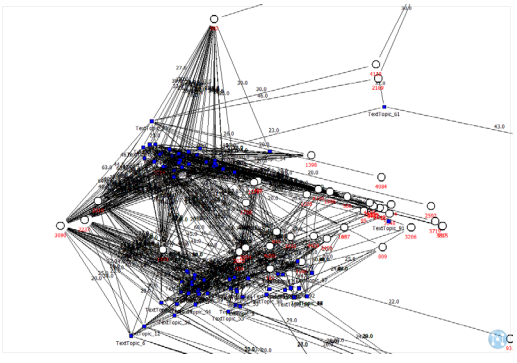
<Figure 7>은 두 개의 이원 네트워크를 병합하여 하나의 이원 네트워크로 만드는 과정을 매트릭스를 통해 설명하고 있다(Hong, 2013). <Figure 7(a)>는 사용자와 기사간의 대응 관계를 보여주며, <Figure 7(b)>는 기사와 이슈간의 대응 관계

를 보여주고 있다. 여기서 각 사용자에게 대응되는 기사의 건수를 이슈별로 합산한 결과가 <Figure 7(c)>에 나타나있다. 즉 <Figure 7(c)>는 각 사용자가 각 이슈에 해당하는 기사 중 몇 건을 방문했는지를 나타내고 있다. 이렇게 생성된 최종 매트릭스에 대해 특정한 임계값 이상을 갖는 경우를 '1'로, 그렇지 않은 경우를 '0'으로 변환함으로써 사용자와 이슈간의 이원 네트워크를 구성할 수 있다.



〈Figure 7〉 Quasi-Network Generation

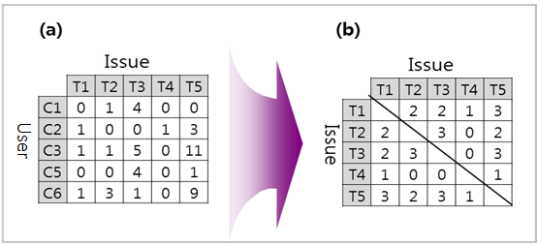
<Figure 8>은 이러한 과정을 통해 구성된 사용자와 이슈간 이원 네트워크의 실제 예를 보여준다. 그림에서 밝은 원형 노드는 사용자를, 어두운 사각형 노드는 이슈를 나타내며, 연결 임계값으로는 '2'를 사용하였다.



〈Figure 8〉 Two-mode Network of Users-Issues

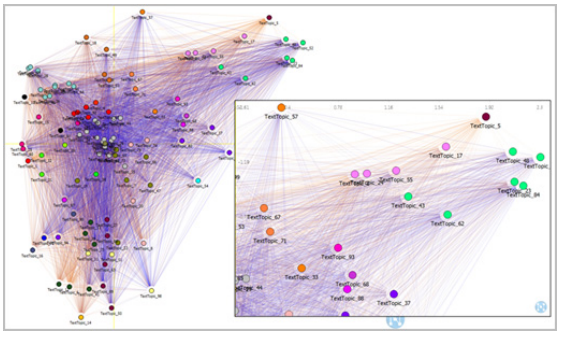
3.5 네트워크 변환을 통한 준 네트워크 생성

본 절에서는 사용자와 이슈간 이원 네트워크를 변환하여 이슈간 준 네트워크를 생성하는 과정을 소개한다(Kwak, 2014). 앞의 절과 마찬가지로 설명의 용이성을 위해 네트워크의 변환 과정을 매트릭스를 통해 설명한다<Figure 9>.



〈Figure 9〉 Adjacency-Network Generation

<Figure 9(a)>는 <Figure 7(c)>와 동일한 내용이며, 사용자와 이슈간의 대응 관계를 나타낸다. 여기서 이슈 T1에 대응되는 사용자 세 명 중 이슈 T2에도 대응되는 사용자는 두 명(C3, C6) 존재함을 알 수 있다. 즉 이슈 T1과 T2는 두 명의 사용자에 의해 공통 방문되었으며, 이러한 방식으로 모든 이슈간 공통 방문자의 수를 구할 수 있다. 이를 정리한 결과가 <Figure 9(b)>에 나타나있다. <Figure 9(b)>에서 셀 값이 클수록 해당



〈Figure 10〉 Adjacency-Network of Issues

셀에 대응되는 두 토픽의 관련성은 높은 것으로 파악될 수 있다. <Figure 10>은 이렇게 생성된 최종 매트릭스에 대해 이슈간 준 네트워크를 구성한 실제 예를 보여준다.

3.6 이슈 클러스터링을 통한 상위 테마 도출

본 절에서는 사용자와 이슈간 관계를 토대로 구성된 이슈간 준 네트워크에 대한 클러스터링을 수행한다. 대상 네트워크의 링크는 이와 연결된 두 노드에 해당되는 이슈에 공통적으로 접근한 사용자가 많음을 의미한다. 따라서 이렇게 생성된 네트워크에 대한 클러스터링 결과는 다수의 사용자에게 동시 접근되는 경향이 있는 이슈들의 군집을 의미하게 된다.

본 실험에서는 클러스터링을 두 가지 방식으로 수행하고 그 결과를 비교하도록 한다. 우선 준 네트워크 생성 이전의 사용자와 이슈간 대응 매트릭에 대한 클러스터링을 수행한다. 즉 <Figure 9(a)>와 같은 형태의 매트릭스에 대한 행/열 변환을 수행한 후, 변환된 행렬에 대해 SAS Enterprise Miner 12.1 상에서 클러스터링을 수행한다. 두 번째 방법으로는 <Figure 10>과 같은 이슈간 인접 네트워크에 대한 클러스터링을 수행하고자 하며, 이 는 소셜 네트워크 분석 및 시각화 도구인 NetMiner 4를 이용하여 수행한다.

CUSID	Cus_13	Cus_17	Cus_25	Cus_26	Cus_28	Cus_29	Cus_30	Cus_31	Cus_32	Cus_33
Topic_1	0	0	0	0	0	0	0	0	0	0
Topic_2	0	0	0	0	3	0	0	0	0	0
Topic_3	6	0	25	0	0	0	8	0	0	0
Topic_4	2	0	0	5	0	0	0	0	0	0
Topic_5	0	0	0	0	3	0	0	0	0	0
Topic_6	22	0	0	0	0	0	0	0	0	0
Topic_7	0	0	0	0	0	0	3	0	0	0
Topic_8	6	0	0	9	0	0	2	0	0	0
Topic_9	0	0	0	0	0	0	0	0	0	0
Topic_10	0	0	0	5	0	0	0	0	0	0
Topic_11	8	0	0	0	0	0	0	0	0	0
Topic_12	7	0	0	2	0	0	0	0	0	0
Topic_13	8	0	0	0	0	0	0	0	0	0
Topic_14	2	0	0	0	0	0	0	0	0	0
Topic_15	8	0	6	0	0	0	5	0	0	0

〈Figure 11〉 Topic-User Matrix for SAS Clustering

우선 사용자와 이슈간 대응 매트릭스에 대해 SAS Enterprise Miner를 사용하여 수행한 클러스터링 과정을 소개하면 다음과 같다. 클러스터링 수행을 위해 100개의 이슈 행과 2,177개의 사용자 열로 구성된 매트릭스를 생성한다<Figure 11>. 이 매트릭스에서 '2' 이상의 값은 '1'로, '2' 미만의 값은 '0'으로 변환한 매트릭스를 도출하고, 이를 클러스터링의 입력 테이블로 사용하였다.

Segment	Theme	Topic ID	Topic Keywords
7	봄 여가활동	Topic_7	마을,숲,여행,풍경,코스
		Topic_19	재료,소금,다진,양파,마늘
		Topic_22	맛,음식,식당,메뉴,재료
		Topic_47	공간,건축,가구,건물,건축가
		Topic_56	관광객,관광,여행,외국인,일본
		Topic_65	축제,벚꽃,자전거,행사,봄꽃
		Topic_73	여행,상품,투어,여행사,호텔
		Topic_100	구장,야구장,부산,응원단,주중
12	건강지식	Topic_6	연구팀,연구,건강,결과,정직한 지식
		Topic_12	다이어트,체중,음식,칼로리,운동
		Topic_13	통증,운동,근육,허리,관절
		Topic_39	암,환자,유방암,수술,검사
		Topic_41	환자,질환,고혈압,위험,병원
		Topic_75	커피,카페인,음료,섭취,mg
17	날씨	Topic_1	기온,내일,지방,날씨,전국
		Topic_26	,구름,많음,구름조금,흐리고
		Topic_31	영하,기온,한파,추위,평년
		Topic_32	기상청,평년,㎜,고기압,강수량
		Topic_38	담배,금연,흡연,흡연자,금연구역

〈Figure 12〉 Clustering Results by SAS E-Miner

SAS 클러스터링에서는 클러스터의 수를 20개로 설정하였다. 물론 클러스터의 수가 결과에 미치는 영향이 매우 크므로 클러스터의 수를 결정하기 위한 사전 실험이 반드시 필요하다. 하지만 사전 실험을 통해 본 연구의 목적을 보이기에 적합한 클러스터의 수를 발견하지 못했다. 따라서 대안으로 하나의 클러스터에 포함되는 평균 이

슈의 수를 기준으로 클러스터의 수를 20개로 임의로 지정하였다. 이렇게 도출된 클러스터링 결과의 일부가 <Figure 12>에 나타나있다.

클러스터링 결과 대부분 유사한 이슈들이 하나의 클러스터를 형성하고 있었지만, 직접적인 연관성이 없는 것으로 보이는 이슈들이 하나의 클러스터로 묶인 경우도 다수 발견되었다. 이들 중에는 이러한 이슈들의 묶음으로 상위 수준의 테마를 정의할 수 있는 경우도 있었으며, 적당한 테마를 정의할 수 없는 경우도 존재했다.

예를 들어 <Figure 12>에서 7번 클러스터의 경우 “여행”, “음식”, “축제”, “스포츠관람” 등 서로 직접적인 연관성을 갖지 않는 이슈들로 구성되어 있다. 하지만 이들 이슈는 “봄 여가활동”이라는 상위 테마로 통합될 수 있다. 즉 단편적인 토픽 분석을 통해서는 발견할 수 없었던 이슈를 사용자 관점, 즉 공통 방문자의 수 관점에서 추가 분석을 실시함으로써 새롭게 발견할 수 있었다. 또한 12번 클러스터의 경우 “다이어트”, “운동”, “환자”, “커피” 등의 이슈로 구성되었다. 이 경우 “건강지식”이라는 새로운 테마가 도출된 것으로 판단할 수도 있지만, 이들 이슈간에는 충분한 관련성이 존재하므로 일반적인 토픽 분석의 결과와 크게 다른 결론을 도출했다고 보기는 어렵다. 마지막으로 17번 클러스터의 경우 대부분 “날씨”에 대한 이슈로 구성되어 있으나 “흡연”에 대한 이슈도 함께 포함되어 있다. 이처럼 모든 이슈를 아우르는 유의미한 신규 테마를 정의하고 설명하기 어려운 클러스터도 일부 파악되었다.

다음으로 소셜 네트워크 분석 도구인 NetMiner를 통한 클러스터링을 진행하였다. <Figure 13>은 각 이슈간의 근접성을 유클리드 거리 (Euclidean Distance)로 측정한 매트릭스를 나타내며, <Figure 10>의 네트워크에 대응된다.

		1	2	3	4	5	6
		TextTopic_1	TextTopic_2	TextTopic_3	TextTopic_4	TextTopic_5	TextTopic_6
1	TextTopic_1		160.8	108.6	141.6	196.7	164.5
2	TextTopic_2	160.8		146.8	135.1	101.8	185.7
3	TextTopic_3	108.6	146.8		123.0	185.1	166.3
4	TextTopic_4	141.6	135.1	123.0		174.4	154.7
5	TextTopic_5	196.7	101.8	185.1	174.4		214.2
6	TextTopic_6	164.5	185.7	166.3	154.7	214.2	
7	TextTopic_7	105.3	146.8	123.7	119.3	184.3	149.2
8	TextTopic_8	119.1	175.4	152.7	152.5	203.9	135.5
9	TextTopic_9	136.7	126.2	122.4	105.9	169.9	163.2
10	TextTopic_10	142.9	146.9	121.5	104.4	184.6	164.7
11	TextTopic_11	148.6	176.2	154.7	136.5	205.9	84.2
12	TextTopic_12	168.1	184.3	169.4	151.9	214.9	59.3

〈Figure 13〉 Distance Matrix between Issues

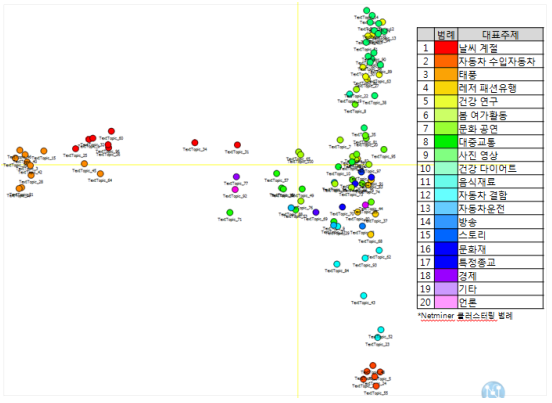
<Figure 13>의 매트릭스에 대해 NetMiner에서 제공되는 PAM (Partitioning Around Medoids) 알고리즘을 적용하여 20개의 클러스터를 도출한 결과가 <Figure 14>에 나타나있다.

Segment	대표 주제	Topicid	Name
1	날씨 계절	Topic_1	기온,내일,지방,날씨,전국
		Topic_25	눈,폭설,중부,지방,많은 눈
		Topic_26	,구름,많음,구름조금,흐리고
		Topic_31	영하,기온,한파,추위,평년
		Topic_32	기상청,평년,㎜,고기압,강수량
		Topic_60	폭염,더위,열대야,기온,최고기온
		Topic_96	수면,밤,잠,습관,한국
5	건강 연구	Topic_6	연구팀,연구,건강,결과,정직한 지식
		Topic_12	다이어트,체중,음식,칼로리,운동
		Topic_27	남성,여성,결혼,조사,남녀
		Topic_41	환자,질환,고혈압,위험,병원
		Topic_50	연구,결과,연구팀,실험,박사
		Topic_63	알코올,음주,소주,농도,혈중
		Topic_86	운동,자전거,걷기,커피,사람
6	봄 여가활동	Topic_7	마을,숲,여행,풍경,코스
		Topic_22	맛,음식,식당,메뉴,재료
		Topic_47	공간,건축,가구,건물,건축가
		Topic_56	관광객,관광,여행,외국인,일본
		Topic_65	축제,벚꽃,자전거,행사,봄꽃
		Topic_73	여행,상품,투어,여행사,호텔
		Topic_76	여수,엑스포,박람회,관람객,조직
		Topic_78	솔로,대첩,솔로대첩,행사,여의도

〈Figure 14〉 Clustering Result by NetMiner

<Figure 14>의 클러스터 1, 5, 6은 각각 <Figure 12>의 클러스터 17, 12, 7에 해당되며, 실험 결과 두 도구를 사용한 클러스터링의 결과에는 큰 차이가 없는 것으로 나타났다. <Figure 14>의 결과를 다차원 척도법 (Multidimensional

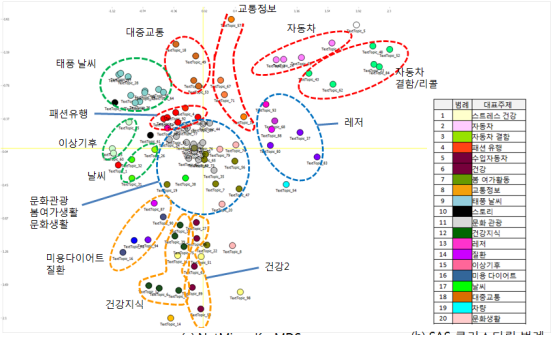
Scaling, MDS)을 통해 시각화한 결과가 <Figure 15>에 나타나있다.



〈Figure 15〉 Visualized Result of NetMiner Clustering

<Figure 15>는 20개의 클러스터를 색상으로 구분하여 나타내고 있다. 즉 동일한 클러스터는 동일한 색상으로 표현되며, 각 클러스터에 해당되는 상위 수준의 테마명은 그림의 우측 상자에 나타나있다. 테마명은 각 클러스터를 구성하는 이슈들을 근거로 도출한 것이다.

마지막으로 <Figure 16>은 SAS 클러스터링의 결과를 NetMiner의 다차원 척도법으로 도식화한 결과를 보여준다.



〈Figure 16〉 Plotting SAS Clusters Using NetMiner

<Figure 16>는 <Figure 12>의 결과, 즉 SAS 클러스터링의 결과를 색상으로 구분한 것이다. <Figure 16>는 전체 이슈를 NetMiner의 Kn-MDS로 표현한 것으로, 그래프 상에서의 색상은 SAS의 클러스터링 결과를 참조하여 수동으로 지정한 것이다. 즉 동일한 색상의 노드가 그래프에서 인접하여 나타난 부분은 SAS와 NetMiner의 결과가 유사하게 나타난 것을 의미하여, 동일한 색상인데도 멀리 떨어져서 나타난 노드는 SAS에서는 같은 클러스터로 묶였지만 NetMiner에서는 서로 다른 클러스터로 분류된 것이다. 두 도구를 통한 클러스터링 결과에 큰 차이가 없음은 본 그림을 통해서도 확인할 수 있다.

4. 결론

본 연구에서는 고객의 평소 인터넷 사용 기록을 통해 최근 방문 사이트들의 주제를 분석함으로써, 고객의 실제 관심 분야를 파악할 수 있는 방안을 제시하였다. 또한 토픽 분석을 통해 각 사이트의 주제를 도출하고 도출된 주제를 다시 동시 방문자 관점에서 군집화함으로써, 고객 관점에서 의미가 있는 상위 수준의 새로운 테마를 발굴하기 위한 방법론을 제안하였다. 제안 방법론의 실무 적용 가능성을 평가하기 위해, 국내 최대 포털 뉴스 사이트의 방문자 2,177명의 1년간 방문 기록 13,652 건에 대한 분석을 수행하였다. 실험 결과 유사한 이슈들이 하나의 클러스터를 형성하는 경우뿐 아니라, 직접적인 연관성이 없는 것으로 보이는 이슈들이 하나의 클러스터로 묶이면서 상위 수준의 새로운 테마를 정의하는 경우도 다수 발견할 수 있었다. SAS E-Miner와 NetMiner를 사용한 두 가지 클러스터링을 수

행하였으며, 비교 결과 두 도구를 통한 클러스터링 결과에는 큰 차이가 없음을 확인하였다.

본 연구의 기여는 크게 두 가지 측면에서 확인할 수 있다. 우선 본 연구에서는 고객이 실제로 방문하여 조회한 웹 페이지에 대한 분석을 통해 해당 고객의 관심 분야를 식별할 수 있는 방안을 제시하였다. 즉 뉴스 기사에 대한 토픽 분석을 실시하여 각 기사가 어떤 이슈에 속하는지를 파악하고, 어떤 사용자가 어떤 이슈의 기사를 방문했는지를 분석함으로써 사용자별 관심 이슈를 식별할 수 있었다. 다음으로 본 연구에서는 다수의 이슈들을 사용자 관점, 즉 공통 방문자의 수 관점에서 추가 분석을 실시함으로써, 직접적인 토픽 분석을 통해서는 파악하기 어려운 상위 수준의 신규 테마를 발굴하기 위한 방안을 제시하였다. 제안 방법론은 향후 사용자 중심의 카테고리 설계, 새로운 관점의 고객군 정의 등 보다 높은 차원의 마케팅 전략 수립에 활용될 수 있을 것으로 기대한다.

하지만 본 연구는 다음과 같은 측면에서 향후 보완이 이루어질 필요가 있다. 우선 토픽 분석 이전의 전처리 단계에서 사용되는 불용어 사전의 품질을 향상시킬 필요가 있다. 반복 실험을 통해 불용어 사전의 품질을 향상시킴으로써, 결과로 도출되는 이슈 키워드를 신뢰할 수 있을 것이다. 또한 본 연구에서 클러스터링 결과를 토대로 신규 테마를 발굴하는 과정은 전적으로 연구자의 주관적 견해에 의해 이루어졌다. 향후 이 과정을 자동화, 또는 반 자동화하기 위한 시도가 이루어져야 할 것이다.

참고문헌 (References)

Albright, R., *Taming Text with the SVD*, SAS Institute Inc., 2006.

Cho, I. and N. Kim, "Recommending Core and Connecting Keywords of Research Area Using Social Network and Data Mining Techniques," *Journal of Intelligence and Information Systems*, Vol.17(2011), 127～138.

Choi, C., "Research on Informal Organizational Network: Social Network Analysis," *Korea Society and Public Administration*, Vol.17, No.1(2006), 1～23.

Choi, K., "Social Big Data Analysis," *Proceedings of the Spring Workshop on Korea Intelligent Information System Society*, (2012).

Fan, W., W. Wallace, S. Rich, and Z. Zhang, "Tapping the Power of Text Mining," *Communications of the ACM*, Vol. 49, No. 9(2006), 76～82.

Hong, S., *Social Network World and Big Data Applications*, Powerbook, Seoul, 2013.

Hyun, Y., H. Han, H. Choi, J. Park, K, Lee, K-Y. Kwahk, and N. Kim, "Methodology Using Text Analysis for Packaging R&D Information Services on Pending National Issues," *Journal Of Information Technology Applications & Management*, Vol.20(2013), 231～257.

Kang, M., and Y. S. Hau, "Multi-level Analysis of the Antecedents of Knowledge Transfer: Integration of Social Capital Theory and Social Network Theory," *Asia Pacific Journal of Information Systems*, Vol.22(2012), 75-97.

Kauffiman, S. A., *The Origins of Order*, Oxford University Press, Oxford, 1993.

Kim, I., "The Value of Big Data and Strategy," *2012 Big Data Search Analysis Technology*,

Insight, 2012.

Kim, Y. H., *Social Network Analysis*, Seoul, 2007.

Kwak, K. Y., *Social Network Analysis*, Cheongram, Seoul, 2014.

Liu, B., *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.

Myung, J., D. Lee, and S. Lee., "A Korean Product Review Analysis System Using a Semi-Automatically Constructed Semantic Dictionary," *Journal of KIISE : Software and Applications*, Vol.35(2008), 392～403.

Sebastiani, F., *Classification of Text, Automatic*, the Encyclopedia of Language and Linguistics 14, 2nd Edition, Elsevier Science Pub, 2006.

Stanvrianou, A., P. Andritsos, and N. Nicoloyannis, "Overview and Semantic Issues of Text Mining," *ACM SIGMOD Record*, Vol. 36(2007), 23～24.

Witten, I, H., *Text Mining*, Practical Handbook of Internet Computing, CRC Press, 2004.

Yoon, S., "A Study of Churn Prediction Model for Department Store Customers Using Data Mining Technique," *Asia Marketing Journal*, Vol.6, No.4(2005), 45～72.

Abstract

User-Perspective Issue Clustering Using Multi-Layered Two-Mode Network Analysis

Jieun Kim* · Namgyu Kim** · Yoonho Cho***

In this paper, we report what we have observed with regard to user-perspective issue clustering based on multi-layered two-mode network analysis. This work is significant in the context of data collection by companies about customer needs. Most companies have failed to uncover such needs for products or services properly in terms of demographic data such as age, income levels, and purchase history. Because of excessive reliance on limited internal data, most recommendation systems do not provide decision makers with appropriate business information for current business circumstances. However, part of the problem is the increasing regulation of personal data gathering and privacy. This makes demographic or transaction data collection more difficult, and is a significant hurdle for traditional recommendation approaches because these systems demand a great deal of personal data or transaction logs. Our motivation for presenting this paper to academia is our strong belief, and evidence, that most customers' requirements for products can be effectively and efficiently analyzed from unstructured textual data such as Internet news text. In order to derive users' requirements from textual data obtained online, the proposed approach in this paper attempts to construct double two-mode networks, such as a user-news network and news-issue network, and to integrate these into one quasi-network as the input for issue clustering. One of the contributions of this research is the development of a methodology utilizing enormous amounts of unstructured textual data for user-oriented issue clustering by leveraging existing text mining and social network analysis.

In order to build multi-layered two-mode networks of news logs, we need some tools such as text mining and topic analysis. We used not only SAS Enterprise Miner 12.1, which provides a text

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-5209, E-mail: ngkim@kookmin.ac.kr
*** College of Business Administration, Kookmin University

miner module and cluster module for textual data analysis, but also NetMiner 4 for network visualization and analysis. Our approach for user-perspective issue clustering is composed of six main phases: crawling, topic analysis, access pattern analysis, network merging, network conversion, and clustering. In the first phase, we collect visit logs for news sites by crawler. After gathering unstructured news article data, the topic analysis phase extracts issues from each news article in order to build an article-news network. For simplicity, 100 topics are extracted from 13,652 articles. In the third phase, a user-article network is constructed with access patterns derived from web transaction logs. The double two-mode networks are then merged into a quasi-network of user-issue. Finally, in the user-oriented issue-clustering phase, we classify issues through structural equivalence, and compare these with the clustering results from statistical tools and network analysis.

An experiment with a large dataset was performed to build a multi-layer two-mode network. After that, we compared the results of issue clustering from SAS with that of network analysis. The experimental dataset was from a web site ranking site, and the biggest portal site in Korea. The sample dataset contains 150 million transaction logs and 13,652 news articles of 5,000 panels over one year. User-article and article-issue networks are constructed and merged into a user-issue quasi-network using Netminer. Our issue-clustering results applied the Partitioning Around Medoids (PAM) algorithm and Multidimensional Scaling (MDS), and are consistent with the results from SAS clustering.

In spite of extensive efforts to provide user information with recommendation systems, most projects are successful only when companies have sufficient data about users and transactions. Our proposed methodology, user-perspective issue clustering, can provide practical support to decision-making in companies because it enhances user-related data from unstructured textual data. To overcome the problem of insufficient data from traditional approaches, our methodology infers customers' real interests by utilizing web transaction logs. In addition, we suggest topic analysis and issue clustering as a practical means of issue identification.

Key Words : Data Mining, Issue Clustering, Social Network Analysis, Topic Analysis

Received: June 15, 2014 Revised: June 20, 2014 Accepted: June 23, 2014

저 자 소 개

김 지 은

중앙대학교 경제학과에서 학사 학위를 취득하고, 현재 국민대학교 비즈니스IT 전문대학원 비즈니스IT 석사 과정에 재학 중이다. 글로벌 무역 포털인 ECPLAZA의 기획팀장으로 국가무역프로세스 혁신사업(BPR/ISP)과 전자무역신뢰체계(TSTA) 구축사업 등 전자무역 관련 프로젝트를 다수 수행하였다. 관련 주요 관심분야는 데이터 마이닝, 소셜 네트워크 분석, 전자무역 등이다.

김 남 규

현재 국민대학교 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학회 이사, 한국IT서비스학회지 편집위원, JITAM 편집위원, 한국인터넷정보학회논문지 편집위원을 역임하였으며, 한국생산성본부 TOPCIT 개발사업 자문위원으로 활동 중이다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 시맨틱 데이터 관리 등이다.

조 윤 호

현재 국민대학교 경영학부 빅데이터경영통계전공 교수로 재직 중이다. 서울대학교 계산통계학과를 졸업하고, KAIST 경영정보공학과에서 석사, KAIST 경영공학과에서 박사학위를 취득하였으며, LG전자(주)에서 6년간 주임연구원으로 재직하였다. 주 연구분야는 비즈니스애널리틱스, 빅데이터마이닝, 추천시스템, 소셜네트워크분석, 고객관계관리 등이다.