

온라인 소셜 네트워크에서 사용자 간의 관계 강도 변화 추정 기법

(Estimating the Variation of Relationship Strength among
the Users in Online Social Networks)

박 지 범 [†] 임 혜 원 [†] 김 형 주 ^{**}
(Jee-bum Park) (Hyewon Lim) (Hyoung-Joo Kim)

요약 온라인 소셜 네트워크는 이진 구조의 그 특성 상 실세계의 소셜 네트워크를 정확하게 반영하지 못하기 때문에, 최근의 온라인 소셜 네트워크 연구에서는 사용자 프로필이나 상호작용 등의 정보를 바탕으로 사용자 간의 관계 강도를 측정하여 실세계의 소셜 네트워크를 표현하는 방법이 주를 이룬다. 하지만, 이러한 기존의 방법들은 현재 시점의 관계 강도만 고려하기 때문에 올바르게 않은 강한 친구 관계 그래프를 만들어 내는 경우가 생긴다. 본 논문에서는 이러한 문제점을 해결하기 위해 유도 상호작용 그래프의 개념과 사용자 간의 관계 강도의 변화 추정 기법을 제안한다. 또한, 제안하는 기법으로 생성한 강한 친구 관계 그래프가 기존의 단순한 상호작용 그래프만을 사용하는 방법에 비해 더 효과적임을 보인다.

키워드 : 소셜 네트워크, 데이터마이닝, 관계 강도

Abstract Online social networks usually consist of the binary relationships; thus, they do not reflect the real-world social network correctly. For this reason, most of the recent studies focus on estimating the relationship strength among users based on the information such as user profiles, user interactions, and user relationships. However, since the previous techniques consider the relationship strength at that point in time, it may generate the incorrect strong friendship graph. In this paper, we propose the concept of a derived interaction graph and estimate the variation of the relationship strength among the users to solve the problems. Through the evaluations, we show that the strong friendship graph generated by our method is more effective than that by the previous work which uses only a simple interaction graph.

Key words : Social Network, Data Mining, Relationship Strength

1. 서론

최근 스마트폰과 같은 모바일 기기가 널리 보급되면서

사용자들은 언제 어디서나 웹 서비스를 자유롭게 이용할 수 있게 되었고, 이러한 변화는 시간적 혹은 지리적 제약에 관계없이 다른 사용자와 자유롭게 의사소통을 하거나 빠른 정보 공유가 가능한 시대를 만들어 주었다[1]. 이러한 시대 변화와 더불어, 사용자 간에 관계를 맺고 대화를 나누거나 비슷한 관심사를 가진 사람들끼리 모여 정보를 공유할 수 있게 하는 웹 서비스의 필요성이 부각됨에 따라, 자유로운 의사소통과 정보 공유, 인맥 확대 등을 통해 사회적 관계를 생성하고 강화시켜주는 다양한 소셜 네트워크 서비스(SNS; Social Network Services)가 등장하였다. Facebook[2]이나 Twitter[3]와 같은 대표적인 소셜 네트워크 서비스의 실제 사용자 수가 9억 명을 넘어 서면서[4], 소셜 네트워크 서비스는 상당히 많은 사용자 정보와 친구 관계, 의사소통 및 상호작용 내역 등을 보유하게 되었다. 이것으로부터 얻어낼 수 있는 다양한 정보

· 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 20120005695)

[†] 비 회 원 : 서울대학교 컴퓨터공학부
jbpark@idb.snu.ac.kr
(Corresponding author)
hwlim@idb.snu.ac.kr

^{**} 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2012년 8월 1일

심사완료 : 2012년 11월 2일

Copyright©2013 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 : 데이터베이스 제40권 제1호(2013.2)

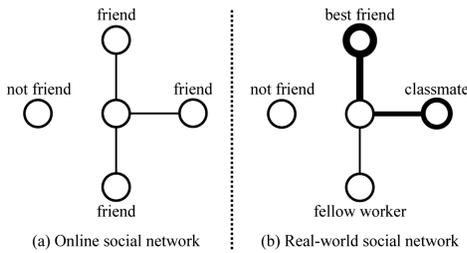


그림 1 온라인과 실제세계의 소셜 네트워크

들은 기술 개발이나 마케팅 등을 위한 비즈니스 활용 가치가 매우 높다고 알려져 있다[5,6].

소셜 네트워크 서비스에서는 사용자 간의 사회적 연결(social connection)을 관계(relationship)라는 개념으로 표현한다[7]. 실제로 서로 친분이 있거나, 비슷한 관심사를 가진 경우에 두 사용자는 소셜 네트워크 서비스 안에서 관계를 형성하게 되는데, 이러한 관계는 소셜 네트워크 서비스마다 “친구(friend)”나 “팔로우(follow)” 등 각기 다른 명칭으로 표현된다.

대부분의 온라인 소셜 네트워크 서비스에서, 두 사용자 간의 관계는 그림 1(a)와 같이 “친구” 혹은 “친구 아님”과 같은 이진(binary) 구조로, 오직 두 가지 종류의 상태를 가질 수 있다. 하지만, 실제세계의 소셜 네트워크에서는 그림 1(b)와 같이 두 사람 사이에 관계의 강도(relationship strength)가 존재한다[8]. 예를 들면, 두 사람 사이가 절친한 친구인 강한 연결(strong ties)로 이루어져 있을 때도 있고, 단순히 서로 아는 사이인 약한 연결(weak ties)일 경우도 있는데[9], 이는 대부분의 온라인 소셜 네트워크 서비스에서 제공하는 단순한 이진 구조 네트워크와는 큰 차이가 있다.

소셜 네트워크 서비스 안에서, 서로 관계가 형성된 사용자 사이에서는 메시지 전송이나 답장(reply), 댓글(comment) 달기 등의 능동적인 사회적 상호작용(active social interaction)[10]이 발생하게 되는데, [11]에서는 이것을 사용자 상호작용(user interaction)이라고 정의하고 이것이 최근 일정 기간 p 동안 발생한 빈도를 바탕으로 관계 강도로 측정하는 방법을 제안하였다. 그러나 이 방법은 p 의 정확한 값을 결정하기가 쉽지 않을 뿐만 아니라, 그 값에 따라서 결과가 크게 달라질 수 있기 때문에 실제로 활용하기는 힘들다는 단점이 있다. 또한, 최근에 사용자 상호작용이 많이 발생했다는 근거만으로도 그 두 사용자의 관계 강도가 높다고 추정하는 것은 문제가 있다. 가령, 별로 친한 사이가 아닌 두 사용자가 우연히 최근에 연락을 해야 할 일이 생겨서 갑작스럽게 사용자 상호작용이 많이 발생한 경우, 이 방법으로는 올바른 관계 강도를 측정할 수가 없다. 무엇보다도 단순히 사용자 간의 관계 강도를 측정하는 기존의 연구[9,11,12]

에서 제안한 방법들은, 그 사용자 간의 관계 강도가 앞으로 어떻게 변화할 것인지에 대한 근거를 전혀 제시해 주지 못하기 때문에 친구 관계의 추세(trend) 분석이나 친구 관계의 변화 예측과 같은 작업을 할 수 없다.

이러한 문제를 해결하기 위해서는 현 시점의 사용자 간 관계 강도뿐만 아니라, 특정 시점의 과거에서 현 시점에 이르기까지의 관계 강도의 변화(variation)를 함께 추정해야 할 필요가 있다. 본 논문에서는 온라인 소셜 네트워크로부터 수집할 수 있는 정보를 바탕으로 사용자 간의 관계 강도와 그것의 변화를 추정하는 방법을 제안하고, 그 변화 값을 기준치로 갖는 유도 상호작용 그래프를 소개한다. 또한, Facebook의 데이터를 수집하여 방법을 적용하고 실험을 수행한다. 그리고 나서 기존 연구와의 비교를 통해 우리가 제안한 방법의 우수성을 보인다.

본 논문의 2장에서는 온라인 소셜 네트워크에서 사용자 간의 관계 강도를 측정하는 방법에 대한 기존의 관련 연구를 소개한다. 3장에서는 제안하는 관계 강도 변화 추정 기법을 구체적으로 소개하며, 4장에서는 실험을 통해 제안된 기법이 얼마나 효과적인지를 설명하고, 5장에서 결론을 맺는다.

2. 관련 연구

소셜 네트워크 서비스에서 수집할 수 있는 정보를 가지고, 사용자 간의 관계 강도를 측정하는 방법에 대한 다양한 연구가 진행되었다. [12]에서는 Facebook에서 수집할 수 있는 사용자 상호작용 데이터의 종류를 담벼락(wall)과 사진, 그룹(groups)으로 구분하여 이를 토대로 로지스틱 회귀분석(logistic regression)과 의사결정 트리(decision trees), 단순 베이지안 분류자(naive Bayesian classifiers)를 이용한 관계 강도를 측정하는 방법을 제안하였다. 그리고 세 가지 각각의 방법에 대한 순위 결과를 대상으로 AUC(area under the ROC curve)를 통해 성능을 분석하였다. AUC는 ROC(Receiver Operating Characteristic)[13] 곡선의 아래 면적으로, 보통 순위 결과를 만들어 내는 학습 모형의 성능을 평가하기 위한 척도로 사용된다[14].

[9]에서는 서로 비슷한 사람들끼리 주로 친구가 형성되고 정보 공유 및 의사소통이 일어난다는 [15]의 연구를 바탕으로, 주로 사용자 프로필이나 관심사가 서로 유사할수록 관계 강도가 높고, 관계 강도가 높은 두 사용자 사이에 사용자 상호작용이 발생한다는 점에 착안하여, 사용자 프로필과 상호작용의 정보를 기준으로 관계 강도를 측정할 수 있는 잠재적 변수 모형(Latent Variable Model)을 제안하였다. 그리고 이를 LinkedIn[16]과 Facebook의 데이터를 가지고 학습시킨 후, AUC를 통해 성능을 평가하였다.

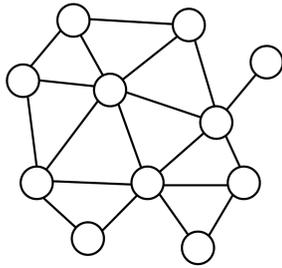


그림 2 친구 관계 그래프의 예

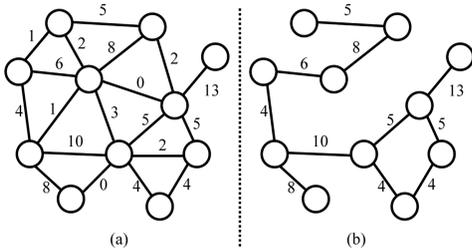


그림 3 상호작용 그래프와 강한 연결($T=3$)만으로 이루어진 강한 친구 관계 그래프

[11]의 연구에서는 Facebook에서 발생하는 메시지 전송이나 답장, 댓글 달기 등의 모든 능동적인 행위를 사용자 상호작용이라고 정의하고, 이 상호작용 데이터가 최근 일정 기간 동안 발생한 빈도에 따라 관계 강도로 측정하여, 이를 기반으로 만들 수 있는 상호작용 그래프(interaction graph)를 제안하였다.

그림 2와 같은 소셜 네트워크 서비스에서 쉽게 수집할 수 있는 친구 관계 그래프(friendship graph)에 간선(edge)의 가중치(weight)를 관계 강도로 나타내면 상호작용 그래프가 된다(그림 3(a)). 여기에서 가중치가 T 보다 큰 간선을 강한 연결이라고 하고, T 보다 크지 않은 가중치를 갖는 간선을 모두 제거하면 그림 3(b)처럼 강한 연결만으로 이루어진 강한 친구 관계 그래프(strong friendship graph)를 만들 수 있다. [11]에서는 강한 친구 관계 그래프가 스팸 필터링(spam filtering) 기법이나 Sybil attack 방지 기법에 대해서 실제로 활용될 수 있는 방안을 제시하고 그 효과를 분석하였다. 또한, 이 강한 친구 관계 그래프가 소셜 네트워크 서비스의 친구 관계 그래프에 비해 작은 세상 효과(small-world effect) [17]가 얼마나 더 나타나는지 실험하여 비교하였다. 작은 세상 효과는 크고 복잡한 네트워크를 이루는 개별적인 요소가 단지 몇 단계만 거치면 모두 서로 연결되는 현상을 일컫는 말로, 이러한 특징을 갖는 네트워크는 그 안에서의 정보 확산 속도가 상당히 빠르며, 실제계의 소셜 네트워크가 이러한 현상을 나타낸다는 것이 증명되

었다[17]. 어떤 네트워크에서 작은 세상 효과가 어느 정도 나타나는지는 특성 경로 길이(characteristic path length)와 군집 계수(clustering coefficient)를 측정하여 알아볼 수 있다[18]. 본 논문에서도 제안한 방법이 기존 연구에 비해 얼마나 더 실제계의 소셜 네트워크에 근접하였는지 이 두 가지 값을 측정하여 실험하고 분석하였다.

3. 유도 상호작용 그래프(Derived Interaction Graph)

이 장에서는 본 논문에서 제안하는 유도 상호작용 그래프를 소개하고, 유도 상호작용 그래프의 가중치를 나타내는 관계 강도 변화 값을 추정하는 기법에 대해 설명한다.

[11]의 연구에서는 상호작용 그래프를 정의하고 이것으로부터 강한 연결만으로 이루어진 강한 친구 관계 그래프를 만들 수 있다는 것을 소개하였다. 하지만, 본 논문의 1장에서 언급한 바와 같이, 상호작용 그래프만을 가지고 실제계의 소셜 네트워크에 근접한 강한 친구 관계 그래프를 만들기에는 몇 가지 문제가 있다. 이러한 문제를 해결하기 위해, 유도 상호작용 그래프를 제안한다.

3.1 정의

유도 상호작용 그래프는 그림 2와 같은 친구 관계 그래프에 그림 4처럼 간선의 가중치를 관계 강도의 변화 값으로 나타낸 그래프이다. 그림 3(a)에서 볼 수 있는 [11]의 상호작용 그래프와 유사하지만, 간선의 가중치가 관계 강도의 변화 값이라는 차이점이 있다.

유도 상호작용 그래프는 순서쌍 $J_{t,p} = (U, E_{t,p})$ 로 정의한다. U 는 그래프의 정점(vertex)을 이루는 사용자의 전체 집합을 의미하며, $E_{t,p}$ 는 그래프의 정점 u 와 v 를 연결하는 간선과 그것의 가중치를 나타내는 순서쌍 $((u, v), \psi_{u,v,p}(t))$ 의 전체 집합을 뜻한다. 정점 u 와 v 를 연결하는 간선의 가중치는 함수 $\psi_{u,v,p}(t)$ 로 표현되며, 이는 $W_{u,v,p}(t)$ 의 미분함수로 다음과 같이 정의한다.

$$\psi_{u,v,p}(t) = \frac{d}{dt} W_{u,v,p}(t) \tag{1}$$

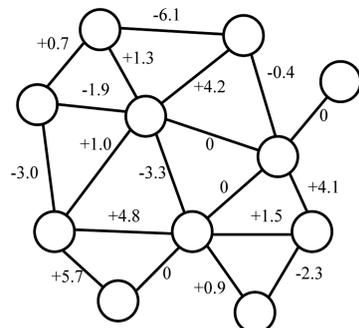


그림 4 유도 상호작용 그래프

함수 $W_{u,v,p}(t)$ 는 특정 시점 t 에서 윈도우 크기(window size) p 로 측정된 사용자 u 와 v 사이의 관계 강도를 나타낸다. 이 관계 강도를 측정하는 방법은 여러 가지가 있으며[9,11,12], 본 논문에서는 [11]에서 한 것과 같이 사용자 u 와 v 사이에 상호작용이 발생한 빈도로 정의한다.

유도 상호작용 그래프는 방향성을 갖는 그래프(directed graph)로 표현할 수도 있으나, 본 논문에서는 방향성이 없는 그래프(undirected graph)라고 가정한다. 따라서, 다음의 식은 항상 성립한다.

$$\psi_{u,v,p}(t) = \psi_{v,u,p}(t)$$

상호작용 그래프에서는 두 사용자 사이의 관계 강도를 알 수 있는 반면에, 유도 상호작용 그래프에서는 두 사용자 사이의 관계 강도의 변화 추세를 알 수 있다. 두 사용자 사이의 관계 강도가 증가하는 추세일 경우 가중치는 양의 값을 갖게 되며, 감소하는 추세일 경우에는 음의 값을 나타낸다. 값이 0일 경우 변화가 없다는 것을 의미한다.

3.2 강한 친구 관계 그래프(Strong Friendship Graph)

강한 친구 관계 그래프를 만들 때, 유도 상호작용 그래프를 사용하면 강한 연결의 기준을 좀 더 구체화할 수 있다. 기존의 방법은 어떤 간선이 강한 연결인지 판별하기 위해 상호작용 그래프의 가중치만을 고려하였으나, 이와 더불어 유도 상호작용 그래프의 가중치도 함께 고려하게 되면, 그림 5에서 볼 수 있듯이 (i) 간선의 관계 강도가 T 보다 큰 경우, (ii) 간선의 관계 강도의 변화 추세가 V 보다 큰 경우, (iii) 간선의 관계 강도가 T

보다 크거나 관계 강도의 변화 추세가 V 보다 큰 경우, (iv) 간선의 관계 강도가 T 보다 크고 관계 강도의 변화 추세가 V 보다 큰 경우, 이렇게 총 네 가지 경우에 대한 강한 연결의 기준이 생긴다.

따라서, 그림 2와 같은 친구 관계 그래프에서 T 와 V 가 주어졌을 때, 그림 3(a)와 같은 상호작용 그래프와 그림 4와 같은 유도 상호작용 그래프를 통해 그림 5와 같은 네 가지 경우의 서로 다른 강한 친구 관계 그래프를 만들 수 있으며, 각 경우의 그래프는 다음의 조건을 각각 만족한다.

$$(i) : \forall u \forall v (W_{u,v,p}(t) > T)$$

$$(ii) : \forall u \forall v (\psi_{u,v,p}(t) > V)$$

$$(iii) : \forall u \forall v (W_{u,v,p}(t) > T \vee \psi_{u,v,p}(t) > V)$$

$$(iv) : \forall u \forall v (W_{u,v,p}(t) > T \wedge \psi_{u,v,p}(t) > V)$$

이렇게 만들어진 강한 친구 관계 그래프 가운데, (i)의 경우는 [11]의 연구에서 소개한 것과 같다. 우리는 (i)을 제외한 나머지 세 경우에 대해 기존의 방법과 어떤 차이가 있는지 분석하였으며, 그 결과는 4장에서 설명한다.

3.3 관계 강도 변화 추정 기법

강한 친구 관계 그래프를 만드는 데 사용되는 유도 상호작용 그래프 $J_{t,p}$ 를 구하기 위해서는 그래프를 이루는 모든 사용자 (u, v)에 대해 관계 강도의 변화 값을 의미하는 함수 $\psi_{u,v,p}(t)$ 를 계산해야 한다. 이를 위해 식 (1)을 다음과 같이 나타낼 수 있다.

$$\frac{d}{dt} W_{u,v,p}(t) = \lim_{\Delta t \rightarrow 0} \frac{W_{u,v,p}(t) - W_{u,v,p}(t - \Delta t)}{\Delta t}$$

하지만 온라인 소셜 네트워크 서비스로부터 수집할 수 있는 정보만으로는 실제로 함수 $W_{u,v,p}(t)$ 가 정확히 어떤 함수인지를 알 수가 없고, 대신 그것의 표본 데이터만 알 수 있다. 표본 데이터는 이산형 분포(discrete distribution)를 따르고 그 개수가 한정되어 있기 때문에 위의 수식에서 극한값을 계산하는 것은 불가능하다. 이 문제를 해결하기 위한 대안으로, 회귀분석[19]과 같은 방법을 사용할 수 있다. 회귀분석은 한 개 또는 그 이상의 독립변수(independent variables)와 종속변수(dependent variable) 사이의 인과관계를 규명하고자 하는 분석 방법으로, 시간에 따라 변화하는 데이터를 예측하거나 표본 데이터를 가지고 변화 추세를 추정하고자 할 때 적합하다. 단, 회귀분석을 사용하기 위해서는 사용자 u 와 v 사이에 상호작용이 발생한 빈도를 나타내는 함수 $W_{u,v,p}(t)$ 가 선형(linear)의 추세를 보인다고 가정해야 할 필요가 있다. 만약 함수 $W_{u,v,p}(t)$ 가 실제로 선형 함수일 경우에는 이 방법으로 거의 정확한 함수 $W_{u,v,p}$

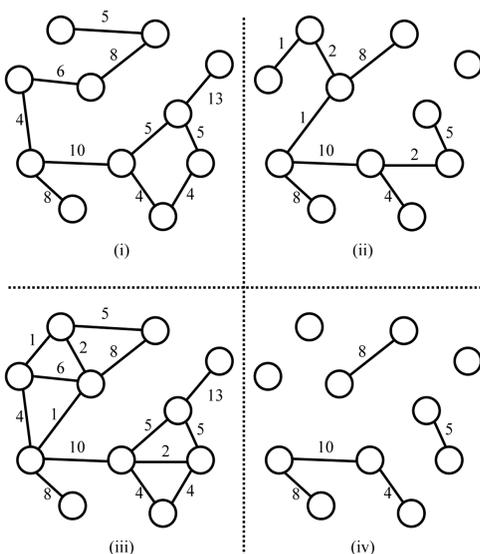


그림 5 강한 친구 관계 그래프($T=3, V=0$)

(t)를 계산해낼 수 있지만, 그렇지 않을 경우에는 표본 데이터의 전역적인 상황 혹은 하향 추세만을 예측할 수 있다.

우선, 함수 $W_{u,v,p}(t)$ 가 기울기 ψ 인 선형의 추세를 보인다고 가정하면, 다음과 같은 일차 식으로 나타낼 수 있다.

$$W_{u,v,p}(t) = \psi t + k + \varepsilon(t)$$

위의 식에서, $\hat{\varepsilon}(t)$ 은 실제 표본 데이터와의 오차를 의미한다. 이 오차를 최소화하는 ψ 와 k 를 구하기 위해, 최소제곱추정법(least squares estimation)[20]을 사용한다. 함수 $W_{u,v,p}(t)$ 의 표본 데이터가 주어졌을 때, 그 표본 데이터 함수의 정의역(domain) 집합을 X 라고 정의하면, 오차의 제곱의 총 합계를 다음과 같이 ψ 와 k 에 대한 함수로 나타낼 수 있다.

$$S(\psi, k) = \sum_{t \in X} \hat{\varepsilon}^2(t) = \sum_{t \in X} (W_{u,v,p}(t) - \psi t - k)^2$$

이 함수를 ψ 와 k 에 대해 각각 편미분하여 그 값이 0이 되도록 식을 세우면 다음과 같다.

$$\frac{\partial S(\psi, k)}{\partial \psi} = - \sum_{t \in X} 2t(W_{u,v,p}(t) - \psi t - k) = 0$$

$$\frac{\partial S(\psi, k)}{\partial k} = - \sum_{t \in X} 2(W_{u,v,p}(t) - \psi t - k) = 0$$

위의 두 식을 전개하여 정리하면,

$$\sum_{t \in X} tW_{u,v,p}(t) - \psi \sum_{t \in X} t^2 - k \sum_{t \in X} t = 0 \quad (2)$$

$$\sum_{t \in X} W_{u,v,p}(t) - \psi \sum_{t \in X} t - k|X| = 0 \quad (3)$$

k 를 소거하기 위해 식 (2)에 $|X|$ 를 곱하면,

$$|X| \sum_{t \in X} tW_{u,v,p}(t) - \psi |X| \sum_{t \in X} t^2 - k|X| \sum_{t \in X} t = 0 \quad (4)$$

식 (4)에 식 (3)을 대입하여 풀면,

$$\begin{aligned} |X| \sum_{t \in X} tW_{u,v,p}(t) - \psi |X| \sum_{t \in X} t^2 - \left(\sum_{t \in X} W_{u,v,p}(t) - \psi \sum_{t \in X} t \right) \sum_{t \in X} t \\ = 0 \\ |X| \sum_{t \in X} tW_{u,v,p}(t) - \sum_{t \in X} t \sum_{t \in X} W_{u,v,p}(t) \\ = \psi |X| \sum_{t \in X} t^2 - \psi \left(\sum_{t \in X} t \right)^2 \end{aligned}$$

ψ 에 대해서 정리하면,

$$\psi = \frac{|X| \sum_{t \in X} tW_{u,v,p}(t) - \sum_{t \in X} t \sum_{t \in X} W_{u,v,p}(t)}{|X| \sum_{t \in X} t^2 - \left(\sum_{t \in X} t \right)^2} \quad (5)$$

ψ 는 함수 $W_{u,v,p}(t)$ 가 선형의 추세를 보인다고 가정했을 때, 그것의 최적 기울기를 의미한다. 따라서, 이 값을 사용자 u 와 v 에 대한 관계 강도의 변화 값으로 사용할 수 있다.

지금까지 설명한 내용을 바탕으로, 유도 상호작용 그

래프를 만드는 과정을 알고리즘 1에 표현하였다. 이 알고리즘은 그래프의 정점을 이루는 사용자의 전체 집합 U 와 친구 관계의 유무를 나타내는 순서쌍의 집합 F , 관계 강도 함수 $W_{u,v,p}(t)$, 그리고 관계 강도 표본의 정의역 집합 X 가 입력으로 주어지면 유도 상호작용 그래프 J 를 생성한다.

알고리즘 1 유도 상호작용 그래프 생성

Input:

user set U , friendship set F , relationship strength function $W_{u,v,p}(t)$, input samples X

Output:

derived interaction graph J

```

1:  $E \leftarrow \emptyset$ 
2: for each  $u \in U$ :
3:   for each  $v \in U$ :
4:     if  $(u, v) \in F$ :
5:       Compute  $\psi$  according to equation (5)
6:        $E \leftarrow E \cup \{(u, v), \psi\}$ 
7:  $J \leftarrow (U, E)$ 

```

4. 실험 결과 및 분석

이 장에서는 수집한 데이터와 이를 기반으로 수행한 실험에 대해 설명한다. 그리고 결과 분석을 통해 본 논문에서 제안한 기법이 기존의 방법에 비하여 어떠한 차이를 나타내는지 살펴본다.

4.1 데이터 수집

3장에서 소개한 관계 강도 변화 추정 기법을 기존 연구와 비교하기 위해서, Facebook의 데이터를 수집하여 실험하였다. Facebook은 전 세계에서 가장 많은 사용자 수[4]를 확보하고 있는 온라인 소셜 네트워크 서비스이다. Facebook의 사용자는 이름이나 성별, 생일 등과 같은 기본적인 정보와, 학벌, 사용 가능 언어, 거주 지역, 연애 상태, 관심사 등과 같은 부가적인 정보를 자신의 개인 프로필 페이지에 설정할 수 있다. Facebook 안에서 사용자는 다른 사용자와 친구 관계를 형성할 수 있고, 각 사용자 마다 최대 5,000 명까지 친구 관계를 가질 수 있다[11]. 사용자는 각각 글을 쓸 수 있는 “답변”을 가지고 있으며, 사용자 간의 정보 공유나 의사소통과 같은 사용자 상호작용은 대부분 답변에서 발생한다. 서로 친구 관계가 형성된 사용자라면 누구나 상대방의 답변을 볼 수 있는 권한을 가진다. 사용자는 자신의 답변이나 친구의 답변에 글을 쓰거나 사진을 게시할 수 있고, 그 글이나 사진에 댓글을 작성할 수 있다.

Facebook 데이터 수집 과정은 친구 관계 그래프 수집과 상호작용 그래프 수집, 이렇게 두 단계로 나누어서

표 1 Facebook에서 발생하는 사용자 상호작용의 종류

상호작용의 유형	설 명
게시(post)	u 가 v 의 담벼락에 글이나 사진을 게시하였을 때
댓글(comment)	u 가 v 의 게시물에 댓글을 작성하였을 때
태그(tagging)	u 가 자신의 게시물이나 댓글에서 v 를 언급하였을 때
좋아요(like)	u 가 v 의 게시물이나 댓글을 좋아할 때

수행하였다. 먼저, 친구 관계 그래프를 만들기 위해 임의적으로 3명의 활동적인 사용자(active-users)들을 선택하고, 그 사용자들을 시작점으로 하여 최대 2단계까지 친구 관계가 형성되어 있는 사용자들을 모두 수집하여 사용자 목록 표본을 만들었다. 그리고 이 사용자 목록 표본을 기반으로 하여 사용자들 간의 친구 관계를 수집하고, 이것으로부터 전체 사용자 3,104 명과 52,913 개의 친구 관계로 이루어진 친구 관계 그래프를 생성하였다. 그 다음 상호작용 그래프를 만들기 위해 앞서, Facebook에서 발생하는 사용자 상호작용을 대략 32 MB 가량 수집하고 표 1과 같이 분류하였다.

게시는 어떤 사용자 u 가 그 사용자의 친구 v 의 담벼락에 글이나 사진을 올렸을 때, 사용자 상호작용이 발생한 것으로 가정할 것이다. 댓글은 어떤 담벼락에 게시된 친구 v 의 게시물에 u 가 댓글을 작성한 경우를 의미한다. Facebook에서는 게시물이나 댓글을 작성할 때 @을 사용하여 어떤 사용자의 이름을 언급할 수 있는, "태그" 기능을 제공한다. 게시물이나 댓글에서 사용자 이름을 태그하게 되면, 자동적으로 태그된 사용자의 프로필 페이지로 향하는 링크가 생기고, 태그된 사용자에게 그 사실이 통보된다. Facebook은 게시물에 댓글을 작성할 수 있지만, 댓글에 댓글을 작성할 수는 없기 때문에, 이 기능은 사용자 참조를 위한 본래의 목적과 함께, 사용자 간에 댓글로 대화를 주고 받는 기능으로도 사용된다. 따라서, 이 기능을 사용한 것도 사용자 간에 상호작용이 발생했다고 볼 수 있다. 그리고 Facebook에서는 어떤 게시물이나 댓글에 "좋아요(like)" 표시를 할 수 있다. 이것은 일종의 추천 기능으로, 사용자가 친구의 게시물이나 댓글에 좋아요 표시를 하였을 때 두 사용자 간에 상호작용이 발생했다고 고려하였다.

대량의 Facebook의 데이터를 자동으로 수집하기 위해서 크롤러(crawler)를 구현하였다. 이 크롤러는 Python으로 구현되었으며, 멀티스레딩(multithreading) 및 분산 환경에서 동작이 가능하도록 설계되었다. 두 사용자 간의 사용자 상호작용 데이터는 Facebook에서 제공하는 API[21]를 이용하여 수집할 수 있었으나, 어떤 사용자의 친구 관계 데이터는 열람 권한이 부여되지 않아 API가 올바르게 작동하지 않는 문제가 있었다. 이 문제를 해결

하기 위해 Facebook의 친구 목록 페이지에서 발생하는 AJAX 요청을 분석하여, 어떤 사용자의 친구 목록을 분석(parsing)하고 그 친구 관계를 추출해 내는 부분을 추가로 구현하였다. 해당 사용자가 접근 권한을 부여하지 않아 열람할 수 없는 데이터는 무시하고, 열람이 가능한 데이터에 대해서만 수집하였으며, 이렇게 수집한 데이터를 보관하기 위한 관계형 데이터베이스로는 Oracle Database 11g를 사용하였다.

4.2 그래프 생성

수집한 Facebook 데이터를 가지고 최종적으로 강한 친구 관계 그래프를 만들기 위해서는 그림 6과 같은 그래프 생성 과정을 거쳐야 한다. 이 때, 각 단계마다 그래프를 생성하기 위해 필요한 변수 값이 있는데, 변수들의 종류는 그림 6의 각 단계의 괄호 안에 표기되어 있으며 이 변수들은 앞서 3장에서 소개된 바 있다. t 는 어떤 특정 시점의 시각을 나타내는 변수로, 친구 관계는 시간에 따라 변화할 수 있기 때문에 이러한 친구 관계 상황을 어떤 특정한 시점으로 고정해야 한다. p 는 사용자 사이의 관계 강도를 측정하여 상호작용 그래프를 만들고자 할 때, 관계 강도 측정 범위를 의미하는 윈도우 크기 변수이다. 어떤 특정 시점 t 부터 p 를 넘지 않는 기간 동안 발생한 사용자 상호작용들을 관계 강도로 측정하여 상호작용 그래프를 생성한다. X 는 유도 상호작용 그래프를 만들기 위해 필요한 관계 강도의 표본으로, 식 (5)를 계산할 때 사용된다. T 와 V 는 강한 친구 관계 그래프를 만들 때 필요한 강한 연결의 기준 값이다. 실험을 수행하기 전에 먼저 이 변수들의 값을 지정해야 할 필요가 있다. 그러기 위해서 우리는 t 의 값을 현재 시점으로 정하고, 이 t 에서의 친구 관계 그래프를 기준으로 p 의 값에 따라 상호작용 그래프의 관계 강도가 어떤 경향을 나타내는지 살펴보았다.

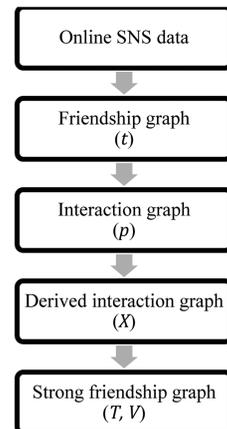


그림 6 그래프 생성 과정

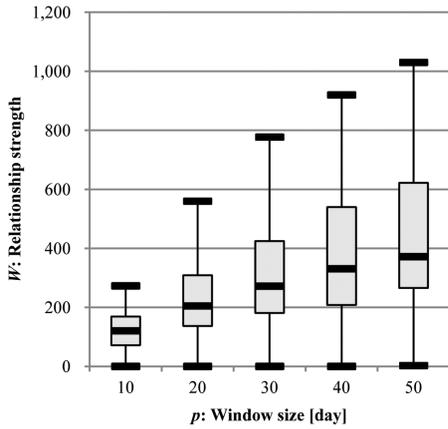


그림 7 p 의 값에 따른 상호작용 그래프의 관계 강도

그림 7은 p 의 값에 따른 상호작용 그래프의 관계 강도를 상자 그림(box plot)으로 나타낸 것이다. p 의 값이 50일 때 두 사용자 사이의 최대 관계 강도는 1,000이 넘지만, 최소의 경우를 살펴보면 0에 가깝다. 이것은 소셜 네트워크 서비스 상에서 어떤 두 사용자가 서로 친구 관계임에도 불구하고, 50일 동안 상호작용이 거의 발생하지 않았다는 것을 의미한다. 이러한 피상적인 관계가 얼마나 존재하는지 알아보기 위해, T 의 값에 따라 강한 친구 관계의 비율이 어떻게 변화하는지 확인해 보았다.

그림 8은 강한 친구 관계 그래프를 만들었을 때, p 와 T 의 값에 따라 그래프의 간선의 수가 원래의 친구 관계 그래프에 비해서 얼마나 남게 되는지 백분율로 표현한 것이다. 이 결과에서 확인할 수 있는 사실은 p 가 10이고 T 가 0인 친구 관계, 즉 10일 동안 상호작용이 한번도 발생하지 않은 친구 관계가 전체의 절반 이상이라

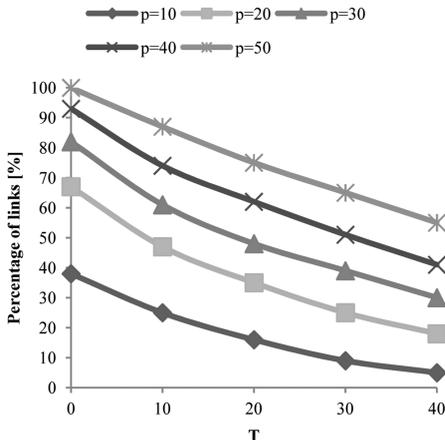


그림 8 T 의 값에 따른 간선 수의 백분율

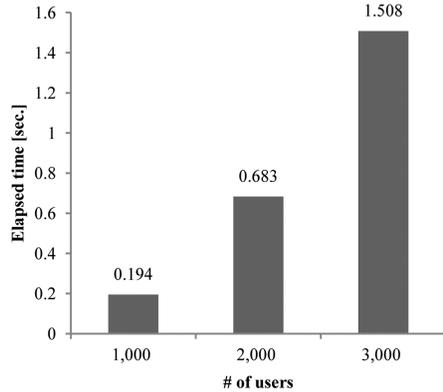


그림 9 사용자 수에 따른 유도 상호작용 그래프의 생성 시간

는 것이다. 심지어, 10일 간 상호작용이 적어도 40번 넘게 발생한 친구 관계가 전체의 10%도 안 되는 것을 확인할 수 있다. 이 사실로 미루어 보아 자주 연락하지 않는 사람과도 친구로 연결되어 있는 온라인 소셜 네트워크는, 최근에 주로 연락하는 사람들로 구성된 실세계의 소셜 네트워크와는 큰 차이를 보인다는 것을 알 수 있다.

실험에서는 상호작용 그래프를 만들기 위한 p 의 값을 30으로 하였으며, 유도 상호작용 그래프를 만들기 위해 p 의 값을 1로 한 관계 강도의 표본 30개를 사용하였다. 또한, 강한 연결의 기준 T 와 V 는 각각 10과 0으로 정하였다.

그림 9는 유도 상호작용 그래프를 생성할 때 사용자 수에 따라서 소요되는 시간을 측정해 본 것이며, Intel® Core™ i3 quad-core 3.07 GHz, 2 GB RAM, Linux 2.6.32의 환경에서 수행하였다. 유도 상호작용 그래프의 생성 시간은 사용자 수에 대해 대략 2차 곡선의 증가 추세를 따르는 것으로 나타났고, 사용자 수가 3,000명일 때 대략 1.5초의 시간이 소요되었다. 이것은 본 논문에서 제안하는 기법이 친구 수가 100명 정도 되는 일반적인 온라인 소셜 네트워크 서비스 사용자나, 사용자 수가 1,000명 가량 되는 사용자 그룹에 대해 매우 빠른 시간 안에 효과적으로 수행되고, 실제 개발이나 분석에 사용될 경우에도 유효하다는 것을 의미한다.

4.3 평가

본 논문에서 제안한 유도 상호작용 그래프가 기존의 방법인 상호작용 그래프에 비하여 어떠한 차이를 나타내는지 살펴보기 위해서, 유도 상호작용 그래프를 사용했을 경우와 상호작용 그래프만 사용했을 경우로 구분해서 강한 친구 관계 그래프를 만들고, 그 강한 친구 관계 그래프들이 각각 얼마나 작은 세상 네트워크[18]에 근접하였는지 비교해 보았다. 작은 세상 네트워크는 그

안에서의 정보 확산 속도가 상당히 빠른 것이 특징으로, 실제계의 소셜 네트워크가 작은 세상 네트워크의 대표적인 예라고 할 수 있다[17]. 따라서, 어떤 친구 관계 그래프가 얼마나 작은 세상 네트워크에 근접했는지 측정하는 것은 그 친구 관계 그래프가 실제계의 소셜 네트워크와 얼마나 유사한 형태를 나타내는지 확인하는 것이라고 할 수 있다. 이를 위해서는 그 친구 관계 그래프의 특성 경로 길이와 군집 계수를 측정해야 한다. 특성 경로 길이 L 은 그래프 상의 임의의 두 정점 사이의 최단 거리를 표현하는 값으로, 그래프를 이루는 정점들의 전형적인 분리(typical separation)의 정도를 나타낸다. 군집 계수 C 는 임의의 정점과 연결된 이웃들이 얼마나 서로 연결되어 있는지 표현하는 값으로, 정점들의 집단(clique)이 형성된 정도를 의미한다. 어떤 네트워크의 L 과 C 의 값을, 그 네트워크와 같은 정점의 수와 평균 차수(degree)로 생성한 임의의 네트워크의 L 과 C 의 값과 비교했을 때 다음의 조건을 만족하면 그 네트워크에서는 작은 세상 현상이 나타난다고 할 수 있다[18].

$$L \gg L_{random} \wedge C \gg C_{random}$$

실험은 표 1과 그림 5에서 볼 수 있듯이 (i) 간선의 관계 강도가 T 보다 큰 경우, (ii) 간선의 관계 강도의 변화 추세가 V 보다 큰 경우, (iii) 간선의 관계 강도가 T 보다 크거나 관계 강도의 변화 추세가 V 보다 큰 경우, (iv) 간선의 관계 강도가 T 보다 크고 관계 강도의 변화 추세가 V 보다 큰 경우, 이렇게 총 네 가지 경우로 구분해서 강한 친구 관계 그래프를 생성하고 이 그래프들의 L 과 C 의 값을 비교하는 방법으로 수행하였다. 이 중에 (i)의 경우가 기존의 방법인 상호작용 그래프만 사용한 경우라고 할 수 있으며, 나머지의 경우는 유도 상호작용 그래프도 함께 사용한 경우이다.

그림 10은 각각의 경우에 대해 강한 친구 관계 그래

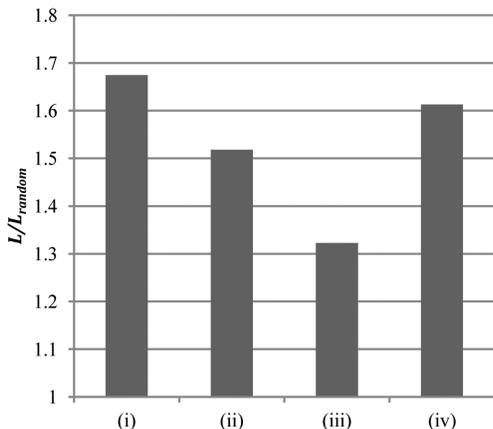


그림 10 강한 친구 관계 그래프의 특성 경로 길이 비교

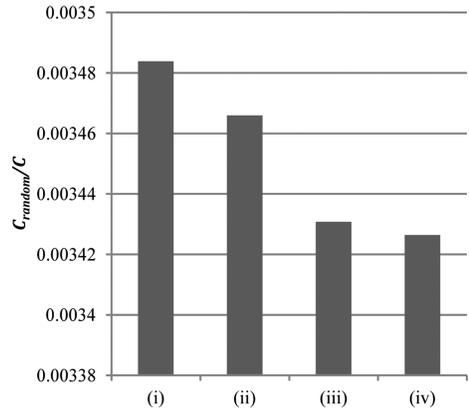


그림 11 강한 친구 관계 그래프의 군집 계수 비교

프의 $\frac{L}{L_{random}}$ 을 나타낸 것이고, 그림 11은 $\frac{C_{random}}{C}$ 을 나타낸 것이다. 두 값 모두 상대적으로 작으면 작을수록 그 그래프는 작은 세상 네트워크에 근접했다고 볼 수 있다. 위의 결과에 따르면, 유도 상호작용 그래프를 사용하여 강한 친구 관계 그래프를 생성한 (ii), (iii), (iv)의 경우가 상호작용 그래프만을 사용한 경우 (i)보다 더 작은 세상 네트워크에 근접한다는 것을 확인할 수 있다. 그 중에서도, 상호작용 그래프와 유도 상호작용 그래프를 함께 사용한 (iii)의 경우가 다른 경우와 비교하여 가장 작은 세상 현상을 보이는 것으로 나타났다.

그림 5에서 볼 수 있듯이, [11]의 연구에서 소개했던 상호작용 그래프만을 사용한 (i)과 본 논문에서 제안하는 유도 상호작용 그래프만을 사용한 (ii)에 서로 OR 연산을 적용한 결과가 (iii)의 경우와 같고 AND 연산을 적용한 결과는 (iv)의 경우가 되는데, 결국 (iii)의 경우 강한 연결의 기준이 느슨해지고 반면에 (iv)의 경우 기준이 엄격해진다. 그러므로 (iii)의 경우에는 보다 상대적으로 간선이 밀집된 강한 친구 관계 그래프가 만들어지며, 반대로 (iv)의 경우에는 간선이 희박한 강한 친구 관계 그래프가 생성된다. 그림 10의 (iii)의 경우를 보면, 간선이 밀집된 정도에 따라 강한 친구 관계 그래프의 특성 경로 길이가 상대적으로 짧아질 수 있지만, (iv)의 경우로 보아 특성 경로 길이가 간선이 밀집된 정도에 항상 비례하지는 않는다는 것을 알 수 있다. 이는 그래프의 간선이 얼마나 잘 밀집되었느냐에 따라 달라질 수 있는 요소이며, 일반적으로는 OR 연산을 적용하였을 경우 특성 경로 길이가 짧아지는 것으로 나타났다. 그림 11에서는 OR 연산을 적용한 (iii)과 AND 연산을 적용한 (iv)의 군집 계수가 (i)이나 (ii)보다 더 나아진 것을 확인할 수 있다. 이 사실로 미루어 보아 (i)에서 정점들이 집단을 형성하도록 하는 데에 방해 요소가 되었던

간선들이 (ii)와의 OR 연산이나 AND 연산을 통해 일부 상쇄되어 결과적으로 강한 친구 관계 그래프의 군집 계수가 향상된 것으로 보인다.

결과적으로, 유도 상호작용 그래프를 사용하여 만들어진 강한 친구 관계 그래프는 그렇지 않은 경우보다 더 실세계의 소셜 네트워크와 유사하다고 볼 수 있으며, 그 안에서의 정보 전파력도 보다 더 강할 것으로 예상해볼 수 있다.

5. 결론 및 향후 연구

온라인 소셜 네트워크는 이진 구조로 이루어진 그 특성 상 실세계의 소셜 네트워크를 정확하게 반영하지 못하기 때문에, 사용자 상호작용이나 사용자 프로필 등의 소셜 네트워크 서비스를 통해 수집한 정보를 최대한 활용하여 측정된 사용자 간의 관계 강도를 가지고 실세계의 소셜 네트워크를 최대한 반영하는 기존의 연구들이 주를 이룬다. 하지만 기존의 연구에서 제안한 방법들은, 값을 측정하는 시점의 관계 강도만 고려함으로 인해 올바른 실세계의 소셜 네트워크를 반영할 수 없는 경우가 생길 뿐만 아니라, 그 사용자 간의 관계 강도가 앞으로 어떻게 변화할 것인지에 대한 근거를 전혀 제시해 주지 못하기 때문에 친구 관계의 추세 분석이나 변화 예측과 같은 작업에는 적합하지 않다. 본 논문에서는 이러한 문제점을 해결하기 위해 유도 상호작용 그래프의 개념과 그 그래프의 간선의 가중치를 이루는 사용자 간의 관계 강도의 변화 추정 기법을 제안하였다. 이 기법은 관계 강도가 선형의 추세를 보인다고 가정하고, 그 추세선의 기울기를 최소제곱추정법을 통해 측정한다. 이러한 방법으로 생성한 강한 친구 관계 그래프는 실험을 통해 기존의 단순한 상호작용 그래프를 사용한 방법에 비해 더 작은 세상 현상이 나타난다는 것을 확인하였다. 또한, 이 기법이 일반적인 환경에서 매우 빠른 시간 안에 효과적으로 수행된다는 것을 보였으며, 실제 개발이나 분석에도 적용할 수 있다는 가능성을 제시하였다.

본 논문에서 제안한 기법을 어떤 사용자의 친구들의 순위를 매기는 데 사용하거나, 공개되지 않은 어떤 사용자의 개인 정보를 그 사용자의 친구들의 공개된 개인 정보를 통해 유추하는 것과 같은 어플리케이션 개발에 활용할 수 있을 것이다. 또한, 본 논문에서는 소셜 데이터의 일부만을 수집하여 비교적 작은 규모의 데이터를 가지고 실험을 수행하였으나, 제안된 기법을 굉장히 큰 소셜 데이터에 적용해 보고, 좀 더 빠른 시간 안에 효율적으로 수행될 수 있도록 알고리즘을 개선시킬 필요가 있다. 또한, MapReduce[22]와 같은 분산 병렬 처리 환경에서도 수행할 수 있게 적용시키는 연구가 필요할 것으로 보인다.

참고 문헌

- [1] M. Maisto, "Twitter Use Growing Daily, Helped By Smartphones: Pew," *eWeek*, Jun. 1, 2012.
- [2] Facebook. <http://www.facebook.com/>.
- [3] Twitter. <http://twitter.com/>.
- [4] E. Eldon, "ComScore: Google+ Grows Worldwide Users From 65 Million In October To 67 Million In November," *TechCrunch*, Dec. 22, 2011.
- [5] J. Nimetz, "Jody Nimetz on Emerging Trends in B2B Social Networking," *Marketing Five*, Nov. 18, 2007.
- [6] P. Domingos, M. Richardson, "Mining the network value of customers," In *Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2001.
- [7] d. m. boyd, N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol.13, no.1, pp.210-230. Oct. 2007.
- [8] M. S. Granovetter, "The Strength of Weak Ties," *The American Journal of Sociology*, vol.78, no.6, pp.1360-1380. May 1973.
- [9] R. Xiang, J. Neville, M. Rogati, "Modeling Relationship Strength in Online Social Networks," In *Proc. of the 19th International Conference on World Wide Web*, WWW, 2010.
- [10] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, "Characterizing User Behavior in Online Social Networks," In *Proc. of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC 2009.
- [11] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, B. Y. Zhao, "User Interactions in Social Networks and their Implications," In *Proc. of the 4th ACM SIGOPS/EuroSys European Conference on Computer Systems*, EuroSys, 2009.
- [12] I. Kahanda, J. Neville. "Using Transactional Information to Predict Link Strength in Online Social Networks," In *Proc. of the 6th International AAAI Conference on Weblogs and Social Media*, ICWSM 2009.
- [13] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, Vol.27, No.8, pp.861-874, Jun. 2006.
- [14] J. A. Hanley, B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol.143, no.1, pp.29-36. Apr. 1982.
- [15] M. McPherson, L. Smith-Lovin, J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, Vol.27, No.1, pp.415-444, 2001.
- [16] LinkedIn. <http://www.linkedin.com/>.
- [17] S. Milgram, "The small world problem," *Psychology today*, vol.2, pp.60-67, 1967.
- [18] D. J. Watts, S. H. Strogatz, "Collective dynamics

of 'small-world' networks," *Nature*, vol.393, pp.440-442, Jun. 4, 1998.

- [19] J. Aldrich, "Fisher and Regression," *Statistical Science*, vol.20, no.4, pp.401-417, 2005.
- [20] J. Aldrich, "Doing Least Squares: Perspectives from Gauss and Yule," *International Statistical Review*, vol.66, no.1, pp.61-81, Apr. 1998.
- [21] Facebook Graph API, <http://developers.facebook.com/docs/reference/api/>.
- [22] J. Dean, S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," In *Proc. of the 6th Conference on Operating Systems Design and Implementation*, OSDI, 2004.



박 지 범

2011년 연세대학교 컴퓨터과학과(학사)
2011년~현재 서울대학교 컴퓨터공학부 석사과정. 관심분야는 데이터베이스, 데이터마이닝, 소프트웨어 공학



임 혜 원

2008년 숙명여자대학교 컴퓨터과학과(학사). 2008년~현재 서울대학교 컴퓨터공학부 석박사통합과정. 관심분야는 데이터베이스, 소셜 네트워크 분석



김 형 주

1982년 서울대학교 전산학과(학사). 1985년 Univ. of Texas at Austin(석사). 1988년 Univ. of Texas at Austin(박사) 1988년~1990년 Georgia Institute of Technology(조교수). 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이터베이스, XML, 시맨틱 웹, 온톨로지