한국빅데이터학회지
제1권 제1호(창간호), 2016, pp. 61-66

페이스북 마케팅 활용 방안에 대한 연구: 페이스북 '좋아요' 기능과 인구통계학적 정보 추출

The Study of Facebook Marketing Application Method: Facebook 'Likes' Feature and Predicting Demographic Information

유성종[1] · 안세은[1] · 이준기[2†]

연세대학교 정보대학원 석사과정[1]
연세대학교 정보대학원 교수[2]

요 약

최근 기업들이 빅데이터를 활용하여 효과적인 마케팅 전략을 전개함에 있어서, 고객의 세부정보를 기반으로 하는 개인화된 마케팅 전략을 활용하고 있다. 하지만 프라이버시 및 개인정보 유출위협이 커짐에 따라 소셜 네트워크 사이트(Social Network Site, 이하 SNS)에서 계정의 개인정보 항목을 삭제하거나 정보공개수준을 통제하는 경향이 높아지고 있다. 이로인해 기업의 마케팅 담당자들은 고객의 세부정보를 파악하는 것에 어려움을 겪고 있다. 본 연구에서는 SNS 중에서 가장 많은 회원 수를 보유하고 있는 Facebook에서 제한된 정보를 바탕으로 성별을 예측하는 분석방법론을 도출하고자 하였다. 본 연구에는 측정도구로 Gaussian RBF, nFactors, randomForest, 그리고 5-fold cross-validation 사용하였다. 그 결과, 성별은 75%, 연령대는 97.85%로 '좋아요' 정보만을 가지고 성별과 연령을 예측할 수 있었다. 즉, 사용자들의 어떠한 세부정보 없이, Facebook의 '좋아요'의 정보를 가지고 인구통계학적인 정보를 추론할 수 있었다. 본 연구의 결과를 바탕으로 개인정보 수집에 어려움을 겪고 있는 기업 및 마케팅 담당자들에게 유용한 가이드 라인을 제시 할 수 있을 것으로 기대한다.

■ 중심어 : 사회 관계망, 좋아요, 기계 학습, 빅데이터, 은닉 마르코프 모델

Abstract

With big data analysis, companies use the customized marketing strategy based on customer's information. However, because of the concerns about privacy issue and identity theft, people start erasing their personal information or changing the privacy settings on social network site. Facebook, the most used social networking site, has the feature called 'Likes' which can be used as a tool to predict user's demographic profiles, such as sex and age range. To make accurate analysis model for the study, 'Likes' data has been processed by using Gaussian RBF and nFactors for dimensionality reduction. With random Forest and 5-fold cross-validation, the result shows that sex has 75% and age has 97.85% accuracy rate. From this study, we expect to provide an useful guideline for companies and marketers who are suffering to collect customers' data.

■ Keyword : Social Networks, Likes, Machine Learning, Big Data, HMM

Ⅰ. 서 론

1.1 연구배경

전 세계인들의 소통에 많은 기여를 하는 SNS가 꾸준한 사랑을 받아오면서, 페이스북 또한 점진적으로 꾸준한 성장을 하고 있다. 페이스북의 자체 3분기 성적 발표에 의하면 2015년 9월, 페이스북을 매일 사용하는 유저(Daily active users; 이하 DAUs)들은 전년 대비 17% 증가하여 평균 약 10억만 명이었고 매일 모바일을 통해 페이스북을 이용 유저(Mobile DAUs)는 전년 대비 27% 상승한 평균 8억 9천 4백만 명이었다[12]. 이렇듯

2015년 12월 1일 접수; 2016년 1월 11일 수정본 접수; 2016년 1월 18일 게재 확정.
† 교신저자 zoonky@gmail.com

페이스북의 사용자가 많아지면서 페이스북의 주 기능인 ‘댓글’, ‘콘텐츠 업로드’, ‘좋아요’, ‘공유’ 등은 그 자체로 양질의 빅데이터가 되었고, ‘좋아요’ 기능으로 인해 자신의 견해를 나타낼 수 있게 되었다[5]. 페이스북의 누적 사용자들이 많아지면서 다양한 흔적과 견해를 남길 수 있는데, 페이스북에서 생성된 데이터를 바탕으로, 기업에서는 제품에 대한 소비자 분석 및 마케팅 채널로서의 활용이 이어지고 있으며, 사회과학 연구에서도 SNS의 데이터를 바탕으로 연구가 진행되어오고 있다.

하지만 최근 개인정보 유출 사고 및 위협으로 인해 이용자가 통제를 통해 개인정보의 제공을 제한하면서 각 기업 및 연구분야에서 데이터 분석에 어려움을 겪고 있다. 퓨 인터넷의 조사결과에 의하면 실제 소셜미디어 사용자의 71%가 자신의 프로파일 수정 및 정보공개수준을 통제하고 있는 것으로 나타났다[16].

1.2 연구목적

기업은 단순히 마케팅뿐만이 아닌 제품 디자인부터 기업의 가치 형성에 있어 고객들의 정보를 통해 성장한다. 기업의 발전에 있어 마케팅 강화 및 전자상거래 등 경영전략차원에서 개인정보의 수집이용 및 제3자 제공이 증가하고 있다[1, 9]. 이처럼 기업에서는 마케팅을 위해 사용자들의 기본정보를 수집하며, 다양한 SNS 채널을 통해서 기업의 상품들을 사용자에게 마케팅한다. 페이스북의 주요 특징으로는 페이지를 통한 고객관계 형성 채널, 친구맺기를 통한 관계 형성, 고객 참여 중심 활용 등이 있다. 마케팅 수단으로써의 페이스북의 주요 활용 방법으로는 기업 소식 및 상품 소개, 기업 및 제품 브랜드 구축, 고객과의 대화 채널, 이벤트 진행 등이 있다[7, 14].

특히 페이스북의 경우, 자체 실적 조사에 따르면 전년 대비 매일 사용하는 유저(DAUs)가 17% 증가하고 있다고 하였으며, 기업들은 이러한 거대 SNS채널인 페이스북을 이용하여 마케팅을 실시하고 있다. 하지만 온라인에서 개인정보의 불성실한 관리 혹은 해킹으로 인해 자신이 모르는 사이에 남용되거나 유통되어 프라이버시의 침해로 인해 피해가 발생하였고, 개인정보의 문제가 중요해졌다[4]. 이런 문제로 인하여 사용자의 정보 수집이 대부분 불가능하게 되면서 마케팅 역시 어려움을 느낄 것이라 예상한다.

기존의 페이스북을 활용한 마케팅의 연구를 살펴보면, 해외 연구인[15] 논문은 성별, 나이, 인종, 정치적 선호도, 성향 등을 유추하였고, 국내 연구인[19]는 정치적 성향을 유추했다.

따라서 본 연구에서는 마케팅 수단에서 가장 중요한 사용자들의 성별과 연령을 페이스북 사용자들의 좋아요(Likes) 정보만을 바탕으로 정확히 추출하여 빅데이터 관리 및 마케터들에게 보다 편리한 정보를 제공 할 수 있는 방안을 제시하고자 한다.

Ⅱ. 이론적 배경

2.1 Social Network Site(Facebook)

소셜미디어는 인터넷 상에서 사용자들이 각자의 콘텐츠를 제작하고(User-Generated Contents) 서로 정보교류를 하는 어플리케이션 집단이다[13]. 다양한 매체들을 통해 사용자들의 개별적 참여로 콘텐츠들은 만들어진다. 소셜 미디어라는 큰 틀 안에는 소셜 네트워킹 사이트, 소셜 콘텐츠 쉐어링 등을 포함하고 있다.

다양한 형태로 나누어진 소셜 미디어 중에서 많은 사람들의 삶과 연관되어 있는 Social Networking Site (SNS)는 웹 상에서 사용자 개개인이 남기는 표현, 타인과의 소통, 소외감에서 벗어나게 도와주는 등 사용자가 사용하는데 있어 많은 공허함은 채워주는 도구로써 사용되고 있다[6]. 프로필과 연결로 기본 구성되어 있는 SNS는 사용자들이 온라인 상의 인맥 네트워크를 확장하여 각자 유용한 정보 교류를 할 수 있도록 도와주는 서비스라고 정의할 수 있다[8]. SNS에서 사용자들은 해쉬태그, 댓글, ‘좋아요’와 같은 기능을 사용함으로써 다른 사용자들과의 콘텐츠 공유를 한다[6].

대표적인 SNS로 많은 사용자를 보유하고 있는 Facebook은 현재 전 세계 사람들의 소통 도구로 사용되고 있다. 또한, 단순히 많은 가입자를 보유함을 떠나서 SNS 중독이라는 표현까지 나오게 한 주범이자 사용자 절반 이상이 최소 1일 1회 로그인을 한다[6].

2.2 Facebook에 관한 분석 연구

SNS의 확산으로 인해 사용자들에 대한 데이터 풀이 형성됨으로써 SNS를 하나의 정보시스템의 형태로 해석되고 있다[2]. 이를 통하여 마케팅, 광고, 사회과학과 같은 다양한 분야에서 데이터를 통하여 전략을 세우는 등 SNS를 통하여 유의미한 결과를 전달할 수 있게 되

었다. 그러던 중 사용자의 데이터를 통하여 유추해낼 수 있는 한계점이 어디인가에 대한 연구가 진행되었다.

오프라인 기록과 설문을 통하여 성향 분석 연구 와는 달리 웹 상의 단순한 데이터만을 통해 성향 유추를 위해서는 대량의 데이터를 통한 분석이 필요하다. 대표적인 비영리단체인 myPersonality는 Facebook 사용자들을 중심으로 웹상에서의 개인정보, '좋아요', 사진, 과거 모든 사용 기록들, 및 간단한 성향 테스트 설문까지 다양하고 많은 데이터를 보유하고 있다.

Kosinski[15] 논문은 myPersonality를 통하여 58,466명의 데이터를 얻어 Facebook에서의 사용자의 '좋아요' 데이터만을 통하여 사용자의 성향에 대한 유추 가능성에 관한 논문을 발표하였다. 사용자들의 신상정보, 및 심리 테스트에 관한 데이터들을 토대로 높은 확률로 성별, 나이, 인종, 정치적 선호도, 및 몇 개의 성향을 유추하는데 성공했다.

국내에 비슷한 사례로는 Wijaya[19]의 web crawling을 이용한 사용자 '좋아요'를 통한 사용자의 정치적 성향을 유추에 관한 연구가 나왔다. 결과적으로 정치적 성향을 어느 정도 유추하는 데에 성공했다.

Ⅲ. 연구방법론

3.1 연구설계 몇 연구대상

SNS상에서 사용자가 기재한 '좋아요' 데이터만을 통하여 사용자의 성별 및 연령대 유추 가능성에 대한 의문점으로 시작하여, 페이스북 사용자들을 중심으로 대상자를 설정했다. 페이스북 사용자 1,000명에게 DM을 발송을하여 '좋아요' 데이터 수집 동의를 한 대상자 814명의 데이터를 수집했다.

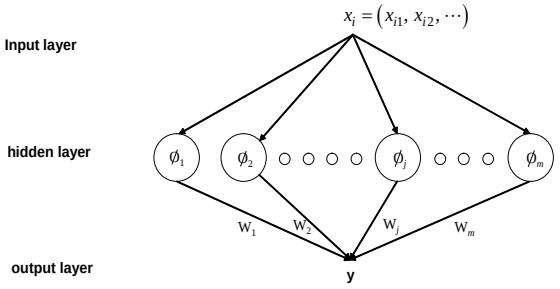
이렇게 모인 대상자들은 연구 목적 정보에 대해 제공받았고 데이터를 수집하는데 있어 사용자들이 데이터를 제공하고 사용자 본인이 느낄 수 있는 개인정보 노출에 대한 민감성 정도를 알려줬다. 이 과정을 통하여 연구 참여에 동의한 사용자들 한에서 페이스북 관련 총 데이터를 제공받았다.

3.2 측정 도구

3.2.1 Gaussian RBF(Radial Basis Kernel Function)

다방면의 분석방법에 적용되어 사용되고 있는 신경망 모델로써, 한 개의 은닉층으로 이루어진 형태로 선형성과 비선형성 둘다 표현 및 분리할 수 있다. 더욱이 가우시안 RBF 모델은 보편 근사(universal approximation)와 최적근사(best approximation)와 같은 수학적 특징을 가지고 있어 다차원 데이터를 차원 축소하는데에 있어 효율적이다[3].

일반적인 RBF 모델은 <그림 1>과 같이 입력층과 비선형 처리를 하는 은닉층 그리고 가중치를 갖는 출력층으로 구성되어 있다[10]. 입력층은 n개의 d차원의 입력 데이터 $\{x_i \in R_d i = 1, \cdots, n\}$를 각각에 대하여 은닉층에 있는 m개의 커널을 기반으로 데이터를 변환하여 커널 출력값 $\phi_1(x_i)$, $\phi_2(x_i)$, $\cdots$, $\phi_k(x_i)$을 생성한다. 커널의 출력값은 가중치 w_j를 곱한 후 더해져 최종 모델 출력값 $y = \sum_{j=1}^{k} \omega_j \phi_j(x_i)$을 생성한다.



〈그림 1〉 RBF 모델 구조

RBF 모델의 최종 출력값 y는 y = f(x) = y = $\sum_{j=1}^{k} \omega_j \exp(-\frac{\|x-\mu_i\|^2}{2\sigma^2})$와 같다. 여기서 μ_i는 가우시안 커널의 중심, σ_i는 폭(width)을 나타낸다. 즉 RBF 모델링은 주어진 데이터 x, y에 대하여 m, μ, σ, ω의 적절한 값을 학습하는 문제이다[3].

3.2.2 nFactors

PCA 혹은 EFA를 사용할때 Factor의 값을 찾는다는 건 차원축소를 하는데 있어 축소, 오류방지, 및 확률 개선이라는 장점을 가진다[20]. Raiche[17]에 의해 만들어진 R 패키지 nFactors는 PA (Parallel Analysis)를 통하여 차원 축소하는데에 있어 적절 n 값을 찾아주어 효과적인 분할하는데 도움을 준다[17].

3.2.3 Random Forest

R version 4.6-12에 나온 논리적인 분류를 하는데에 있어 효과적인 패키지로 한 개의 의사결정 나무를 다

수의 형태로 확장하여 예측하는 방식이다. 기본적으로 일어날 수 있는 상황을 대비하기에 과적합 현상이 일어나지 않는 고 정밀도 알고리즘이다[11].

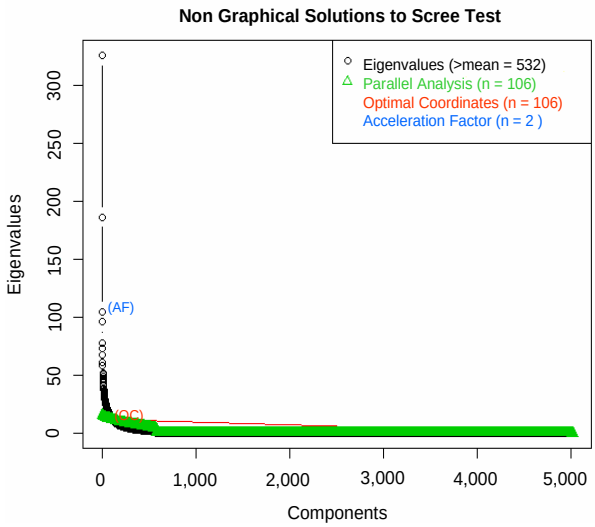
3.3 연구 절차 및 분석 방법

수집한 모든 데이터들은 총 두 차례의 축소 단계를 거쳐 여러 단계를 통하여 처리되는 방식으로 진행됐다.

처음 사용자와 사용자들의 '좋아요' 데이터에 대한 관계를 각 사용자마다 그 페이지를 '좋아요' 했으면 1, 안했으면 0으로 나타내는 희소행렬 형태로 만들어 나열했다. 이때 사용자들 중에 '좋아요' 총 개수가 30 미만이거나 <그림 2>와 같이 사용자 '좋아요' 총 수가 5 미만인 페이지들은 모두 제외했다.

페이지들을 또 한 번 차원축소 방법으로 Gaussian RBF를 선택했고 그러기 위해서는 nFactors 패키지를 사용하여 <그림 3>과 같이 n = 106이라는 값을 구하여 차원 축소했다.

마지막으로, 차원 축소된 값을 randomForest를 이용하여 성별과 연령대 분류를 시도했다. 시도한 결과에 대한 정확도를 평가하기 위해 5-fold cross-validation를 실시했다.

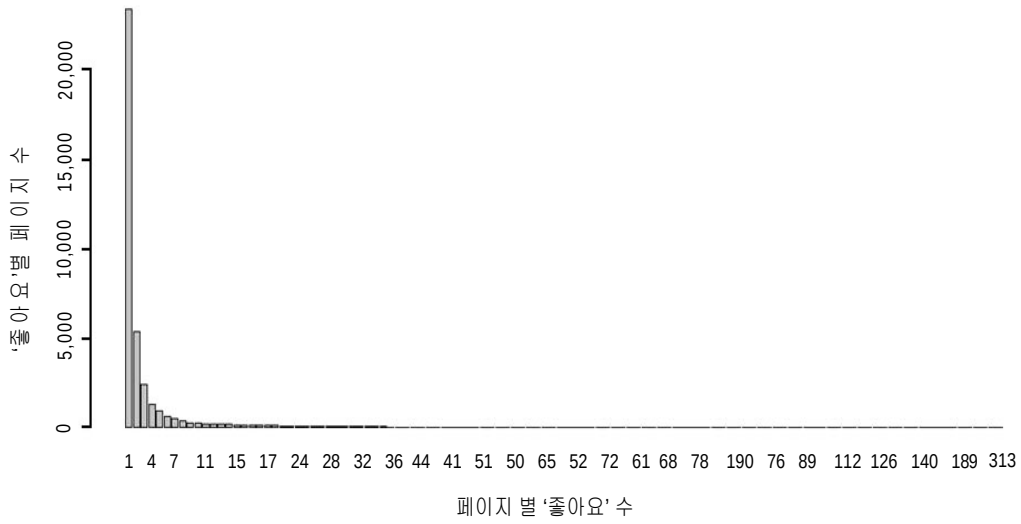


〈그림 3〉 Non Graphical Solutions to Scree Test

Ⅳ. 분석결과 및 연구 결과 토의

4.1 분석결과

본 연구에 참여한 총 814명의 사용자로부터 페이스북 데이터를 받았고 그 가운데 라이크 30개 미만인 샘플은 분석하는 데에 있어 부적합하다 판단하여 제외하고 분석을 실시했다. 총 580개의 최종 샘플은 성별 분석에 사용하였고 이 중 연령 재확인이 가능한 샘플은 466명은 연령 분석에 사용됐다. 실험 표본의 성별 구성은 남성 250명(43%), 여성 330명(57%)이다. 연령대 분포는 10대 232명(50%), 20대 208명(45%), 30대 이상은 26명(5%)이다. 총 페이지 수는 37,443개에서 '좋아요' 수가 5 미만인 페이지를 제외한 5013개 페이지를 변수로 사용했다. Gaussian RBF를 이용하여 20개 변수로 차원 축소를 실행했다. 집단 분류기로 효과적인 randomForest 알고리즘을 통하여 분류를 실시한 결과 성별은 75% 연령대는 97.85%의 정확도를 보였다.



〈그림 2〉 사용자 '좋아요' 수별 페이지 수

4.2 연구 결과 토의

본 연구의 결과는 페이스북 사용자들의 어떠한 개인정보의 수집 없이, SNS에서 '좋아요' 이용 현황을 바탕으로 성별과 연령을 추출해낼 수 있다는 점에 대해서 의의를 둘 수 있다.

측정도구인 Gaussaian RBF, nFactors, randomForest, 그리고 5-fold cross-validation 등을 통해서 성별은 75% 연령대는 97.85%라는 비교적 정확한 사용자들의 개인정보를 얻을 수 있었으며, 이러한 결과 값은 추후 인구통계학적 데이터를 필요로 하는 기업의 빅데이터 관리 및 마케터들에게 유용하게 쓰일 것이라는 결론이다.

Ⅴ. 연구의 한계점 및 향후 연구 방향

첫 번째, 본 연구에서 연령이 97.85%라는 높은 결과로 유의미한 연구를 내었지만, 주로 학생들을 대상으로 하여 10대와 20대에 95% 이상 편중이 되었다. 따라서 앞으로의 연구에서 10대 20대뿐만 아니라 다양한 연령층의 데이터를 확보하여 넓은 폭으로 연구를 진행한다면 폭 넓은 연구 결과물을 얻을 수 있을 것이라고 예상한다.

두 번째, 본 연구에서는 페이스북 사용자에 대한 '좋아요' 데이터 샘플이 선행연구보다 적게 수집함에 따라 성별, 연령은 추론하였지만, 이외에 여러가지 성향 분석의 한계점을 느꼈다. 따라서 향후 '좋아요' 데이터의 샘플을 확보하여 연구를 진행한다면 성별, 연령 이외의 알고자 하는 사용자의 정보를 추출해 낼 것이라고 예상한다.

이러한 한계점에도 불구하고, 본 연구는 다음과 같은 학술적, 실무적 시사점을 갖는다는 데 연구의 기여점이 있다. 학술적인 측면에서는 기존 연구 방법이 문화적 차이 및 데이터 샘플에 대한 제약이 있음에도 불구하고 randomForest를 통한 높은 정확도를 낼 수 있다는 점을 알아낼 수 있었다. 실무적인 측면으로는 사용자들이 개인 정보공개수준을 통제함으로써 기업의 입장에서 사용자 정보에 접근하는 방법이 어려워진 상황이지만 머신러닝을 통한 SNS 사용자 유추가 가능하다는 것을 증명하였다. 이로 인해 기업 가치 창출 및 마케팅하는 데에 자동화된 소비자 분석을 활용함으로써 사용자에게 폭 넓은 제품 및 서비스를 제공할 수 있게 됐다. 또한, 향후에는 단순히 사용자에 대한 분석만이 아닌 페이스북 페이지 및 블로그 자체의 성향 및 가치가 유추 가능할 것으로 기대한다. 본 연구가 향후 페이스북 마케팅 연구 분야의 발전에 소정의 기여를 할 것으로 기대한다.

참 고 문 헌

[1] 배성일, "외식 소비자의 개인정보를 이용한 마케팅 활용이 관계품질과 성과에 관한 연구", 경기대학교 박사학위논문, 2011.

[2] 손달호, "SNS의 사회인지요인이 사용의도에 미치는 영향", 정보시스템 연구, 제23권, 제3호, pp.73-97, 2014.

[3] 신미영, 박준구, "Monk's Problem에 한 가우시안 RBF 모델의 성능 고찰", 전자공학회논문지, 제43권, 제6호, pp.34-42, 2006.

[4] 유종락, "디지털시대의 개인정보보호: 새로운 개인정보보호법을 중심으로", 디지털융복합연구, 제9권, 제6호, pp.81-90, 2011.

[5] 이상민, "페이스북 이용자의 제품 정보메시지 구전의도 연구: 메시지 공급자 유형과 계획된 행동이론(TPB)을 중심으로", 중앙대학교 석사학위논문, 2015.

[6] 전성민 역, "페이스북 시대: 소셜 네트워크를 활용한 비즈니스와 마케팅", 서울: 한빛미디어: Shih, Clara, The Facebook Era, Tapping Online Social Networks to Market, Sell, and Innovate(2nd Ed.), Pearson Education, NJ: Prentice-Hall, 2010.

[7] 전유진, "기업의 SNS 마케팅 활용에 관한 연구", 숭실대학교 석사학위논문, 2015.

[8] 최재용, "SNS(소셜네트워크 서비스)를 활용한 유통업체 온라인 마케팅 활성화 방안에 관한 연구", 한국유통학회 학술대회발표논문집, pp.183-201, 2010.

[9] 최재혁, "형사법상 개인정보보호에 관한 연구-사이버공간 상의 개인정보보호를 중심으로", 한양대학교 박사학위논문, 2008.

[10] 허성현, 장석준, 박승권, "비선형 채널상에서 RBF 신경망을 사용한 채널 등화기의 성능분석", 한국통신학회 학술대회논문집, pp.1709-1713, 1998.

[11] Breiman, L. and A. Culter, "randomForest: Breiman and Culter's Random Forests for Classification and Regression. R package version 4.6-12, 2015, URL:

https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

[12] Facebook Reports Third Quarter 2015 Results, 2015, URL: http://investor.fb.com/search.cfm?keyword=Facebook+ Reports+2015.

[13] Kaplan, A.M. and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media", *Business Horizons*, Vol.53, No.1, pp.59-68, 2010.

[14] KB 금융지주경영연구소. 금융권의 소셜 마케팅 활용, 2013.

[15] Kosinski, M., D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior", *Proceedings of the National Academy of Sciences*, PNAS, pp.5802-5805, 2013.

[16] Madden, M. and A. Smith, *Reputation Management and Social Media*, Washington, DC: Pew Internet and American Life Project, 2010 URL : http://pewinternet.org/~/media/Files/Reports/2010/PIP_Reputation_Management.pdf.

[17] Raiche, G. and D. Magis, "Package 'nFactors' Parallel Analysis and Non Graphical Solutions to the Cattell Scree Test, 2010, URL : http://cran.r-project.org/Web/packages/n Factors/n Factor.

[18] Raiche, G., M. Riopel, and J.-G. Blais, "Non graphical solutions for the Cattell's scree test. Paper presented at the International Annual meeting of the Psychometric Society, Montreal, 2006, URL http://www.er.uqam.ca/nobel/r17165/RECHERCHE/COMMUNICATIONS/.

[19] Wijaya Muhammad, E., "페이스북 사용자의 '좋아요', 패턴을 통한 정치적 성향 예측", 서울과학기술대학교 석사학위논문, 2015.

[20] Zwick, W.R. and W.F. Velicer, "Comparison of five rules for determining the number of components to retain", *Psychological Bulletin*, Vol.99, pp.432-442, 1986.

저 자 소 개

유 성 종(Seong Jong Yu)

- 2013년 : Rensselaer Polytechnic Institute EMAC학과 (이학사)
- 2015년~현재 : 연세대학교 정보대학원 (석사과정)
- 관심분야 : Big data, Open collaboration, Machine learning

안 세 은(Seun Ahn)

- 2015년 : 숭실대학교 스토리텔링 경영학과 (경영학 학사)
- 2015년~현재 : 연세대학교 정보대학원 (석사과정)
- 관심분야 : Big data, Open collaboration, healthcare, IoT

이 준 기(Zoonky Lee)

- 1985년 : 서울대학교 전산통계학과 (학사)
- 1991년 : 카네기멜론대학 사회심리학과 (석사)
- 1999년 : 남가주대학교 경영정보학과 (박사)
- 2004년~현재 : 연세대학교 정보대학원 (교수)
- 관심분야 : Web2.0, E-Transformation, Dynamic pricing, KM, Open Innovation