지능정보연구 제17권 제1호 2011년 3월(pp.153~169)

웹사이트 중복회원 관리 : 소셜 네트워크 분석 접근

강은영
국민대학교 비즈니스IT 전문대학원
(01244972@hanmail.net)

곽기영
국민대학교 경영대학 경영정보학부
(kykwahk@kookmin.ac.kr)

오늘날 기업의 마케팅에 있어 인터넷 환경의 이용은 필수적이며, 좀 더 효율적인 마케팅을 위해 다양한 방법들이 시도되고 있다. 기업들은 온라인마케팅을 통해 다양한 경품이나 포인트 등의 마케팅 비용을 사용하는 것으로 제품이나 서비스를 알려왔다. 특히 웹 2.0의 등장과 함께 기업은 좀 더 적극적으로 고객과 소통하기 위한 노력을 아끼지 않고 있다. 고객들은 회사의 웹사이트에 개인정보를 제공하는 형태로 회원가입을 하여 회사가 제공하는 혜택을 받으면서 제품 광고나 프로모션에 참여하게 된다. 그러나 온라인 마케팅의 운영측면에서 볼 때 현재의 회원관리 시스템은 회원의 모집과 운영에 있어서 효과적이지 못한 문제점이 나타나고 있다. 온라인 환경에서의 고객들은 오프라인 환경에서보다 명확한 자아를 덜 드러내기 때문에 회원가입 과정 중에 일부 악의적인 목적을 가진 고객들이 주변인의 개인정보를 이용하거나 조작하여 중복 아이디를 만들어 활동할 수 있게 된다. 이러한 취약점을 이용하여 중복가입 회원들은 고객들에게 돌아가야 할 경품이나 포인트 등을 가로채어 기업 마케팅 비용의 효율을 떨어뜨리고 있다. 그러나 증가하고 있는 마케팅 비용에 비해 중복회원의 선별 및 이들에 대한 제재를 위한 효과적 방법은 뚜렷하게 제시되지 않고 있다. 따라서 이를 방지하기 위한 체계적인 회원관리 시스템이 요구된다. 본 연구에서는 소셜 네트워크 분석 기법을 이용한 중복회원 식별방법을 제시하고 실제 온라인 고객데이터를 이용하여 그 효과성을 검증한다. 소셜 네트워크는 노드들의 관계를 표현하며, 관계의 유무, 방향 및 강도 등으로 연결 형태를 나타낼 수 있다. 특히 컴포넌트 분석방법은 소셜 네트워크 하위그룹 분석방법으로 네트워크의 내부 그룹을 구분하여 다양한 네트워크 특성을 식별하여 준다. 회원정보 분석에 있어 컴포넌트 분석방법은 전제회원 데이터 내의 의미 있는 정보를 이루고 있는 그룹을 식별하게 된다. 본 연구는 H사의 서로 다른 회원가입 기준을 가진 3개 웹사이트의 회원정보를 사용하여 진행되었다. 제안된 분석방법은 중복회원의 실체를 분석하고 시각화함으로써, 실무적인 측면에서 효율적인 마케팅의 증진을 도울 뿐만 아니라 신뢰성 있는 고객의 의견수렴 및 의사결정에도 도움이 될 것으로 기대된다.

논문접수일 : 2010년 12월 23일 논문수정일 : 2011년 01월 04일 게재확정일 : 2011년 01월 16일
투고유형 : 2010 추계학술대회 우수 교신저자 : 곽기영

1. 서 론

최근 개인의 프라이버시 보호차원에서 웹사이트에 대한 회원가입이 실명인증 등을 통해 엄격히 관리되고 있는 추세이다. 하지만 여전히 대부분의 기업 웹사이트에서는 가입자 편의를 위해 이메일이나 사용자 생성 아이디를 이용하여 회원을 구분하고 이름과 주소 등의 추가적인 회원정보를 입력받는 형태로 회원가입절차를 시행하고 있다.

특히 홍보 목적으로 운영되는 기업 웹사이트의 경우 개인정보에 대한 실명인증을 요구하지 않거나 실명인증 절차가 있다 하더라도 주변인의 정보를 사용해도 회원가입에 크게 어려움을 겪지 않는다. 따라서 이를 악용한 일부 고객들이 여러 개의

* 본 연구는 2011년도 국민대학교 우수연구센터사업비를 지원받아 수행된 연구임.

이메일과 회원정보로 동일한 서비스(웹사이트)에 가입하여 이벤트에 응모하거나 포인트를 적립하는 등 고객관리의 취약점을 악용하는 사례가 나타나고 있다.

정상적인 제품마케팅 활동에 참여하는 순수목적을 가진 회원이 아닌 중복가입고객은 가족이나 친구의 개인정보를 이용하여 경품을 부당 수령하거나 해당 서비스의 포인트를 비정상적인 방법으로 축적하기도 한다. 이러한 회원의 경우 악의적인 목적으로 타인의 개인정보를 도용하여 여러 개의 아이디를 관리하는 경우도 있다. 중복가입고객은 제품이나 서비스의 홍보를 목적으로 하는 해당 웹사이트의 마케팅 효율성을 떨어지게 하는 것은 물론 저조한 경품 및 이벤트 당첨률로 인해 기업에 대한 고객의 신뢰마저 저하시킨다.

웹사이트 관리자 입장에서 볼 때 중복가입고객으로 인해 발생하는 허수회원은 보안이나 시스템 운영과 같이 웹사이트에 직접적으로 영향을 미치는 요인에 비해 사소한 문제라고 여겨져 왔다. 또한 데이터베이스(DB)를 이용하여 중복회원을 쉽게 걸러낼 수 있다고 판단하거나 허수회원은 수치상 웹사이트의 마케팅 효과가 증대된 결과로 보여질 수 있기 때문에 이러한 현상은 어느 정도 묵인이 되어왔다. 이로 인해 중복회원에 대한 관리의 필요성은 지금까지 상대적으로 덜 중요하게 고려되어 왔으나, 인터넷을 통한 마케팅의 효율성 및 효과성이 점차 중요하게 요구되면서 중복가입회원에 대한 관리 필요성이 증대하고 있다.

마케팅 목적으로 웹사이트를 운영하는 기업에서는 중복가입고객으로 인해 다음과 같은 측면에서 운영상 비효율성이 발생할 수 있다.

첫째, 상대적으로 적극적으로 포진해 있는 중복가입회원으로 인해 실제 순수고객들의 이벤트 당첨률이 떨어진다. 따라서 당첨되지 않는 경품 이벤트 행사는 고객의 신뢰를 떨어뜨리기도 하고, 고객들로 하여금 중복당첨자에 대한 의심을 불러 일으키기도 한다. 그러나 해당 웹사이트의 관리자는 이 사실을 잘 모르는 경우도 있다.

둘째, 기업에서는 고객을 상대로 한 불미스러운 일을 최소화 한다는 입장에 있기 때문에 확실한 증거 없이 고객을 상대로 이벤트 당첨에서 불리한 조치를 내리기 어렵다.

셋째, 웹사이트의 시스템적인 규칙이나 운영내용을 파악하고 있는 중복가입회원은 장기적으로 활동할 가능성이 높다. 따라서 이로 인해 단기적으로뿐만 아니라 장기적 관점에서도 마케팅 효과가 감소된다.

넷째, 중복가입고객은 또 다른 중복가입고객을 발생시키기 쉽다. 고객의 전파경로와 마찬가지로 가까운 사람들에 의한 입소문을 통해 퍼질 가능성이 높기 때문이다. 이러한 중복가입고객의 제품/서비스에 대한 충성도는 낮을 것으로 예상된다.

중복가입고객은 해당 사이트의 허술하게 관리되는 부분을 잘 알고 이용하고 있기 때문에 주소나 아이디 비교 등 단순 DB 비교만으로는 중복가입고객을 추출해 내기가 어렵고, 의심이 가는 블랙리스트도 공식적으로 사용하기는 어렵다.

단순 DB 비교 방법으로는 경품수령을 위한 주소를 비교해 보는 방법을 주로 사용한다. 그러나 당첨자가 많을 경우 의도적으로 변형된 실질적으로 동일한 주소를 사용하거나 회사나 거주지 등 서로 다른 장소를 주소로 입력한 중복가입고객을 찾아내기는 쉽지 않다(어떤 중복가입고객은 당첨된 후 주소변경을 통하여 이러한 과정을 피해가기도 한다). 또한 중복가입고객이라 하더라도 전화조사 등을 통해 고객의 감정을 상하게 할 수는 없기 때문에 이러한 문제를 명확하게 조사하는 것은 기업에게 매우 어려운 일이다.

본 연구에서는 소셜 네트워크분석을 이용하여 이러한 중복회원 정보 사이의 연계성을 파악하고, 관계정보의 연계성을 바탕으로 중복가입고객을 식별하는 방법론을 제시한다. 이를 위해 본 연구에서는 H사에서 운영하는 세 개 웹사이트의 회원데이터를 이용하여 제안된 방법론의 효과성을 검증하고자 한다.

2. 이론적 배경

2.1 소셜 네트워크분석의 개념

사회적 관계로 성립된 구조의 의미를 분석하는 소셜 네트워크분석(social network analysis)은 분석초점에 따라 '에고 네트워크(ego-centric network)', '양자 네트워크(dyadic network)', '전체 네트워크(total network)' 등 다양한 네트워크 형태로 분류되며, 관계모형이나 관계의 강약, 밀도의 높고 낮음에 따른 다양한 사회적 역할 및 영향을 분석할 수 있다. 이러한 네트워크의 영향력과 효과의 가치는 주로 정보를 주고받거나 정서적, 물질적 지원으로 인해 발생되는 관계데이터의 분석을 통해 파악된다(김용학, 2003b; 손동원, 2002; Butts, 2008; Newman, 2006). 따라서 소셜 네트워크분석은 행위자 사이의 상호작용(interaction)이 구체적인 '실체'로 나타나는 것으로 볼 수 있다.

네트워크의 가장 기본적인 구성인 사람, 지역, 자원과 같은 행위자 노드(node)는 다양한 관계에 의한 연결형태 링크(link)로 나타나게 된다. 이때 관계데이터는 노드 사이의 관계 유무, 방향 및 빈도와 강도 등을 나타낼 수 있다. 특히 '아마존' 사이트의 추천시스템과 같이 행위자들의 직접적인 연결관계가 아닌 준연결망(김용학, 2003a; 박종학 등, 2009)을 통해서도 행위자에 대한 네트워크 분석이 가능하다. 이러한 준연결망을 통해 협업필터링(김형도, 2009), 관계의 추론(이승훈 등, 2007; 이승훈 등, 2009)등 다양한 연구가 이루어지고 있다. 또한 네트워크 안에서는 다른 노드보다 더 많은 연관성을 가진 다양한 형태의 하위집단(sub group)을 발견할 수 있다. 하위집단은 네트워크 내부에서 관찰되는 그룹들로 네트워크의 커뮤니티 구조를 파악할 수 있다(Girvan and Newman, 2002).

2.1.1 하위집단

한 네트워크 내부에는 다양한 하위집단이 존재하게 된다. 하위집단은 동질적인 이해관계나 성격을 가진 네트워크 내의 부분집합으로 구성되며, 하위집단에 대한 분석을 통해 구성원들 사이의 이해관계나 파벌 및 역할, 소속 등 다양한 집단특성을 파악할 수 있다. 또한 이를 통해 해당 네트워크가 가진 응집력, 갈등, 협력, 신뢰 등도 파악이 가능하다(손동원, 2002).

Newman(2004)에 의하면 하위집단은 그 안에서 노드 간의 강한 연결을 발견할 수 있고, 하위집단과 하위집단 사이에는 비교적 약한 연결을 보이는 노드들의 집합이다. 이러한 그룹을 발견하기 위한 방법으로 스팩트럼 분할(spectral bisection), 계층트리(hierarchical clustering), 상호관계 관찰 등에 의한 컴퓨터 알고리즘이 소개되고 있다(Girvan and Newman, 2004; Newman, 2003).

하위집단은 크게 계층적 의미를 갖는 클러스터링(Newman, 2003; Newman, 2004)과 응집노드에 기반을 두는 파당분석(손동원, 2002; Palla et al., 2005), 그리고 연결된 노드들이 구축하는 특정 영역인 응집지역에 초점을 맞추는 컴포넌트(김용학, 2003a; 손동원, 2002) 형태로 관찰할 수 있다. 소셜 네트워크의 그래프이론(Albert and Barabasi, 2002;

Faloutsos et al., 2004)에서 정의되는 하위집단인 서브그래프(subgraphs)에서도 그 구성요소로써 클러스터링과 컴포넌트가 사용된다.

클러스터링은 집단 사이의 계층적인 의미를 갖는 하위집단을 의미하며 광범위하고 다양한 연구가 이루어지고 있다. 최근 포털에서 서비스되고 있는 뉴스 클러스터링 같은 경우 텍스트 마이닝을 이용한 뉴스 안의 그룹 발견(Joshi and GaticaPerez, 2006)과 토픽 캡쳐(Mccallum et al., 2005)에 의한 관계연구를 통해 발달해왔고, 커뮤니티 분야에서도 텍스트 마이닝을 사용한 커뮤니티 클러스터링(Velardi et al., 2008)을 비롯한 클러스터링을 이용한 개인적 지역감지 알고리즘(Zhou et al., 2004) 등의 연구가 이루어지고 있다.

파당분석은 한 네트워크에서 결속집단을 찾는 것을 의미하고, 세 점 이상으로 구성된 하위그래프로 표현된다. 모든 노드들이 반드시 직접적으로 연결되어 있어야 하며 서로 다른 파당그래프에 다른 그래프와 중복되는 노드가 나타날 수 있다(김용학, 2003a; 손동원, 2002). 따라서 파당분석에서 네트워크는 노드를 공유하는 하위집단들의 조합이라고 할 수 있다(Palla et al., 2005).

컴포넌트는 연결된 모든 점들을 포함하는 하위 네트워크이며, 계층적 의미나 결속집단을 고려하지 않는다. 한 점에서 출발하여 그 점과 연결된 모든 점들을 찾아내는 절차를 지속하여 연결고리를 끝까지 추적하는 방법인 '눈덩이 굴리기의 방법'을 사용하여 컴포넌트를 찾을 수 있다(손동원, 2002). 발견된 컴포넌트 외에 어떠한 컴포넌트에도 속하지 않게 되는 점들은 고립점(isolated points)이라고 한다.

2.1.2 컴포넌트 구성과 의미

컴포넌트는 <그림 1>의 A, B, C와 같이 네트워크 내부에서 연결이 끊어지지 않은 모든 하위그룹을 의미한다. A, B, C 각 컴포넌트는 각각의 관계데이터에 의해 연결된 하위그룹이며 네트워크 내부에서 개별적인 특성이나 의미를 갖게 된다. 일반적인 네트워크에서 컴포넌트가 많이 발견된다는 것은 정보의 흐름이 원활하게 흐르지 않고 하위집단간 파편화될 가능성이 있음을 나타낸다. 이렇듯 컴포넌트의 수와 규모를 파악하는 것으로 한 네트워크의 자원흐름 원활 여부와 자원의 편중적 사용 여부를 알 수 있다(손동원, 2002).

<그림 1> 소셜 네트워크와 컴포넌트

또한 정보의 연결성 측면에서 네트워크 내의 컴포넌트들은 의미 있는 정보를 이루고 있는 그룹을 나타내게 된다. 본 연구에서는 소셜 네트워크의 컴포넌트 분석을 통해 각 노드의 정보 연결성을 파악하여 웹사이트의 중복고객을 식별하는 방법을 제안하고자 한다.

2.2 관련 연구 고찰

네트워크 안에서 커뮤니티 구조의 발견은 타당한 실용성을 갖는다(Girvan and Newman, 2002). 특히 소셜 네트워크에서 관계데이터의 분석은 행위자에 대한 직접적인 연결이나 조사과정이 없어도 클러스터링이나 컴포넌트 등의 네트워크분석에 의해 커뮤니티 구조를 발견함으로써 행위자의 실체를 파악할 수 있게 한다.

소셜 네트워크 안에서 컴포넌트들은 실제 사회그룹에서 나타나는 현상으로도 표현될 수 있다. 인

용문 네트워크의 커뮤니티에서 페이지들이나 단일 토픽 등의 연관성을 나타낼 수 있으며(Chen, 1999), 범죄수사에서도 그 연관성은 범죄수사의 시각화(Kerschbaum and Schaad, 2008; Xu and Chen, 2005)를 통해 나타나기도 한다.

또한 소셜 네트워크 클러스터링 알고리즘은 스팸처리에서 스팸 메일 예측모형(안수산과 신경식, 2000; Fawcett, 2004)을 구축하거나 스팸 이미지에 대한 데이터 처리 방법으로 스팸 소스를 추적(Zhang et al., 2009)하는 방법으로 이용되고 있다.

최근 소셜 네트워크의 다양한 연구 중에서 시멘틱 웹 패러다임의 발달과 함께 소셜 네트워크 모델링과 분석의 접근방법 중 하나로 온톨로지 이용이 큰 관심을 얻고 있다(이승훈 등, 2007). Wennerberg (2005)에 의하면 소셜 네트워크는 사회적 엔티티(entities)에 대한 모델(사람, 조직, 발생하는 이벤트)로 이루어진 온톨로지 기반 관계에서 추론 메카니즘을 통해 노드 간의 새로운 관계를 발견한다. 특히 드러나지 않은 잠재된 관계에서 소셜 네트워크를 이용하여 명확한 관계를 나타냄으로 시멘틱 웹의 추론에 사용된다.

시맨틱 웹과 소셜 네트워크의 합성어인 시맨틱 소셜 네트워크(이승훈 등, 2009)는 이러한 사회적 연결관계에 의미를 부여한다. 리소스와 리소스의 준연결망에는 관련된 사회적 노드(사람, 조직 등)가 함께 연결되어 있기 때문에, 노드 간의 다양한 의미적 연결관계를 가지는 컴포넌트나 클러스터링을 이용한 표현과 추론이 가능하다. 이를 바탕으로 아마존 서점의 추천 시스템이나 협업 필터링(위키백과사전)을 기반으로 한 신규추천 문제(김형도, 2009; 박종학 등, 2009; Jaewon et al., 2008) 등이 연구되었다.

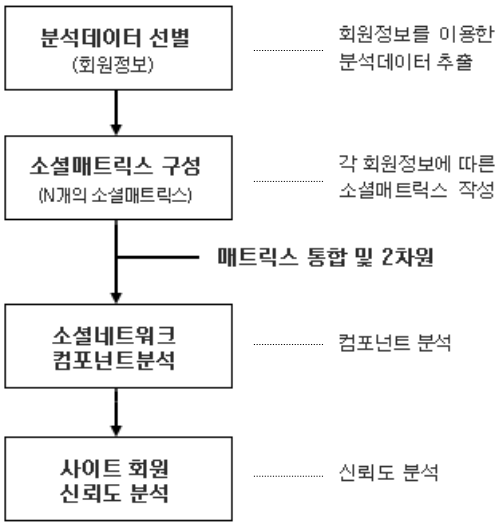
서브그룹과 준연결망에 의한 소셜 네트워크분석방법은 데이터 마이닝을 통해서도 다루어 지고 있다. Velardi(2008)에 의해 제시된 컨텐츠 베이스 소셜 네트워크분석(CB-SNA)은 커뮤니케이션의 캡쳐와 식별을 위해 텍스트 마이닝(Frey and Dueck, 2007)을 사용하여 관계의 양이 아닌 질적인 관계에 대해 고찰하였다. 이것은 구글 뉴스에서 그룹을 추출하거나(Joshi and GaticaPerez, 2006), 메시지 교환 환경에서 송·수신자에게서 토픽과 규칙을 발견하는(Mccallum et al., 2005) 연구에서 발전한 것으로 클러스터링 알고리즘을 사용한다. 또한 협업 필터링을 기반으로 고객의 네트워크 가치를 마이닝하여 잠재고객의 마케팅 효과를 높이는 모델도 제시되고 있다(Domingos and Richardson, 2001).

응용분야인 컴퓨터과학, 전자공학, 생체공학, 의학 등 다양한 분야의 과학자들에 의해 만들어진 BSNs(Body Sensor Networks)에서도 소셜 네트워크 하위그룹을 이용한 연구가 진행되고 있다. 특히 무선네트워크를 접목시킨 환경에서 수집된 감각데이터는 소셜 네트워크 하위그룹에 의해 매핑되어 신체활동을 감지하거나 모니터링 할 수 있고, 이러한 내용은 의료 응용에서 매력적인 어플리케이션을 발견하는데 이용된다(Rahman et al., 2010).

3. 소셜 네트워크분석 기반 중복고객 식별 방법

소셜 네트워크를 이용한 중복고객 식별은 고객들의 가입정보를 이용하여 <그림 2>와 같은 절차를 따라 수행된다.

첫 번째 과정인 분석데이터 선별 단계에서 추출되는 데이터는 아이디, 이름, 전화번호 등 일반적으로 중복 가능성이 낮고, 개인의 특성을 나타내는 자료로 선별된다. 선별된 각각의 데이터는 두 번째 단계에서 각 변수에 대한 1 모드 매트릭스(손동원, 2002)를 도출하고, 세 번째 단계에서 도출된 모든



<그림 2> 중복고객 식별 절차

매트릭스의 합산된 값에 대해 소셜 네트워크 컴포넌트 분석을 수행한다. 매트릭스의 합산된 값은 회원에 대한 의미 있는 식별 고유성으로 네트워크에서의 연결관계를 보여준다. 끝으로 고립 노드와 중복회원(컴포넌트)의 비율을 바탕으로 중복회원으로 인한 사이트 신뢰도를 평가한다.

3.1 분석데이터 선별

일반적으로 보이는 중복가입자의 패턴은 아이디의 일부를 변경하거나 비밀번호를 동일하게 작성하는 등 기존 개인정보의 일부를 변경하여 사용하는 것이다. 또한 가족 혹은 타인의 정보를 도용하더라도 경품수령을 위한 주소 등 최소한의 가입정보는 일치하는 점을 보이고 있다.

그 내용은 아래와 같이 정리할 수 있다.

- 아이디의 일부가 비슷하다.
- 비밀번호가 같다.
- 전화번호가 같거나 비슷하다.
- 주소가 동일하다(가족 회원이 있을 가능성과 가족의 정보 이용가능성이 있다).
- 게시글 작성시간과 IP등 외부적 정보가 동일하다.

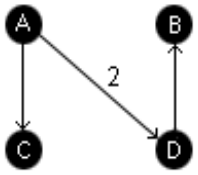
따라서 아이디의 일부, 이름, 비밀번호, 전화번호, 주소 및 IP정보에 대한 <표 1>과 같은 자료 추출이 가능하다. 그러나 비밀번호의 경우 시스템 상에서 암호화 되어 있는 경우가 많고, 개인정보의 유출 우려가 있어 제외한다.

<표 1> 회원 데이터와 분석데이터

회원데이터	분석데이터
회원 아이디	이메일 앞에서 3자리
비밀번호	분석제외
이름	이름전체
전화번호	뒷번호 4자리
휴대전화	뒷번호 4자리
주소	우편번호 뒷자리

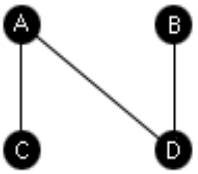
3.2 소셜매트릭스 구성

이 단계에서는 이전 단계에서 추출된 데이터를 바탕으로 각 리소스와 리소스 사이의 소셜매트릭스를 생성한다. 소셜매트릭스는 노드간 관계데이터의 값을 보여주는 (n×n)행렬이다. 한 네트워크에 A, B, C, D회원이 있을 때, <그림 3>(a)와 같이 방향이 있는 관계는 정보나 재화 등의 흐름을 나타낸다. 매트릭스에서 행은 영향이나 정보를 주는 노드의 상태를 나타내고, 열은 영향이나 정보를 받는 노드의 상태를 나타낸다. 또한 수치를 통해 관계의 강도(intensity)를 표현할 수 있다. 관계의 방향이 없이 연관성만을 나타내는 <그림 3>(b)의 네트워크에서도 행렬 매트릭스를 통해 관계를 표현할 수 있으며, 이때 행과 열은 같은 정보를 나타내게 된다.



	A	B	C	D
A	-	0	1	2
B	0	-	0	0
C	0	0	-	0
D	0	1	0	-

(a) 방향이 있는 관계



	A	B	C	D
A	-	0	1	1
B	0	-	0	1
C	1	0	-	0
D	1	1	0	-

(b) 방향이 없는 관계

<그림 3> 소셜 매트릭스

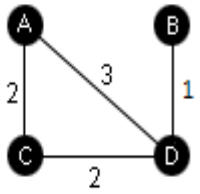
이번 단계에서는 앞 단계에서 추출한 회원정보를 바탕으로 아이디(이메일), 이름, 휴대전화번호, 전화번호, 우편번호에 대한 소셜 매트릭스를 생성하고 이들을 합산한다. 이 과정은 동일한 회원의 준연결망(김용학, 2003a)을 리소스 별로 생성하여 합산하는 것으로 합산매트릭스는 회원정보의 연관성을 최소 0에서 최대 분석대상 회원정보 개수까지의 수치로 나타내게 된다. 따라서 합산된 매트릭스(S_{ij})는 강도를 가지는 방향이 없는 관계매트릭스이며 강도가 높을수록 회원간의 관계가 더 강하게 있다고 판단할 수 있다.

합산매트릭스(S_{ij})에서는 관계에 대한 기준 값(p)을 설정하여 기준 값에 따라 관계의 존재유무를 결정한다. 본 연구에서는 기준 값인 p에 따라 p보다 작으면(S_{ij} < p) '0', 같거나 크면(S_{ij} >= p) '1'로 코딩하여 2진 매트릭스로 변환한다. p의 값은 네트워크의 특성과 데이터의 종류에 따라 설정한다. 매트릭스 합산 과정에서 제거된 기준 값(p) 미만의 관계강도를 갖는 노드들은 컴포넌트 분석에서 나타나지 않으므로 의도하는 안정적인 관계를 가진 컴포넌트 도출이 가능하게 된다.

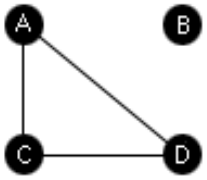
3.3 소셜 네트워크 컴포넌트 분석

컴포넌트 분석은 2진 소셜매트릭스의 '1'과 '0'의 값으로 연결성을 파악한다. 아래 <그림 4>과 같은 경우 <그림 4>(a)는 각각 관계가 합산된 소셜매트릭스와 그 연결관계를 소셜다이어그램으로 표현한 것이다. {B, D} 사이에 하나의 관계가 있고 {A, C}, {C, D} 사이에 각각 2개의 관계 그리고 {A, D}에 3개의 관계가 있음을 알 수 있다. 관계에 대한 기준 값을 2(p = 2)로 설정한 <그림 4>(b)에서는 2이상의 관계로 이루어진 하나의 컴포넌트 {A, C, D}와 고립점 B가 나타나며, <그림 4>(c)와 같이 기준 값을 3(p = 3)으로 설정하였을 경우 {A, D}로 이루어진 컴포넌트와 B, C 두 개의 고립점이 나타나게 된다. 기준 값을 높게 설정할수록 컴포넌트 분석은 엄격한 기준을 갖게 된다.

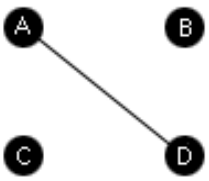
	A	B	C	D
A	-	0	2	3
B	0	-	0	1
C	2	0	-	2
D	3	1	2	-



(a) 방향이 없고 강도가 있는 관계



(b) p = 2



(c) p = 3

p가 연결수치라고 가정하였을 때 <그림 4>(b)는 연결이 2개(p >= 2)이상일 때 인정되는 컴포넌트이고, <그림 4>(c)는 연결이 3개 이상(p >= 3)일 때 인정되는 컴포넌트이다. 컴포넌트 분석을 위하여 매트릭스는 기준 값 p의 설정에 따라 p값 이상일 때는 '1', p값 미만일 때는 '0'의 값을 가지는 2차원 배열로 변환되어야 한다.

<그림 4> 컴포넌트 분석

3.4 사이트 회원 신뢰도 분석

이전 단계에서 나타난 컴포넌트 분석결과의 고립점과 컴포넌트의 수를 이용하여 식 (1)과 같이 실제회원의 수를 계산할 수 있다.

실제회원 = 고립점+컴포넌트(2개 이상의 노드) (1)

또한 전체 노드와 실제회원의 노드 수를 이용하여 식 (2)와 같이 전체 사이트에 대한 회원 신뢰도를 정의할 수 있다.

$$\text{사이트 회원 신뢰도}(\%) = \frac{\text{실제회원}}{\text{전체노드}} \times 100 \quad (2)$$

4. 중복회원식별 파일롯 연구

본격적인 실증연구에 앞서 M사이트의 중복회원에 대한 실험적인 컴포넌트 분석을 실행하였다. 중복회원이 포함된 실제 회원정보 20개의 이름, 주소, 전화번호, 휴대폰번호, 이메일에 대한 5개의 매트릭스(i*j)를 작성하고, 컴포넌트 분석을 위해 합산매트릭스 M(i*j)의 결과값에 대해 p = 3을 기준으로 '1'과 '0'으로 매핑한 2진 매트릭스(i*j)를 도출하여 파일롯 연구를 수행하였다.

4.1 분석데이터 선별

<표 2>와 같이 독립된 20개의 회원정보를 추출하였다. 아이디 부분중복 확인을 위해 아이디 앞자리 3자리와 이름, 전화번호 뒷자리, 핸드폰 뒷자리, 우편번호 뒷자리를 선정하였다.

<표 2> 회원정보 추출

회원정보	분석데이터
아이디	앞자리 3자리
이름	이름전체
전화번호	뒷자리 4자리
휴대폰	뒷자리 4자리
우편번호	뒷자리 3자리

4.2 소셜매트릭스 구성

선별된 분석데이터에 따라 각각의 연관성을 파악하여 회원정보 개별항목(아이디, 이름, 전화번호, 핸드폰, 우편번호)의 소셜매트릭스를 구한 후 합산된 매트릭스를 도출하였다. 이를 바탕으로 컴포넌트 분석을 위해 기준 값 3(p = 3)을 기준으로 크거나 같으면 '1', 기준 값보다 작으면 '0'의 값을 가진 2진 매트릭스를 <표 3>과 같이 도출하였다.

<표 3> M 사이트의 2진 매트릭스

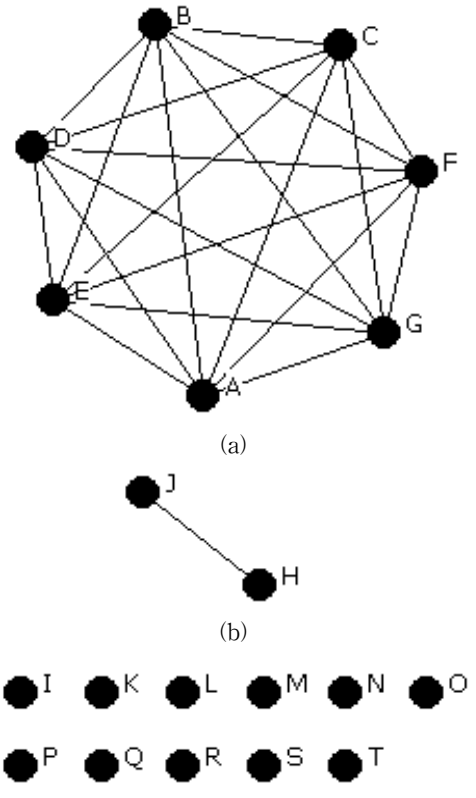
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	-	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
B	1	-	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C	1	1	-	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
D	1	1	1	-	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
E	1	1	1	1	-	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
F	1	1	1	1	1	-	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1	1	1	1	1	1	-	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	1	0	-	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-

4.3 소셜 네트워크 컴포넌트 분석

변환된 매트릭스를 기반으로 UCINET 6.187을 이용하여 컴포넌트 분석을 수행하였다. 본 연구에서는 최소 2개 이상의 노드로 연결되는 방향성 없는 노드들의 집합을 컴포넌트로 정의하였다. 그 결

과 <그림 5>와 같은 그래프가 도출되었다.

<그림 5>와 같은 경우 {J, H}, {A, B, C, D, E, F, G} 두 개의 컴포넌트는 각각 한 회원이 소유하고 있는 중복가입고객 아이디로 간주할 수 있다. 특히 <그림 5>의 (가)컴포넌트는 7개의 모든 노드가 3(p = 3) 이상의 관계로 구성되어있는 것을 알 수 있다.



<그림 5> M사이트의 회원 소셜 다이어그램

4.4 사이트 회원 신뢰도 분석

소셜 네트워크 컴포넌트 분석을 통해 2개의 컴포넌트와 11개의 고립점이 발견되어 실제 회원은 총 13명의 회원으로 구성된 것을 알 수 있었다.

<표 4> M사이트 컴포넌트 분석결과

구 분	M회원
회원정보	20(명)
컴포넌트	2
고립점	11
분석된 회원	13
사이트 회원신뢰도	0.65

식 (2)에 의한 전체 회원에 대한 사이트 회원 신뢰도는 65%로 신뢰도가 낮고 중복회원의 비율이 높게 나타나 있음을 알 수 있다. 또한 단순한 주소 비교를 통해서는 16명의 회원이 존재하는 것으로 나타나서, 본 연구에서 제시하는 방법이 더 우수한 중복회원 식별력을 보이는 것을 알 수 있었다. <표 4>에 M사이트의 중복회원 식별결과를 요약하였다.

5. 실증분석

5.1 데이터 수집

H사는 일반 음료를 비롯한 의약제품의 홍보사이트를 장기간 운영하고 있다. 사이트의 가입자는 점차 늘어나고 있는 추세이고, 지속적인 온라인 마케팅에 대한 투자가 이루어지고 있으나 마케팅 효과에 대한 효과적인 검증은 실질적으로 이루어지지 않고 있다. 따라서 소셜 네트워크를 이용한 중복회원 분석을 통한 회원신뢰도 분석을 위해 H사에서 운영하는 3개 사이트의 회원정보를 수집하였다. 3개 사이트의 회원은 각각 주민번호 인증등 실명인증과 이메일 확인방법을 통한 이메일 인증, 단순 이메일 가입 등 서로 다른 방법으로 모집되었다.

본 연구에서 제안하는 방법론을 검증하기에 앞서 해당 회원을 대상으로 H사의 음료 이용에 대한 설문조사 이벤트를 실시하였다. 이벤트의 목적은

<표 5> H사의 브랜드 사이트

구 분	사이트 분류	회원수	참여자	회원가입방법	운영기간
A(mi)	음료 브랜드 사이트	23,200	1303	실명인증	2년
B(m)	탈모샴푸	4,100	642	이메일가입	1년 미만
C(hb)	통합 브랜드 사이트	21,500	435	이메일인증	5년

참여(현재 활동)하는 고객을 대상으로 제안된 방법론의 효과성을 검증하기 위해 실시되었으며, 웹사이트 이벤트 형식으로 이메일을 통해 배포되었다. 설문조사와 함께 작성자에 대한 IP정보를 취득할 수 있도록 설정하였다. 이벤트 참여자는 A사이트 1,303(5.6%)명, B사이트 642(15.6%)명, C사이트 435(2.0%)명이었다. 연구에 사용된 H사의 브랜드 사이트에 대한 현황 및 참여자를 <표 5>에 정리하였다.

각 사이트 이벤트 참여자의 데이터를 활용하여 <그림 2>의 중복고객 식별절차에 따라 우선 <표 6>과 같은 회원정보를 H사의 각 브랜드 사이트로부터 추출하였다.

<표 6> 회원정보 추출

회원정보	분석데이터
아이디/이메일	앞자리 3자리
이름	이름전체
전화번호	뒷자리 4자리
핸드폰	뒷자리 4자리
우편번호	뒷자리 3자리
설문참여IP	IP전체

5.2 데이터 측정 및 분석

회원아이디의 유사아이디는 iblue, iblue1, iblue2 등 일정 패턴이 앞/뒤로 반복된다고 가정한다. 따라서 아이디의 문자 앞에서 3자리를 추출하여 같으면 '1'로 표시하고, 다르면 '0'으로 표시하여 ID 매트릭스 작성하였다(본 연구에서 아이디의 뒷부분의 반복은 계산하지 않았다).

고객 이름에 대한 관계데이터 매트릭스는 이름이 완전히 일치하는가에 따라 '1' 또는 '0'으로 코딩하였다. 휴대전화와 집전화 번호도 마찬가지로 각각 번호의 뒷자리 4자리가 완전히 일치하면 '1' 그렇지 않으면 '0'으로 매핑하였고, 주소에 대한 데이터는 우편번호 뒷자리를 사용하여 코딩 하였다. 회원자료 중 비빌 번호 중복이 많을 것으로 예상되었으나 개인정보 보호차원에서 시스템에서 암호화되어 있어 분석데이터에서 제외되었다.

컴포넌트 분석을 사용하기 위하여 아이디(이메일), 이름, 휴대전화번호, 전화번호, 우편번호, IP 매트릭스를 구하고 이를 합산한 합산 매트릭스(S_{ij})를 구성하였다. 합산 매트릭스(S_{ij})에서 기준 값(p)보다 작으면($S_{ij} < p$) '0', 같거나 크면($S_{ij} >= p$) '1'로 코딩하여 2진 매트릭스로 변환하였다. UCINET 6.187를 이용하여 2진 매트릭스의 컴포넌트 분석을 실시하였다.

5.3 분석결과

<표 7>은 관계데이터에 대한 기준 값이 각각 2(p = 2), 3(p = 3), 4(p = 4)일 때의 중복회원 컴포넌트와 단순 주소비교에 의한 중복회원 도출결과를 나타낸것이다. 본 연구에서는 p = 2일 때 가장 효과적인 중복회원 도출 결과가 나타났으며, 각 사이트 별로 A사이트는 123개, B사이트는 27개, C

사이트에서는 60개의 중복회원 컴포넌트가 나타났다. 이는 주소비교 방법에 비해서 평균 79% 정도의 중복회원이 더 발견된 결과이다. 또한 비교적 엄격한 기준인 p = 3을 적용하였을 때도 평균 12%의 중복회원 컴포넌트를 더 발견할 수 있었다.

<표 8>은 <표 7>에서 발견된 컴포넌트 분석결과와 식 (1)에 따라 실제 회원의 수를 정리한 내용이다. A사이트의 경우 회원가입정보에 의하면 1,303명의 회원정보가 존재하는 것으로 나타나있다. 그러나 일반적으로 사용하는 중복회원 추출을 위한 주소비교 결과 93명이 제거된 1,210명이 존재하는 것으로 나타났으며, p = 2를 기준으로 적용한 소셜 네트워크 컴포넌트 분석결과에서는 206명이 제거된 1,097명이 실제 회원으로 나타났다. 따라서 전체노드 1,303개, 고립노드 974개, 컴포넌트 123개로 84.19%의 사이트 회원 신뢰도를 보이고 있다.

B사이트에서는 회원가입정보에 의한 642명의 회원가운데 주소비교를 통해 25명이 제거된 617명이 주소비교기준의 실제회원으로 나타났다. p = 2를 기준으로 적용한 컴포넌트 분석을 통해서는 41명이 제거된 601명이 실제회원으로 도출되었다. B사이트는 전체노드 642개, 고립노드 574개, 컴포넌트 27개로 93.61%의 사이트 회원 신뢰도를 보이고 있다.

C사이트에서는 회원 가입정보에 의한 435명의 회원 정보가운데 주소비교를 통해 45명이 제거된 390명이 주소비교 기준의 실제회원으로 나타났다. p = 2를 기준으로한 컴포넌트 분석을 통해서는 114명이 제거된 321명이 실제 회원으로 집계되었다. 따라서 전체노드 435개, 고립노드 261개, 컴포넌트 60개로 73.79%의 사이트 회원 신뢰도를 나타냄으로써 분석한 3개의 사이트 중 가장 낮은 신뢰도를 보이고 있다.

소셜 네트워크 컴포넌트를 이용한 분석결과에 대한 검증을 위해서 설문조사 이벤트를 통해 수집

<표 7> 발견된 잠재 중복회원의 컴포넌트 수

사이트			A(mi)	B(m)	C(hb)	합계
주소비교			70	16	31	117
소셜 네트워크 컴포넌트 분석	p = 2	컴포넌트	**123**	**27**	**60**	**210**
		(노드)	(329)	(68)	(174)	(571)
	p = 3	컴포넌트	89	17	31	131
		(노드)	(220)	(43)	(80)	(343)
	p = 4	컴포넌트	49	11	16	76
		(노드)	(117)	(29)	(43)	(189)

<표 8> 컴포넌트 분석에 의한 실제회원

사이트		A(mi)	B(m)	C(hb)	합계
(회원정보의 회원 수)		1303	642	435	2380
주소비교		1210	617	390	2217
소셜 네트워크 컴포넌트 분석	p = 2	**1097(84.19%)**	**601(93.61%)**	**321(73.79%)**	**2019(84.83%)**
	p = 3	1172	616	386	2174(91.34%)
	p = 4	1235	624	408	2267(95.25%)

한 IP 컴포넌트와 주소비교 컴포넌트를 합산하여 대조하는 방법을 사용하였다. IP 컴포넌트는 IP전체가 동일해야만 '1'값을 가질 수 있는 매트릭스로부터 도출되었다. 본 연구에서 실시한 분석결과(p = 2)에는 각 사이트별로 90% 이상 IP 중복회원 컴포넌트가 포함되었다. 그러나 IP 컴포넌트는 사용자가 PC를 변경하였거나 재부팅 혹은 다른 장소에서 중복응모 하였을 경우 올바르지 못한 결과를 나타낼 수 있다. 따라서 유동IP에 대한 보완으로 IP 컴포넌트와 주소비교 방식의 컴포넌트를 그룹을 합산하여 1이상인 결과를 사용하였다.

검증데이터는 분석결과(p = 2)에서 나타나는 컴포넌트가 얼마나 정확하게 그룹화 되었는지 파악하기 위해 사용되었다. 결과적으로 소셜 네트워크 분석 결과(p = 2)로 나타난 전체 컴포넌트 중 5개의 그룹은 검증데이터와 상이한 것으로 나타났다. 상이한 컴포넌트에는 동일회원이 아니면서 연관성이 높은 노드가 연결 관계로써 포함되었다. 검증데이터에서 IP 및 주소가 동일하나 소셜 네트워크 분석결과에서 나타나지 않은 결과는 가족회원이나 동료 등 서로 다른 회원으로 파악하고 고려하지 않았다.

결론적으로 본 연구에서는 소셜 네트워크 컴포넌트 분석을 이용하여 검증데이터를 바탕으로 97.6%의 정확성을 같는 컴포넌트로 중복회원을 찾아내는 것으로 나타났으며, 전체 사이트의 평균적인 회원신뢰도는 85%로 15%의 회원이 중복회원으로 활동하는 것으로 판명되었다. 이는 주소비교 방식에서의 중복회원비율 7% 보다 두 배 이상 증가한 결과를 보여준다. 또한 관계데이터에 대한 엄격한 기준인 p = 3을 적용하였을 때도 9%의 중복회원 비율을 나타냄으로 주소비교 방식에 비해 높은 중복회원 발견비율을 보여주었다.

각 사이트의 중복회원에 대한 신뢰도는 실명인증이나 이메일인증 등 회원가입의 엄격함에 따라 비율이 달라질 것으로 예상할 수 있다. 그러나 회원 신뢰도는 가입방법에 관계없이 유지기간이 5년인 C사이트에서 74%의 낮은 비율을 보였고, 비교적 최근에 개설된 B사이트에서 94%로 나타났다. 따라서 사이트의 신뢰도는 실명인증과 같은 회원가입 제한사항보다 사이트 유지기간과 더 많은 관계가 있는 것으로 나타났다. 따라서 이러한 중복회원을 줄이기 위해서는 운영적인 측면에서 꾸준한 회원관리가 필요할 것으로 보인다.

6. 결 론

인터넷 환경에서는 많은 제품과 서비스들에 대한 적극적인 마케팅이 이루어지고 있다. 이러한 마케팅 목적의 사이트에서는 SNS(social network service) 커뮤니티 공간과는 달리 수많은 고객들이 명확한 자아를 드러내지 않고 있다. 따라서 많은 중복고객이 발생할 개연성이 존재하며, 이러한 중복고객은 마케팅의 효율성을 저하시키는 경향이 있다.

본 연구에서는 소셜 네트워크의 컴포넌트 분석을 통하여 중복고객 관리에 대해 연구하였다. 기존 회원관리에서는 실명인증 단계에서 중복회원 가입을 최소화하려는 노력을 하였으나 사이트가 장기적으로 운영될수록 중복회원은 점차적으로 증가하는 것을 알 수 있었다. 본 연구의 소셜 네트워크분석 방법을 사용하여 회원가입 시점에서 컴포넌트를 생성하고 컴포넌트별로 회원관리를 한다면 회원들에게 신뢰성 있는 사이트로 유지될 것으로 기대된다. 소셜 네트워크를 이용한 중복회원관리는 고객의 실체를 분석하고 시각화함으로써, 실무적인 측면에서 향후 효율적인 마케팅 뿐만 아니라 어떤 고객의 의견과 조언을 더 많이 들어야 할지에 대한 의사결정에도 도움이 될 수 있을 것이다.

더불어 소셜 네트워크 컴포넌트 분석에서는 중복가입자의 컴포넌트 그룹을 발견할 뿐만 아니라 주소가 동일하더라도 서로 다른 사람임을 나타내어 주는 사례도 발생하였다. IP 분석 및 사이트에 게재된 내용, 작성시간 등을 비교해 본 결과 소셜 네트워크분석으로 분석한 결과가 더 신뢰성 있는 것으로 파악되었다. 따라서 소셜 네트워크분석은 중복가입고객 서브그룹의 추출도 가능할 뿐 아니라 가족회원과 같이 동일한 주소라 할지라도 서로 다른 사람이 가입했을 경우 이를 분리해 줄 수 있어 주소비교에 의한 오류도 보완가능한 것으로 나타났다.

본 연구에서 소셜 네트워크 분석방법으로 기준(p = 2)를 제시하였다. 이는 회원정보의 검증데이터 확보를 위한 최소한의 변수 사용으로 최대의 효과를 보기위한 장치이다. 그러나 기준(p = 2)이 낮은 상태에서 변수로 사용한 우편번호 뒷자리는 '도'차원의 지역이 다를 경우 명확한 컴포넌트 추출이 어려울 수 있다. 이러한 결과는 본 연구의 한계점으로써 향후 진행되는 연구나 실무에서는 검증데이터로 사용한 변수를 비롯한 사용가능한 다양한 변수를 추가하고 기준을 강화 한다면 더 신뢰성 있는 결과가 있을 것으로 예상된다. 또한 본 연구에서는 모든 관계의 값을 '1', '0'으로 설정하여 평가하였으나, 분석데이터에 따라서 가중치를 둔다면 더 정교한 결과가 나올 것으로 보인다. 추가로 본 연구의 실험에서는 중복가입고객의 특성으로서 데이터를 입력하는 패턴의 유사성(예를들면, 중복가입한 회원은 'null'값을 많이 입력하거나 '0000'을 전화번호로 입력하는 것), 글 작성시간의 유사성, 주로 사용하는 문자 등도 발견할 수 있었다. 향후 이러한 행위나 패턴을 데이터화하는 측정장치를 마련하여 행위적인 측면에서도 보완이 이루어지면 본 연구에서 제시한 방법론이 보다 다양한 분야에서 응용될 수 있을 것으로 기대된다.

참고문헌

김용학, *사회연결망 분석*, 박영사, 2003a.

김용학, *사회연결망 이론*, 박영사, 2003b.

김형도, "일관성 기반의 신뢰도 정의에 의한 협업 필터링", *한국전자거래학회지*, 14권 1호(2009).

박종학, 조윤호, 김재경, "사회연결망 : 신규고객 추천문제의 새로운 접근법", *지능정보연구*, 15권 1호(2009).

손동원, *사회 네트워크 분석*, 경문사, 2002.

안수산, 신경식, "데이터마이닝 기법을 활용한 스팸메일 분류 및 예측모형 구축에 관한 연구", *한국지능정보시스템학회*, 7권 1호(2000).

위키백과사전, http://ko.wikipedia.org/wiki/협업 필터링.

이승훈, 김지혁, 김홍남, 조근식, "가상 커뮤니티에서 사회 관계 추론을 위한 시맨틱 웹 접근 방법", 한국지능정보시스템학회 2007년도 추계학술대회(2007), 343~352.

이승훈, 김지혁, 김홍남, 조근식, "웹 기반 소셜 네트워크에서 시맨틱 관계 추론 및 시각화", *지능정보연구*, 15권 1호(2009).

Albert, R. and A.-L. Barabasi, "Statistical mechanics of complex networks", *Rev. Mod. Phys.*, Vol.74(2002).

Butts, C. T., "Social network analysis : A methodological introduction", *Asian Journal of Social Psychology*, 2008.

Chen, C., "Visualizaing Semantic Spaces and Author Co-Citation Networks in Digital Libraries", *In Information Processing Management*, Vol.35, No.3(1999).

Domingos, P. and M. Richardson, "Mining the network value of customers", KDD, (2001), 57~66.

Faloutsos, C., K. McCurley and A. Tomkins, “Connection Subgraphs in Social Networks”, SIAM International Conference on Data Mining, 2004.

Fawcett, T., “‘In vivo’ spam filtering : A challenge problem for data mining”, Hewlett-Packard Laboratories 1501 Page Mill Road Palo Alto, CA USA, 2004.

Frey, B. J. and D. Dueck, “Clustering by Passing Messages Between Data Points”, *Science*, Vol.315(2007), 972~976.

Girvan, M. and M. E. J. Newman, “Community structure in social and biological networks”, Proc Natl Acad Sci USA, 2002.

Jaewon, C., H. J. Lee and Y. C. Kim, “The Influence of Social Presence on Evaluating Personalized Recommender System”, 한국경영과학회 추계학술대회, 2008.

Joshi, D. and D. GaticaPerez, “Discovering Groups of People in Google News”, Proceedings of the 1st ACM international workshop on Human-centered multimedia, 2006.

Kerschbaum, F. and A. Schaad, “Privacy-Preserving Social Network Analysis for Criminal Investigations”, Alexandria, Virginia, USA, 2008.

Mccallum, A., A. Corrada-Emmanuel, and X. Wang, “Topic and Role Discovery in Social Networks”, IJCAI, 2005.

Newman, M. E. J., “The structure and function of complex networks”, *SIAM Review*, Vol. 45, No.2(2003), 167~256.

Newman, M. E. J., “Detecting community structure in networks”, *Eur. Phys. J. B.*, Vol.38 No.2(2004), 321~330.

Newman, M. E. J., “Finding community structure in networks using the eigenvectors of matrices”, *Physical Review E.*, (2006), 74.

Palla, G., I. Derényi, I. Farkas and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society”, *Nature*, Vol.433(2005), 392~395.

Rahman, M. A., A. El Saddik and W. Gueaieb, “Building Dynamic Social Network From Sensory Data Feed”, *IEEE Transactions on Instrumentation and Measurement*, Vol.59, No.5(2010), 1327~1341.

Velardi, P., R. Navigli, A. Cucchiarelli and F. D’Antonio, “A New Content-Based Model for Social Network Analysis”, Proceedings of the 2008 IEEE International Conference on Semantic Computing, (2008), 18~25.

Wennerberg, P. O., “Ontology Based Knowledge Discovery in Social Networks”, Final Report, JRC Joint Research Center, 2005.

Xu, J. and H. Chen, “Criminal Network Analysis and Visualization”, *Communications of the ACM*, Vol.48, No.6(2005).

Zhang, C., W. B. Chen, X. Chen, R. Tiwari, L. Yang and G. Warner, “A Multimodal Data Mining Framework for Revealing Common Sources of Spam Images”, Journal of Multimedia, 2009.

Zhou, C., D. Frankowski, P. Ludford, S. Shekhar and L. Terveen, “Discovering Personal Gazetteers : An Interactive Clustering Approach”, Proceedings of ACM GIS, 2004.

Abstract

Managing Duplicate Memberships of Websites : An Approach of Social Network Analysis

Eun-Young Kang* · Kee-Young Kwahk**

Today using Internet environment is considered absolutely essential for establishing corporate marketing strategy. Companies have promoted their products and services through various ways of on-line marketing activities such as providing gifts and points to customers in exchange for participating in events, which is based on customers' membership data. Since companies can use these membership data to enhance their marketing efforts through various data analysis, appropriate website membership management may play an important role in increasing the effectiveness of on-line marketing campaign. Despite the growing interests in proper membership management, however, there have been difficulties in identifying inappropriate members who can weaken on-line marketing effectiveness. In on-line environment, customers tend to not reveal themselves clearly compared to off-line market. Customers who have malicious intent are able to create duplicate IDs by using others' names illegally or faking login information during joining membership. Since the duplicate members are likely to intercept gifts and points that should be sent to appropriate customers who deserve them, this can result in ineffective marketing efforts. Considering that the number of website members and its related marketing costs are significantly increasing, it is necessary for companies to find efficient ways to screen and exclude unfavorable troublemakers who are duplicate members. With this motivation, this study proposes an approach for managing duplicate membership based on the social network analysis and verifies its effectiveness using membership data gathered from real websites. A social network is a social structure made up of actors called nodes, which are tied by one or more specific types of interdependency. Social networks represent the relationship between the nodes and show the direction and strength of the relationship. Various analytical techniques have been proposed based on the social relationships, such as centrality analysis, structural holes analysis, structural equivalents analysis, and so on. Component analysis, one of the social network analysis techniques, deals with the sub-networks that form meaningful information in the group connection. We propose a method for managing duplicate memberships using component analysis. The procedure is as follows.

* Graduate School of Business IT, Kookmin University
** School of Management Information Systems, College of Business Administration, Kookmin University

First step is to identify membership attributes that will be used for analyzing relationship patterns among memberships. Membership attributes include ID, telephone number, address, posting time, IP address, and so on. Second step is to compose social matrices based on the identified membership attributes and aggregate the values of each social matrix into a combined social matrix. The combined social matrix represents how strong pairs of nodes are connected together. When a pair of nodes is strongly connected, we exepct that those nodes are likely to be duplicate memberships. The combined social matrix is transformed into a binary matrix with '0' or '1' of cell values using a relationship criterion that determines whether the membership is duplicate or not. Third step is to conduct a component analysis for the combined social matrix in order to identify component nodes and isolated nodes. Fourth, identify the number of real memberships and calculate the reliability of website membership based on the component analysis results. The proposed procedure was applied to three real websites operated by a pharmaceutical company. The empirical results showed that the proposed method was superior to the traditional database approach using simple address comparison. In conclusion, this study is expected to shed some light on how social network analysis can enhance a reliable on-line marketing performance by efficiently and effectively identifying duplicate memberships of websites.

Key Words : Duplicate Membership Management, Social Network Analysis, Sub Group, Component

저 자 소 개

강은영

현재 국민대학교 BIT 전문대학원 석사과정 재학 중이며 현대 아이엔에스(주)에서 웹서비스 기획을 맡고 있다. 관심분야는 소셜 네트워크 서비스(SNS), 소셜 네트워크 분석 및 응용, 지식경영, 사회적지식활용이다. 2010 지능정보시스템학회 추계학술발표대회에서 우수논문상을 수상하였다.

곽기영

현재 국민대학교 경영대학 경영정보학부 교수로 재직 중이다. 서울대학교 경영학과를 졸업하고 KAIST 경영과학과 및 테크노 경영대학원에서 석사 및 박사학위를 취득하였다. 주요 연구관심분야는 IT-enabled organizational agility, Social media use in organizations, Knowledge management, Social network analysis and its application 등이다.