Journal of the Korea Industrial Information Systems Research Vol. 21 No.4, Aug. 2016 :1-11
http://dx.doi.org./10.9723/jksiis.2016.21.4.001 ISSN:1229-3741

온라인 소셜네트워크를 통한 한국인의 정치성향 예측 기법의 연구

(A Study on Political Attitude Estimation of Korean OSN Users)

무하마드 에카 위자야[1], 안 희 준[2]*
(Muhammad Eka Wijaya and Heejune Ahn)

요 약 본 연구는 Facebook 사용자들의 Like활동 정보를 사용하여 정치성향을 예측하기 위한 분석 모델과 프로그램를 개발하였다. Facebook의 Ajax사용 특성 을 반영한 Facebook 크로울러를 개발하였으며, 이를 사용하여 수집된 성기고 방대한 데이터의 상관 매트릭스 정보를 효과적의 축소하기 위한 카테고리 레벨 필터링 기법을 개발하였다. 대한민국 사용자들을 대상으로 LCA (Latent class analysis) 분석한 결과 28 개의 기준 (전체 대상페이지의 3% 미만) 으로 사용자의 정치적인 극성을 상당히 정확하게 (AUC of 0.82) 예측할 수 있음을 확인하였다.

핵심주제어 : 온라인 소셜 네트워크, 성향 분석, 데이터 마이닝, 빅 데이터, 웹 크로울러

Abstract Recently numerous studies are conducted to estimate the human personality from the online social activities. This paper develops a comprehensive model for political attitude estimation leveraging the Facebook Like information of the users. We designed a Facebook Crawler that efficiently collects data overcoming the difficulties in crawling Ajax enabled Facebook pages. We show that the category level selection can reduce the data analysis complexity utilizing the sparsity of the huge like-attitude matrix. In the Korean Facebook users' context, only 28 criteria (3% of the total) can estimate the political polarity of the user with high accuracy (AUC of 0.82).

Key Words : Online Social Network, Attitude analysis, Data mining, Big-date, Web crawler

1. 서 론

1970년대 군사용으로 시작된 인터넷은 1990년대 웹을 통하여 상용화 되었으나, 2000년대 초까지도 콘텐트의 생산자와 소비자의 역할이 구분되어 있는 형태였다. 그러나 최근 10여년사이 일반 개인들의 인터넷을 통한 정보공유를 하는 형태로 발전되고 있으며, 이러한 변화의 가장 눈 에 뜨이는 현상이 Facebook, Twitter, Instagram과 같은 OSN (online social network) 서비스의 등장이다. 조사[4]에 따르면, 현재 전체 인터넷 인구의 59%이상이 하나이상의 OSN 계정을 가지고 생활하는 것으로 보고되어 있다.

이러한 OSN의 등장은 그동안 사회과학자들이 궁금해왔던 사람들의 행동패턴 분석을 대규모로

* Corresponding Author : heejune@seoultech.ac.kr
+ 이 논문은 미래창조과학부 및 정보통신기술진흥센터의 해외 ICT 전문인력활용촉진사업(IITP-2016-R0134-16-1030)의 지원을 받았음.
Manuscript receivrd Aug 15, 2016 / revised Aug 30 ,2016 / accepted Aug 31, 2016
1) 서울과학기술대학교 전기정보공학과, 제1저자
2) 서울과학기술대학교 전기정보공학과, 교신저자

쉽고, 정확하게 할 수 있는 토대를 만들어주고 있다. 본 연구의 시점에서는 Myspace, Twitter, LinkedIn, Facebook, Instagram 등이 대표적인 OSN 서비스로, 이 중에서 유독 Facebook에 대하여, 벌률, 경제, 사회, 경영, 마케팅, 정치관련 많은 연구관심 들이 모이고 있다. 그 이유는 두 가지 요인에 기인한다고 판단된다. 우선 현재 Facebook이 실제적으로 가장 많은 사용자와 일간 활동이 발생하는 OSN이다. Facebook사의 통계[8]에 따르면 2015년 5월 현재 9억 명 이상의 일간 활동사용자와, 그중 80%인 7억9천명이상의 모바일 사용자가 존재한다. 한국의 경우 2013년 6월 이후 한 달에 한번이상의 방문을 하는 사용자는 인구의 25%가 넘는 천백만 명에 달한다. 반면 초기에 강세였던 Myspace와 Twitter 등은 그 비중에 줄어들어 있다. 또한, Twitter나 Instagram은 제공하는 서비스가 단조로운데 비하여, Facebook은 매우 다양한 서비스를 제공하여 있어서, 사용자들 간의 관계와, 개인의 상태, 특성, 활동 내용을 다양한 정보를 포함하고 있다.

본 연구에서는 Facebook에서의 사용자의 활동 내역 중 Like 표현을 통하여 사용자의 정치적인 성향을 예측하는 분석 모델과 프로그램을 개발하고, 이를 바탕으로 예측성능을 확인하려고 한다. 정치성향을 본 연구에서 선택한 이유는 정치적인 성향이 사회과학에서 오랜 동안 인간관계 형성과 선거 등에 흥미로운 주제였으며, 상대적으로 분명한 극성을 나타내는 특성이 있기 때문이다. 이전부터 계속적으로 개인성격, 생물학적인 특징, 사회적인 특성 등에 대한 상관관계에 대한 연구가 있어왔다 [1, 2]. Caprara 등은 [1] 이태리의 선거결과를 개인의 성행을 바탕으로 67.2% 정확도로 예측할 수 있음을 발표하였으며, Gerber 등은 [2] 더 나아가서 인종이나 교육수준 등의 특성에 대한 영향을 같이 분석하여 발표하였다. 그러나 모두 자가보고를 통한 연구로서 자료수집의 규모의 제약, 보고 편향성 등의 한계를 가지고 있다.

기존 조사[5, 6, 7]에 따르면 상당수의 넓은 연령대의 인터넷 사용자들이 정치와 관련된 OSN 활동을 하고 있는 것으로 파악되고 있다. 본 논문은 Facebook의 그룹을 통한 정치성향을 판단 기준으로 하여 Like 활동을 사용한 사용자의 정치적 선호도를 예측하는 방법과 성능을 평가하는 것에 초점을 맞추고 있다. 본 연구의 기여는 두 가지로 주장할 수 있다. 우선 공학적으로는 Facebook에서 빅데이터 분석을 위한 데이터 수집 방법과 프로그램을 개발한 것이다. 최근 Facebook의 웹페이지 구성 특성과 보안정책의 강화로 Facebook 데이터 수집에 어려워진 환경에서 효과적인 방안을 제시한다. 두 번째는 수집된 빅 데이터를 효과적으로 분석하기 위하여 페이지들을 범주화하는 방법을 사용하여 계산 양을 줄였다는 점이다. 이를 통하여 분산 빅데이터 환경이 아닌, 공유메모리 기반 환경에서도 SVD (singular value decomposition)가 가능한 크기의 매트릭스 데이터로 구성할 수 있다. 사회과학적인 입장에서는 아시아, 특히 한국을 대상으로 OSN을 통한 정치성향 예측을 한 첫 번째 연구라는 점이다. 앞서 제시한 연구들은 주로 미국이나 유럽을 대상으로 한연구로, 저자들이 알기로는, 아시아나 한국의 환경에서 분석한 결과는 발표된 바가 없다[9, 10, 11, 12, 13]. 이러한 배경에는 중국에서 Facebook을 비롯한 주요인터넷 웹사이트를 검열 등의 이유로 허용하지 않고 있는 것[4, 14]도 원인이다.

본 논문은 제 2절에서 Facebook 관련 기존 연구들에 대하여 소개하고, 제3절에서 데이터 컬렉션하기 위하여 개발된 Facebook 크로울러에 대하여 설명하며, 수집된 데이터의 통계특성에 대하여 설명한다. 제 4장에서 데이터의 분석을 위하여 적용한 데이터 크기 감축 방법과 추정방법에 대하여 설명하고, 제 5장에서 본연구의 의미와 추후과제에 대하여 논의 한다.

2. Facebook을 사용한 연구들

2.1 시스템 구성

2004년에 등장한 Facebook은 현재 세계적으로 가장 인기 있는 OSN으로 엄청난 수의 가입자 수와 사회적 영향력, 기업의 가치를 누리고 있다. Facebook은 친구나 가족, 동료 등 아는 사람들의

소식 글과 사진 등을 공유하는 기능으로 시작하였고, 발전해나가면서 메시지 기능, 광고를 제공, 게임 등 응용 포털 기능도 제공하고 있다. 사용자들은 자신의 프로필을 설정하는 것 이외에, 1) News Feed, 2) Friends, 3) Wall, 4) Timeline, 5) Likes, 6) Message, 7) Notifications, 8) Networks, 9) Groups 등의 기능을 사용할 수 있다. 특히 본 연구에서 관심을 두는 Facebook Like는 다른 사용자들의 글, Facebook 그룹, 페이지 (영화나 음악 포함) 등에 대하여 긍정적인 감정(공감/동감/호감)을 표현하는 방식으로, 사용자들의 표현기회를 통하여 활동을 증대시킨다. 본 연구에서는 Facebook이 거의 모든 영역에 Like를 사용할 수 있으며, 데이터를 수집한 시점(2015년)에 바이너리 정보로 관심도 해석이 명확한 특성을 보이기 때문에, 분석의 대상으로 삼았다[22].

Facebook이 관련되어 최근 10여 년간 매우 다양한 연구가 행해졌으며, 지금까지의 연구를, Wilson 등은 서베이 논문[15]을 통하여 분류하여 보면, 크게 다음의 5가지로 분류할 수 있다. 1) Facebook 사용자들의 특성 분석. 예를 들어 어떤 사람들이 주고 Facebook을 사용하며, 어떤 기능을 중점적으로 사용하는 지에대한연구 (24%), 2) Facebook 사용의 동기에 대한 분석. 예를 들어 왜 Facebook을 사용하게 되는지에 대한 연구 (19%), 3) Facebook에서의 자기 표출 방식. 즉 Facebook에서 자신을 표출하는 방식과 실제 자신의 모습과의 차이 여부 (12%), 4) 인간관계 형성에 있어서 Facebook의 역할과 영향. 즉, Facebook 에서의 관계가 인간관계형성에 어떤 영향을 미치는지에 대한 연구 (27%), 5) 개인 사생활 및 정보공개에 대한 연구. 사용자들의 정보공개여부, 공개하는 정보의 종류, 그 이유 및 영향 (18%).

사회정치와 관련된 연구 중에서 가장 관심을 많이 받은 연구는 'Facebook의 정치에 대한 영향'[16, 17, 18, 19, 20]으로 보인다. 그러나 역으로 정치적인 특성이 Facebook 활동에 미치는 영향에 대한 연구는 찾아보기 힘들다. 최근에 Kosinski 등이 [21]에서 Facebook의 Like를 기반으로 개인의 특성 (나이, 인종, 정치, 종교, Big5로 대표되는 성격) 등을 예측하는 연구를 발표하였다. 이 연구는 미국인 사용자들을 대상으로 하였고, 성격설문 앱(MyPersonality)을 통하여 오랜 동안 사용자들을 리크루트 하는 방식을 사용하였다. Like를 판단기중으로 삼은 점에서는 본 연구와 공통점이 있으나, 본 연구에서는 앱 방식에 따른 문제점들을 해결하기 위하여 크로울러방식을 사용하여 분석자가 수집 조건 등을 결정할 수 있도록 하였고, 분석 방법도 대규모 분산처리 방식이 아닌, R과 Matlab과 같은 단일 PC용 분석 툴을 사용할 수 있도록 하였다.

3. 데이터 수집 방법

데이터 수집방법 대하여 설명하기 전에 우선, 대상으로 하는 대한민국 정치구조에 대하여 정의하고자한다. 대한민국의 정당제도는 다당제를 바탕으로 하고 있고 실제로 데이터를 수집한 2015년 현재에도 다수의 당이 존재한다. 하지만 데이터 수집당시 실제적으로 19대 국회(임기 2012-2016)는 여당인 새누리당이 160석 제2야당인 새정치당이 130석으로, 두 당이 298석 중에 290을 차지하고 있어 분석의 목적으로 보아 양당제 형식으로 모델링하였다. 정당의 일반적인 성향은 새누리당은 보수성향이 강한 것으로, 그리고, 새정치당은 진보성향이 강한 것으로 [23] 평가되고 있다.

3.1 Facebook에서의 정치 성향 정보의 수집의 대상 결정

Facebook 사용자의 정치적인 성향을 알아볼 수 있는 방법은 크게 4가지로 볼 수 있다. 첫 번째는 자신이 개인의 프로필에 'political views'를 표시하는 것이고, 두 번째는 정치인과 친구가 되거나 폴로우어가 되는 경우이며, 세 번째가 Facebook 페이지를 Like하는 것이고, 마지막이 정치적인 Facebook 그룹에 가입하는 것이다. Fig. 1은 각 해당 방식에 해당하는 Facebook의 페이지의 예를 표시한다. 각 방식에 대한 정의와 장단점을, Facebook API지원여부, 정보접근 범

위, 정보소스 등으로 구분하여 비교할 수 있다. 사용자 프로필을 통한 정치성향은 매우 직접적이고 정확한 방법이라고 볼 수 있으나, 실제로 아주 낮은 비율의 사용자들이 자신을 정치성향을 설정/공개 (미국 사용자의 8%, 한국사용자의 3% 미만) 하고 있기 때문에 이를 바탕으로 데이터 확보하는 것은 항계가 있을 뿐더러, 바이어스가 항한 데이터가 될 수 있다. 두 번째 원천은 국회 위원들의 친구 또는 팔로우어 리스트를 통한 방법이다. 이 경우 Facebook API를 사용하여 수집이 가능하나, 해당 정보에 접근이 가능한 권한을 확보하여야한다. 또한, 실제 해당 국회위원의 친구나 팔로우어가 국회의원과 같은 정치적 성향이 있는가에 대하여도 추가적인 확인이 필요하다. 세 번째는 정당이나 정치인의 정치페이지를 Like 한 경우를 확인하는 방법이다. 그러나 이 경우는 현재 Facebook은 API를 제공하고 있지 않아 정보를 확보하기가 쉽지 않다. 마지막 방법은 정치적 Facebook 그룹에 조인한 커뮤니티 사용자들의 정보를 확보하는 방법이다. Facebook 그룹은 공통의 관심사나 성향을 갖는 사람들이 모여서 커뮤니티를 형성하고 의견을 공유하는 공간이다. 또한 해당 그룹멤버를 확보할 수 있는 API가 존재한다. 위 의 4가지 장단점을 바탕으로 본 연구에서는 실험에 사용할 데이터 수집 대상으로 Facebook 그룹을 사용하기로 하였다.

Fig. 1 Political information Source in Facebook

3.2 Facebook 정보수집

실험에서 기존의 연구에서 Facebook 상의 데이터 수집은 크게 두 가지 방법으로 분류된다. 하나는 Facebook 앱을 개발하고, 사용자들에게 이 앱을 광고하여, 사용하도록 할 때 얻어진 권한을 통하여 사용자의 개인 정보를 확보하는 방법[21]이 있으며, 나머지는 크로울러를 사용하여 웹에 접근하듯이 Facebook페이지에 접근하여 데이터를 수집하는 방법[24, 25]이다. 앱 기반의 방식은 사용자가 앱을 허용하고 권한을 부여하는 과정을 수행하여야하기 때문에 데이터 수집에 오랜 시간이 필요하며, 사용자들에게 해당 앱을 알리고, 사용하도록 하는 과정이 필요하여 리쿠르트 과정이 필요하다. 또한, 사용자를 선별하는 과정이 사용자 측에 있으므로, 사용자의 특성이 전체 사용자를 대표하지 못하고 편향될 수 있는 문제가 발생한다. 따라서 본 연구에서는 Facebook 크로울러를 개발하여 데이터를 확보하였다.

개발된 크로울러는 Scrapy 프래임워크 [27]와 Selenium [28]을 결합하여 구성하였다. Fig. 2은 본 연구를 위하여 제작한 3 계층 구조의 Facebook 크로울러를 도시한다. 백엔드 계층은 수집한 데이터 저장소이고, 중간 계층은 Scrapy 프레임워크와 처리 모듈인 Spider 들이다. 상단 계층은 Facebook 웹상에서 자동화된 웹 서핑을 시뮬레이션 하는 selenium 브라우저이다. Scrapy는 확장이 가능한 오픈소스 웹크로울러로서 동시에 여러 개의 HTTP 요청을 사용할 수 있고, 사용자가 기능을 쉽게 황장이 가능하도록 구성되어 있다. Selenium은 확장이 가능한 웹브라우저 엔진으로, Facebook 페이지가 여러 개의 Ajax앱의 mash-up으로 구성되어 있기 때문에 필요하다. 즉, Scrapy를 사용하여 JavaScript 와 html 문서를 다운로드하고, javascript 해당 기능은 Selenium을 사용하여 처리하는 구조를 가지고 있다.

API를 사용하여 Facebook 접근하기 위해서 (프라이버시) 토큰을 사용하여야하는데, 이 토큰의 생명주기가 수 시간으로 비교적 단시간이다. 이를 해결하기 위하여 본 연구에서는 Facebook 앱을 개발하고, 이를 통하여 장시간 사용이 가능한 토큰을 확보하여 사용하였다. Facebook 싸이트는 웹브라우저에서 오는 요청들의 패턴을 분

석하여 크로울러인지를 감시하는 기능을 사용하고 있다. 크로울러로 감지되는 경우 해당 IP 주소로 오는 요청을 블록하기 때문에, 이를 방지하기 위하여 약 5 초의 간격을 두고 요청을 하도록 프로그램을 구성하였다.

Facebook 그룹은 공개그룹과 비공개 그룹 두 가지 형태가 있다. 본 연구에서는 공개 그룹에 가입된 사용자들만을 대상으로 정보 수집을 수행하였다 (Table 2). 이렇게 수집된 사용자들의 Like 정보를 API를 사용하여 확보하고, Like수가 5개미만인 경우는 분석 대상에서 제외하였다.

- Facebook 정치 그룹 리스트: 한국 정치 그룹들의 리스트는 두 정당의의 이름으로 Facebook 서치를 사용하여 얻었다 (Table 1).
- 그룹의 멤버 리스트: 자동 데이터 수집 절차는 이 단계에서부터 이루어지는 데, Facebook API를 사용하면 해당 그룹의 사용자들을 얻을 수 있다.
- 사용자 정보: 사용자의 신정 정보는 사용자 신상 페이지 웹으로 접근하여 얻을 수고 있다. 이를 위하여 Selenium 브라우저를 사용하고, 앞서 획득한 id를 사용하여 'About' 페이지에서 사용자 명이나, 아이디를 확보한 후 공개한 교육, 성, 사는 곳 등을 덩을 수 있다
- 사용자의 Like 페이지 정보: 사용자가 Like한 페이지에 대한 정보는 사용자의 페이지에서 확보가 가능하다. 추가로 해당 페이지와 관련된 정보, 즉, 페이지의 범주정보, Likes 수 등도 얻을 수 있다.

Table 1 Crawled group and the corresponding members' numbers.

Group name (English Translation)	Group Id	Supporting Party	# of Members
Saenuri, For big success	120869327985461	Saenuri	1846
Saenuri party	233928133357455	Saenuri	4640
Saenuri revolutionary	772486989449229	Saenuri	1137
Disorganization of Saenuri	1469075306695014	Saejungchi	298
Saejungchi, Gwanju-city	175703049123755	Saejungchi	35
Saejungchi, new Korea	202265119855332	Saejungchi	9286
New political party of the people	222703814585120	Saejungchi	3146
Saejungchi, Ganwon-province	277645229012536	Saejungchi	203
Saejungchi, Geoje	472198762886131	Saejungchi	56
Saejungchi Union	667965906656555	Saejungchi	49

3.3 수집된 데이터의 특성

수집된 사용자는 총 1332 명으로 681명은 새누리당 지지자, 651명은 새정치연합 지지자들로 구성되었다. 성비로는 725 명은 남성 246 명은 여성, 나머지는 성별정보를 확보하지 못하였다. 성

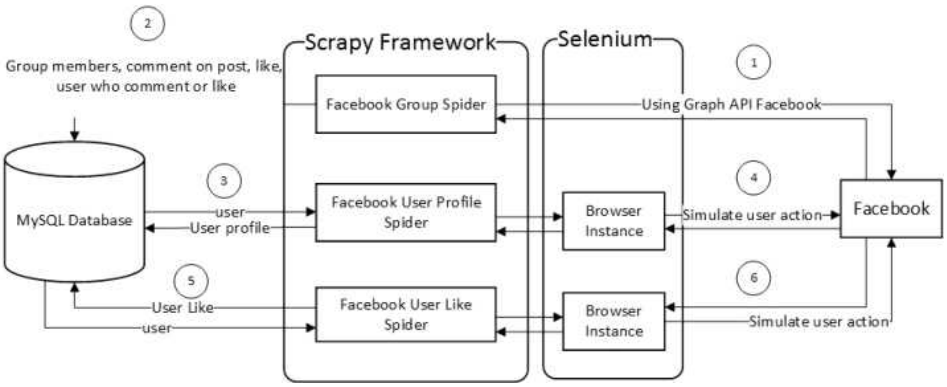

Fig. 2. Facebook Crawler Designed

별의 불균형은 아마도 성에 따른 정치에 대한 관심도가 차이가 있음을 반영하는 것으로 보인다. 이들의 페이지 Like의 평균 262, 표준편차는 644이다.

수집된 데이터의 중요한 특징은 성기다(sparse)는 점과 크기가 매우 크다는 점이다. 성긴 매트릭스의 특성은 전체 페이지에 비하여 한 명의 사용자가 선택한 Like의 수가 작기 때문이다. 이런 경우에 일반적인 회귀 모형방식으로는 예측하기 위한 정보가 충분하지 않은 문제가 발생한다. Kosinski 등은 그들의 연구[21]에서 SVD (singular-value-decomposition) 방법을 사용하였다. LCA (Latent Class analysis) 분석 방법인 SVD을 사용하면 유사한 특성을 갖는 페이지들의 영향을 합하여 고려가 가능해진다.

그러나, 대상 매트릭스가 갖는 또 다른 특성이 크기에 따른 계산량 문제는 SVD을 사용하여도 여전히 남는다. 본 연구에서는 데이터 수집의 한계로 1332x95675 개를 수집하였으나, 일반적으로 이보다 훨씬 큰 데이터를 확보할 수도 있다. 이러한 빅데이터 분석을 위해서는 Hadoop과 같은 분산 병렬처리환경을 사용하거나, 데이터양을 줄이기 위한 알고리즘을 개발하여야한다. Kosinski의 연구에서는 주요 관심사가 사회 심리학적인 결과 예측에 있었고, 계산기술적인 언급은 전혀 없어서, 어떤 방식을 사용했는지 파악하는 것이 불가능하다. 본 연구에서도 SVD 방식을 이용하나 Facebook 페이지들이 갖는 범주 특성을 반영하여 계층적인 페이지 감축 방법을 사용한다. Fig. 3은 범주 선택알고리즘 방식을 도식화하고 있다.

4. 정치 성향 예측 모델 개발

본 연구에서는 출력값이 이진값을 갖기 때문에, Likes 정보를 통한 정치성향 예측을 위하여 로지스틱 회귀분석 방법에 기초한 분류알고리즘을 개발하였다.

$$a_u \sim l_{u,1} + l_{u,2} + l_{u,3} + \cdots + l_{u,p} + \cdots + l_{u,M} \quad (1)$$

여기서 $l_{u,p}$은 입력 변수로 해당 페이지에 대한 Like 여부(0, 1) 이고,

$$l_{u,p} \equiv \begin{cases} 1 \; user\, u\, liked\, page\, p. \\ 0 \; o.w. \end{cases} \quad (2)$$

출력변수 a_u는 사용자의 정치성향으로 다음과 같이 정의된다.

$$a_u \equiv \begin{cases} 1 \; ruling\; party^{\,port} \\ 0 \; opposite\; party^{port} \end{cases} \quad (3)$$

수집된 데이터들은 사용자-페이지의 2진 행렬로 표현된다.

$$L = \begin{pmatrix} l_{1,1} \cdots l_{1,M} \\ \vdots \;\; \ddots \;\; \vdots \\ l_{N,1} \cdots l_{N,M} \end{pmatrix} \quad (4)$$

Table 2 User-Page matrix example

Facebook ID	Chu Shin su[1]	Yuna Kim[2]	Lee Jaes u[3]	President[4]	Supporting Party
1	1	0	0	0	Saenuri
2	0	0	0	0	Saenuri
3	1	1	0	1	Saenuri
4	0	1	0	1	Saenuri
5	0	1	0	0	Saejungchi
6	0	0	1	0	Saejungchi
7	0	0	1	0	Saejungchi
8	0	0	0	0	Saejungchi

[1]Korean US major league baseball [MLB] player, [2]Famous Olympic Gold Medalist,[3]Citizen is Answer,[4]For President Geunhae Park

4.1 사용자 범주 매트릭스

Facebook 페이지는 주제에 따라서 범주를 가지고 분류되어 있다. 예를 들어서, 현재 미국 대통령인 Barack Obama나 한국 대통령인 Park Geun-Hye 페이지는 '정치'라는 범주로 구분이 되어 있으며, 영화배우 Kim Hee-sun, Emily

Watson 등은 'actor/director' 범주로 분류되어 있다. 의미상으로 각 범주에 따라 정치적인 성향에 크게 영향을 미칠 수도 그렇지 않을 수도 있을 것이라는 추론이 가능하다. 따라서 본 연구에서는 각 카테고리 $c=1,2,...,K$와 페이지의 연결 $C_k=\{p(c)\}$관계를 통하여, 사용자-페이지 매트릭스를 사용자-범주 매트릭스로 전환하였다. Table 3은 사용자-범주 매트릭스의 예이다. 아이디. 98501288485XXX의 는 'actor/director' 분류에서 두 개의 Facebook 페이지를 'athlete' 페이지에서 한 개의 페이지를 'automobiles'에서 6개의 페이지를 'community' 범주에서 9 개의 페이지를 Like 하였다.

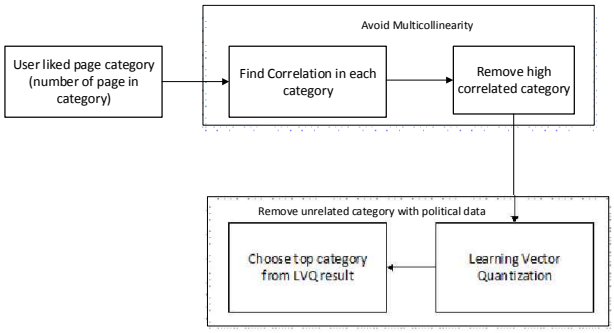

Fig. 3 Like page category selection analysis flow.

Table 3 User-category matrix example

Facebook ID	actor	athlete	auto-Mobile	community	related party
1	2	1	6	9	Saenuri
2	12	4	18	42	Saenuri
3	1	2	3	88	Saenuri
4	4	1	0	2	Saenuri
5	0	1	0	0	Saejungchi
6	6	5	6	18	Saejungchi
7	0	2	10	35	Saejungchi
8	0	0	0	0	Saejungchi

4.2 Multi-collinearity 문제 처리

로지스틱 회기 방법을 적용하기 전에 다중공선성(multi-colinearity)현상을 제어하기 위하여, 범주 데이터간의 상관도를 확인하는 것이다. 상관도는 다음 식으로 정의된다. 모든 조합의 범주에 대하여 상관도를 계산하여 상관도가 크게 되면 (본 연구에서는 0.5를 사용), 그 중에서 Like 수가 많은 것만을 변수로 남기는 방식을 사용하였다. 예를 들어 'community/government' 범주와 'cause' 범주의 상관성이 높았다. 'Cause'는 사람들의 생활 방식, 예를 들어 'save the earth'나 'save the children'등과 같은 페이지들로 구성되어 있다. 따라서 이 두 가지 범주가 상관성이 높을 것임을 추론이 가능하다고 생각된다. 다중공선성 제거 과정을 통하여 총 226 개의 Facebook 범주는 109개로 정리되었다.

4.3 사용자-범주 매트릭스

다음 단계는 출력 값인 정치성향과 상관도가 극히 작은 입력 범주를 제거하는 단계이다. 이 과정에는 회기 결정 영역을 강화하기 위한 신경망 알고리즘인 'learning vector quantization' [29]를 사용하였다. 그 결과 109개중 단 25개의 범주만이 출력과 영향이 있음을 보였다. 이들 중 가장 높은 상관도를 보이는 5개는 Fig. 4와 같이 TV network, sport league, games toys, personal website로 나타났다. 이 과정을 거쳐 얻어진 남은 페이지는 처음 95.675개에서 2.793개로 3%정도로 줄었다.

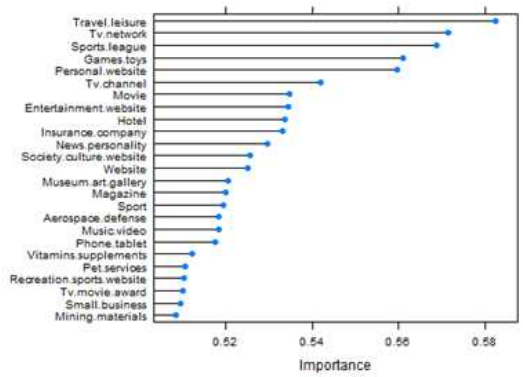

Fig. 4. Importance to political attitude of Facebook page categories.

4.4 사용자-페이지 매트릭스 SVD

Table 3 는 8개의 사용자-페이지 매트릭스는 각 정당 지지자들이 다른 페이지 선호도를 보이는 것의 예이다. 예를 들어 'Lee Jaesu- Citizen is the answer' 페이지는 주로 새정치 지지자들이 좋아하는 것으로 보였다. 그러나 여전히 페이지수가 많기 때문에, 성분 분석을 위하여 SVD을 수행하였다.

$$L^c = U \Lambda V^T \quad (5)$$

Λ 매트릭스의 차원을 줄여가면서, 전체 페이지 수로부터 줄여가면서 실험한 결과 Fig. 5와 같이 90% 정보량이 유지되는 값은 약 450개 정도인 것으로 확인되었다.

4.5 회기 분석 및 예측 성능

SVD을 통하여 얻은 284 기준요소를 바탕으로 하여 AIC(Akaike Information Criterion) [30]을 최소화하기 위한 절차를 적용하였고, 그 결과를 10-folds 상호검증방법에 의하여 로지스틱 회기 분석을 적용하였다. 최종 성능은 AUC (Area Under Curve)를 사용하였다. AIC 최적화 방법 [30]을 사용한 결과 신뢰도 $p < 0.05$에서 28 개의 기준을 얻었다. Fig. 6는 AUC 그래프로서, AUC = 0.82, sensitivity=0.86, specificity = 0.79를 보였다.

회귀 모델의 계수는 새누리당의 지지정도에 대한 기여도를 나타낸다. 모든 기준이 다수개의 페이지로 부터의 영향이 결합되어 있긴 하지만, 의미 분석을 위하여 SVD의 V행렬로부터 기여도 분석을 해볼 수 있다. 기여도 0.4를 기준으로 분석하여 본 결과 새누리당 지지자들은 'Pop music', 'ZIEN ART SPACE', 'Pig-year born woman', 'GMA Network' 과 같은 페이지들을 Like하였고, 반면 새천년지지자들은 'KNN', 'God's Not Dead', 'Jetsetter', 'The Guy coming', 'Uppy Look', 'bed scene, how deep did you see?)' 등의 페이지를 Like 한 것으로 분석되었다.

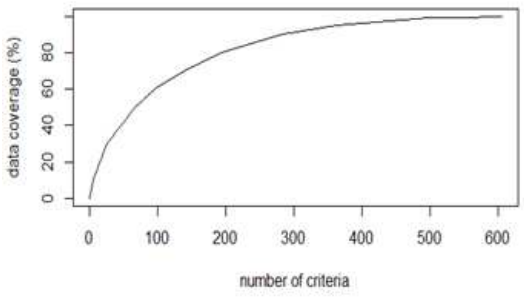

Fig. 5 Selected criteria and data coverage (spectral powers).

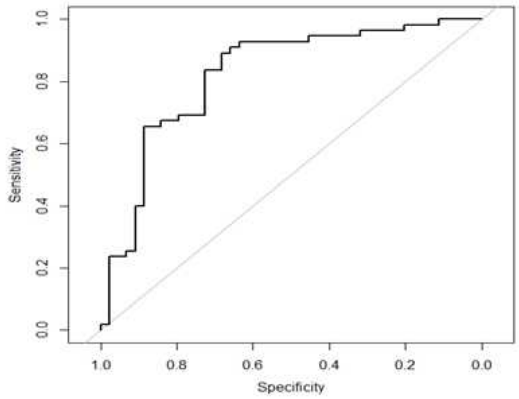

Fig. 6 Classification Accuracy of Political attitudes from Facebook Like (AUC)

4.6 계산 요구량 평가

본 연구에서 제안한 범주기반 매트릭스 축소방법은 계산에 필요한 메모리와 CPU 사용량을 크게 줄임을 확인할 수 있다. 사용자의 수는 변화가 없으나, 분석에 사용할 페이지의 수는 3%로 줄었다. 본 연구에서의 계산은 공개 데이터 마이닝 및 통계 툴인 R을 사용하여 Intel Core i7 CPU, 8GB RAM, 1TB HDD, Windows 7 OS 환경에서 처리되었으며, 10-fold 검증에서 한 번의 계산에 필요한 시간은 10분 이하로 측정되었다. 반면, 일반적으로 동일한 컴퓨터 환경에 매트릭스 축약방법 없이 SVD를 사용하면 메모리 문제로 계산이 불가능할 뿐 아니라, 계산이 되더라도 수 시간 이상의 계산을 필요로 한다[31].

5. 결 론

본 연구의 성과는 공학과 사회과학적인 특면의 두 가지 측면으로 정리할 수 있다. 공학적인 측면에서 본 논문에서 Ajax기반의 메쉬업 형태의 Facebook의 웹 데이터를 수집하는 툴을 개발하였고, 이를 바탕으로 사용자들의 공개된 Like와 신상정보를 획득할 수 있음을 확인하였다. 또한 수집된 빅데이터의 볼륨을 줄이고 이를 효과적으로 처리하는 알고리즘을 개발하였다. 최근 Hadoop이나 Spark과 같은 분산처리기반의 빅데이터 플랫폼이 제안되어 확산되고 있는 것은 사실이나, 일반적으로 많은 사람들이 이러한 분산처리환경을 접근하기가 어렵고, 이러한 환경에서 제공되는 해법들도 아직은 제한적이다[3, 31]. 반면 R이나 MATLAB을 사용한 툴은 보다 보편적이며, 데이터의 크기가 일정 범위 이하로 제한한다면 월등히 높은 성능을 보여준다. 본 연구에서는 이러한 점을 고려하여 범주 기반으로 매트릭스 크기를 줄이는 방법을 제안하였고, 이를 통하여 방대한 양의 수집된 데이터를 기존의 패키지들을 활용하여 처리할 수 있도록 하는 절차를 개발하였다. 따라서 이를 활용하면, 연주자들이 관련 연구를 수행하고 확장할 수 있으리라 생각된다.

사회학적인 측면에서 본 연구는 아시아, 특히 한국에서의 Facebook을 사용한 정치성향 분석기법을 적용한 첫 번제 연구로서 Facebook과 같은 OSN의 데이터가 사용자의 정치 성향을 매우 정확히 예측할 수 있음을 확인하였다. 정치적인 페이지를 제거하고 나서 평가한 성능 평가에서 AUC = 0.82의 상당히 좋은 성능을 확인하였다. 저자들은 이러한 Like를 사용한 성향분석이 정치적인 특성 뿐 아니라 상품이나 회사의 선호도 등을 포함한 다양한 분야로도 확장할 수 있다고 생각한다. 예를 삼성 폰과 애플폰의 선호도, 브라질 축구팀과 독일 축구팀의 호감도 등을 예측하는데 사용할 수 있다고 보인다.

본 연구의 한계점은 최선을 다했음에도 불구하고 전체 국내 Facebook 사용자들을 대상으로 데이터 수집과 분석을 수행하지 못해서 데이터의 대표성에 제약이 있다는 점과 2개의 극성을 갖는 환경에서 데이터 분석을 하였다는 점이다. (데이터 수집 당시 2개 이상의 정당이 있었으나, 편의상 두 개의 당만을 대상으로 하였음)

References

[1] G. V. Caprara, S. Schwartz, C. Capanna, M. Vecchione, and C. Barbaranelli, “Personality and politics: Values, traits, and political choice,” Political psychology, Vol 27, pp. 1-28, 2006.

[2] A. S. Gerber, G. A. Huber, D. Doherty, C. M. Dowling, and S. E. Ha, “Personality and political attitudes: Relationships across issue domains and political contexts,” American Political Science Review, Vol. 104, pp. 111-133. 2010.

[3] J. H. Lee, J. Choi, and D. H. Koo, "An Empirical Evaluation Analysis of the Performance of In-memory Bigdata Processing Platform", Journal of the Korea Industrial Information System Society, Vol. 21, No. 3, pp. 13-19, 2016.

[4] K. Hampton, “Social networking sites and our lives. PEW Internet and American Life Project,” http://pewinternet.org/Reports/ 2011/ Technology-and-social-networks.aspx, 2015.

[5] E. Quintelier and Y. Theocharis, “Online political engagement, Facebook, and personality traits,” Social Science Computer Review, Vol. 31, pp. 280-290, 2012.

[6] M. Conroy, J. T. Feezell, and M. Guerrero, “Facebook and political engagement: A study of online political group membership and offline political engagement,” Computers in Human Behavior, Vol. 28, pp. 1535 - 1546, 2012.

[7] G. J. Gulati, and C. B. Williams, “Social Media and Campaign 2012: Developments and Trends for Facebook Adoption,” Social Science Computer Review, Vol. 31, pp. 577-588, 2012.

[8] Facebook Inc., “Facebook Company Info”, http://newsroom.fb.com/company-info/, 2015.

[9] C. L. Hsu and H. W. Park, “Mapping online social networks of Korean politicians,” Government Information Quarterly, Vol. 29, pp. 169-181, 2012.

[10] A. French, “An Empirical Analysis Evaluating Trust in Friends Versus the Community in the Social Networking Context.” J. Internet Electronic Commerce Research, Vol. 13, No. 1, pp 173-197, 2013.

[11] D. L. Painter, “Online political public relations and trust: Source and interactivity effects in the 2012 U.S. presidential campaign,” Public Relations Review, 2015.

[12] L. Vesnic-Alujevic, “Political participation and web 2.0 in Europe: A case study of Facebook,” Public Relations Review, Vol. 38, pp. 466-470, 2012..

[13] J. Vitak, P. Zube, A. Smock, C. T. Carr, N. Ellison, and C. Lampe, “It’s Complicated: Facebook Users’ Political Participation in the 2008 Election,” Cyber-psychology, Behavior, and Social Networking, Vol. 14, pp. 107-114, 2011.

[14] Wikipedia, “Websites blocked in mainland China,” https://en.wikipedia.org/wiki/Websites_blocked_in_mainland_China, 2015.

[15] R. E. Wilson, S. D. Gosling, and L. T. Graham, “A Review of Facebook Research in the Social Science,” Perspectives on Psychological Science, Vol. 7, pp. 203-220, 2012.

[16] B. Williams and G. Gulati, “The political impact of Facebook: Evidence from the 2006 midterm elections and 2008 nomination contest,” Politics and Technology Review, Vol. 1, pp. 11-24, 2008.

[17] D. Kim, and T. J. Johnson, “A Victory of the Internet over Mass Media? Examining the Effects of Online Media on Political Attitudes in South Korea,” Asian Journal of Communication, Vol. 16, pp.1-18, 2006.

[18] Y. Kim, “The contribution of social network sites to exposure to political difference: The relationships among SNSs, online political messaging, and exposure to cross-cutting perspectives,” Computers in Human Behavior, Vol. 27, pp. 971-977, 2011.

[19] T. Macafee, “Some of these things are not like the others: Examining motivations and political predispositions among political Facebook activity,” Computers in Human Behavior, Vol. 29, pp. 2766-2775, 2013.

[20] N. Pennington, K. L. Winfrey, B. R. Warner, and M. W. Kearney, “Liking Obama and Romney (on Facebook): An experimental evaluation of political engagement and efficacy during the 2012 general election,” Computers in Human Behavior, Vol. 44, pp. 279-283, 2015.

[21] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” in Proceedings of the National Academy of Sciences, Vol. 110, pp. 5802-5805. 2013.

[22] B. S. Kim, S.-H. Han, and Y.-S Kang, “Exploring Purchase Behavior of Digital Items and Actual Usage in a Social Network Site: A Longitudinal Perspective,” The Journal of Information Systems, Vol. 21, No. 2, pp. 97-114, 2012.

[23] M. E. Manyin, “South Korean Politics and Rising "Anti-Americanism": Implications for U.S. Policy Toward North Korea,” Congressional Research Service, 2003

[24] M. Gjoka, M. Kurant, C. T. Butts, and A.

Markopoulou, “Practical recommendations on crawling online social networks”, IEEE Journal on Selected Areas in Communications, Vol. 29, pp. 1872-1892, 2011.

[25] S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, (2011). “Crawling Facebook for social network analysis purposes,” in Proceedings of the International Conference on Web Intelligence, Mining and Semantics, pp. 52:1-52, 2011.

[26] J. W. van Dam, and M. van Vekdeb, “Online Profiling and Clustering of Facebook Users,” Decision Support Systems, Vol. 70, pp. 60-72, 2015.

[27] Scrapy, “Scrapy Architecture” http://doc.scrapy.org/en/latest/topics/architecture.html, 2015.

[28] Selenium, “Selenium Introduction,” http://www.seleniumhq.org/docs/01_introducing_selenium.jsp#introducing-selenium, 2015.

[29] A. Rajaraman, and J. D. Ullman, Mining of Massive Datasets, New York, Cambridge University Press, 2012.

[30] M. B. Reddy and S. Reddy, “Dimensionality Reduction: An Empirical Study on the Usability of IFE-CF (Independent Feature Elimination by C-Correlation and F-Correlation) Measures,” International Journal of Computer Science, Vol. 7, pp. 74 - 81, 2010,

[31] J. H. Lee and H.-K. Lee, “A study on unstructured text mining algorithm through R programming based on data dictionary,” Vol. 20, No. 2, pp. 113-124, 2015.

무하마드 에카 위자야 (Muhammad Eka Wijaya)

- 정회원
- 인도네시아, Institut Technology Sepuluh Nopember, 정보공학 공학사.
- 서울과학기술대대학교 전기정보공학과 공학석학사
- (현) 인도네시아 프리랜서 개발자
- 관심분야 : 데이터 마이닝, 웹서비스

안 희 준 (Heejune Ahn)

- 정회원
- KAIST 전기정보공학과 박사
- 서울과학기술대학교 전기정보공학과 정교수
- 관심분야 : 인터넷 프로토콜, 영상처리, 데이터 마이닝