블로그 연결망에서 파급 파워 유저의 파악

손호용*, 임승환*, 김상욱*, 박선주**
*한양대학교 전자컴퓨터공학과
**연세대학교 경영대학
e-mail: iso434@agape.hanyang.ac.kr

Determining Diffusion Power Users in Blog Networks

Ho-Yong Son*, Seung-Hwan Lim*, Sang-Wook Kim, Sunju Park**
*Dept. of Elctronics and Computer Engineering, Hanyang University
**School of Business, Yonsei University

요 약

최근 인터넷 기술의 발달로 인해서 온라인에서 다양한 사회연결망이 출현하였고, 이를 분석하기 위한 연구가 활발히 진행되고 있다. 온라인 사회연결망의 대표적인 예로 블로그 연결망을 들 수 있다. 블로그 연결망에서는 블로그 사용자들이 작성한 게시글들이 다양한 방식을 통하여 다른 사용자들에게 전파된다. 본 논문에서는 이를 게시글이 파급되었다고 부르고, 게시글을 파급한 사용자는 게시글을 소유하고 있는 사용자에게 동화되었다고 부른다. 블로그 내에는 다수의 사용자들에게 컨텐츠를 파급시키는 영향력 있는 사용자들이 존재한다. 본 논문에서는 이러한 사용자들을 파급 파워 유저라고 정의하고, 이러한 사용자들을 파악하는 효과적인 방법에 대하여 논의한다. 본 논문에서는 블로그 연결망에서 파급 파워 유저를 파악하기 위해서 독립 전파 모델을 이용한다. 독립 전파 모델의 수행을 위해서는 사용자들 간의 동화확률을 부여하는 것이 필수적이다. 따라서 본 논문에서는 사용자의 컨텐츠 파워, 재생산 파워의 개념과 이를 계량화하는 방법을 제안하고, 이 값들을 이용하여 사용자간의 동화확률을 부여하는 방안을 제안한다. 끝으로, 실제 블로그 연결망에서 제안하는 기법과 기존의 기법을 이용하여 파워 유저들을 파악하는 실험을 수행하고, 실험결과를 비교 및 분석한다.

1. 서론

블로그(blog)는 사용자가 자신의 글을 온라인상에 게시할 수 있는 일종의 개인 웹사이트이며, 블로그 서비스는 사용자가 블로그를 생성하고 운영할 수 있도록 지원해주는 서비스이다[1][2]. 블로그 서비스 내의 사용자들은 서로 관계를 맺을 수 있으며, 이를 통해서 사회연결망(social network)이 형성된다. 본 논문에서는 이러한 블로그들로 이루어진 사회연결망을 블로그 연결망(blog network)이라고 정의한다.

사회연결망 분석(social network analysis)은 사회연결망을 분석함으로써 그 사회의 특성을 도출하는 연구 분야이다[3]. 기존의 사회연결망 데이터들은 구성원간의 관계 유무에 대한 정보만을 포함하고 있었기 때문에, 기존의 대부분의 연구들은 주로 사회연결망의 위상 구조적인 특성만을 대상으로 분석을 수행하였다[4][5][6][7][8].

블로그 서비스 제공자(blog service provider)는 블로그사용자들이 서비스를 이용한 주요 기록들을 데이터베이스에 저장하고 있다. 본 연구에서는 데이터베이스를 분석하여 블로그 사용자들 간의 관계의 유무에 대한 정보뿐만 아니라 관계의 정도에 대한 정보를 파악하고자 한다. 이러한 구성원간의 관계의 정도에 대한 정보는 기존의 사회연결망 데이터들에는 존재하지 않는 것으로서, 이를 이용하여 사회연결망의 위상 구조만을 분석하던 기존 연구에 비하여 실제적이고 정밀하게 분석하기 위한 연구를 가능하게 한다.

본 논문에서는 블로그 연결망 내의 일반 사용자들과 달리 분석가의 관심측면에서 높은 파워를 가진 사용자를 파워 유저(power user)라고 정의한다. 예를 들어, 파워유저의 한 예로써 다른 사용자들에게 컨텐츠를 파급 시키는 능력이 큰 사용자를 들 수 있다. 본 논문에서는 이러한 사용자를 블로그 연결망에서의 파급 파워 유저(diffusion power user)라고 정의한다. 블로그 연결망에서 파급 파워 유저를 파악함으로써 이들을 중심으로 블로그 서비스 활성화를 위한 다양한 정책을 적용하여 비즈니스의 성공을 도모할 수 있다.

파급 파워 유저를 파악하기 위한 대표적인 방법으로서 독립 전파 모델(independent cascade model)을 이용한 방법을 들 수 있다[9]. 독립 전파 모델은 파급 파워 유저를 파악하기 위해서 사용자간에 파급이 일어날 확률 값을 필요로 한다. 그러나 이 실질적인 확률값을 부여하는 기법에 대한 연구는 미흡한 상태이다.

따라서 본 논문에서는 블로그 연결망에서 독립 전파 모델을 이용하여 파급 파워 유저를 파악하기 위해서 사용자간에 파급이 발생할 확률을 부여하는 방안에 대하여 논의한다. 이를 통해서 실제 응용에서 필요로 하는 파급 파워 유저들을 파악할 수 있다. 또한, 사용자간에 파급이 발생할 확률을 이용하여 커뮤니티 식별[10][11], 사회연결망 클러스터링[10] 등의 그래프 마이닝 연산을 수행할 수 있다. 본 논문의 공헌은 다음과 같다. 첫째, 블로그 연결망 내에서 파급 파워 유저를 파악하기 위해서 사용자간에 파급이 발생할 확률로서 실질적인 값을 부여하는 방안을 제안한다. 둘째, 이를 위하여 사용자의 컨텐츠 파워, 재생산 파워의 개념과 이를 계량화하는 방안을 제안한다. 셋째, 사용자간에 파급이 발생할 확률의 값을 계산하고, 이 값의 정확도를 향상시키기 위한 기법을 제안한다. 넷째, 제안하는 기법과 기존의 기법들을 이용하여 실제 블로그 연결망에서 파급 파워 유저들을 파악하는 실험을 수행하고, 이 결과를 비교 및 분석한다.

2. 관련 연구

사회 연결망에서 파급 파워 유저를 식별하는 문제는 바이럴 마케팅(viral marketing) 분야에서 최대의 마케팅 효과를 거두기 위하여 공략할 소수의 고객들을 식별하기 위한 목적으로 오랫동안 논의되어 왔던 문제이다[12][13]. 파워 유저를 파악하기 위한 기존의 방법으로는 사용자의 위상 구조를 이용한 방법[3], 사용자가 연결망을 통해서 직간접적으로 다른 사용자들에게 영향을 미치는 정도를 계량화한 방법이 제안되었다[14][15][16].

사용자의 위상 구조를 이용하는 방법으로는 사회연결망에서 해당 사용자가 중앙에 위치한 정도를 측정하는 방법들이 제안되었다[17][18][19]. 그러나 이와 같이 사회연결망의 위상 구조적 특징만을 고려하는 기준을 사용하는 경우, 블로그 연결망에서 영향력을 크게 발휘하는 파워 유저를 올바르게 선정할 수 없다. 이는 블로그 사용자의 영향력은 단순히 관계의 개수 보다는 관계의 친밀한 정도에 좌우되기 때문이다. 예를 들어, 블로그 연결망에서 다수의 사용자와 관계를 갖고 있는 사용자보다 적은 사용자와 관계를 갖고 있는 사용자가 다른 사용자들에게 큰 영향력을 행사할 수도 있다.

어떤 사용자가 연결망을 통해서 직간접적으로 다른 사용자들에게 영향을 미치는 정도를 계량화하는 기존의 방법들이 공유하고 있는 기본 아이디어는 다음과 같다. 사용자들은 상호간에 영향을 주고받을 수 있으며, 이 영향에 의해서 특정 사용자의 성향이 영향을 준 사용자의 성향과 같아질 수 있다. 본 논문에서는 이러한 경우에 이 사용자는 영향을 준 사용자에 의해서 동화되었다(assimilate)고 부른다.

참고문헌 [14]에서는 연결 가치(network value)를 이용하는 기법을 제안하였다. 연결 가치는 특정 사용자가 다른 사용자들을 동화시킨 경우에 얻게 되는 이득의 양이며, 이 값이 큰 사용자를 영향력 있는 사용자로 선정한다. 이 기법은 특정 사용자로 인해 얻게 되는 이득의 양에 대한 분석은 가능하지만, 이 사용자로 인해 어떤 사용자들이 동화되었는지는 식별할 수 없기 때문에 이득의 양에 관계없이 최대한 많은 사용자들을 동화시키는 사용자를 파악하기 위해서는 적당하지 않다.

참고문헌 [20]에서는 선형 임계값 모델(linear threshold model)을 제안하였다. 선형 임계값 모델은 사용자마다 임계값을 부여하고, 사용자간의 관계에 가중치를 부여하여 특정 사용자가 주변 사용자들로부터 받은 영향의 정도(가중치)를 누적한 값이 해당 사용자가 갖고 있는 임계값 이상이면, 이 사용자는 영향을 미친 사용자들에 의하여 동화된 것으로 간주한다. 사용자의 파급 파워로서 해당 사용자에 의해서 직간접적으로 동화된 사용자의 수를 측정한다. 그러나 실제 블로그 연결망 내에서 게시글의 파급은 사용자간의 독립적인 관계에 의해서 이루어지는데 반해, 선형 임계값 모델은 여러 사용자들의 영향의 합에 의한 파급을 설명하기 위한 모델이므로 블로그 연결망에 적용하기에는 적당하지 않다.

참고문헌 [9]에서는 독립 전파 모델(independent cascade model)을 제안하였다. 독립 전파 모델은 사용자간의 관계에 확률을 부여하여 사용자 간에 영향을 미칠 때, 이 확률에 의하여 동화 여부를 결정한다. 본 논문에서는 이 값을 사용자간의 동화확률이라고 부른다. 독립 전파 모델은 선형 임계값 모델과 마찬가지로 사용자의 파급 파워로서 해당 사용자에 의해 직간접적으로 동화된 사용자의 수를 측정한다. 특정 사용자가 임의의 게시글을 파급하는 것은 자신의 이웃 사용자들에게 영향을 받아서가 아니라, 해당 게시글을 소유하고 있는 사용자에게만 영향을 받아서 이루어진 것이다. 따라서 독립 전파 모델을 통해서 블로그 연결망에서의 게시글의 파급 현상을 적절하게 설명할 수 있으므로, 독립 전파 모델은 블로그 연결망에서 파급 파워 유저를 파악하기 위한 적당한 방법이다.

참고문헌 [16]에서는 일반화된 전파 모델(general cascade model)을 제안하

였다. 일반화된 전파 모델은 독립 전파 모델에서 특정 사용자를 동화시키기 위해서 이웃의 사용자들이 독립적으로 영향을 미친다는 조건을 제거함으로써 선형 임계값 모델과 독립 전파 모델의 특성을 일반화한 것이다. 따라서 일반화된 전파 모델은 선형 임계값 모델과 독립 전파 모델의 특성을 모두 갖고 있는 파급현상을 설명하기에 적당한 방법이다.

본 논문에서는 앞서 언급한 이유로 인해서 블로그 연결망에서 파워 유저를 파악하기 위해서 독립 전파 모델을 이용한다. 그러나 독립 전파 모델을 이용하여 파급 파워 유저를 파악하기 위해서는 사용자간의 동화확률을 필요로 한다. 정확한 분석 결과를 위해서는 이러한 사용자간의 동화확률은 정확해야 한다. 그러나 기존의 연구들에서는 주로 연결망 내에서의 파급 관계를 설명하는 모델을 제안하는 데에 초점을 맞추고 있었으므로, 사용자간의 동화확률 값을 부여하기 위한 방안에 대한 연구는 미흡한 상태이다. 따라서 본 논문에서는 파급 파워 유저를 정확하게 파악하기 위해서 사용자간의 동화확률을 부여하기 위한 방안에 대하여 논의한다.

3. 제안하는 기법

3.1. 용어 정리

표 1은 앞으로의 논의 전개를 위해서 필요한 용어 및 기호들을 정리한 것이다. U_i는 식별자가 i인 사용자를 의미한다. D_i는 U_i가 소유한 게시글들의 집합을 의미하고, $D_{i,j}$는 U_i의 j번째 게시글을 의미한다. 게시글 $D_{i,j}$가 사용자들에게 미치는 컨텐츠 영향력을 이 게시글의 컨텐츠 파워(document contents power)라고 정의하며, $DCP(D_{i,j})$로 표기한다. 또한, 사용자 U_i가 다른 사용자들에게 미치는 컨텐츠 영향력을 사용자의 컨텐츠 파워(user contents power)라고 정의하고, $UCP(U_i)$로 표기한다.

만일, 사용자 U_i가 큰 컨텐츠 파워를 갖는다면, 사용자 U_i는 양질의 컨텐츠를 소유하고 있다고 볼 수 있다. 이에 착안하여 본 논문에서는 사용자의 컨텐츠의 질의 척도로서 사용자의 컨텐츠 파워를 이용한다. 또한, 블로그 사용자들은 자신이 직접 게시글을 작성하는 것 외에도 다른 사용자들의 게시글들을 스크랩하거나 엮인글을 달 수 있는데, 본 논문에서는 사용자의 이러한 행동을 컨텐츠 재생산이라고 부른다. 사용자가 컨텐츠를 재생산하는데 적극적인 정도를 사용자의 재생산 파워(user delivery power)라고 정의하며, 사용자 U_i의 재생산 파워를 $UDP(U_i)$로 나타낸다.

사용자가 블로그 서비스를 이용할 때, 취할 수 있는 액션으로는 게시글 작성(write), 조회하기(read), 덧글 남기기(comment), 스크랩 하기(scrap), 엮인글 달기(link)의 다섯 가지가 있으며, 이러한 액션을 각각 W, R, C, S, L로 표기한다. 게시글의 컨텐츠 영향력을 계량화할 때, 각각의 액션에 다른 의미를 부여하기 위하여 서로 다른 가중치를 할당할 수 있다. 액션 W, R, C, S, L을 위한 가중치는 w_w, w_r, w_c, w_s, w_l로 표기한다.

<표 1> 용어정의

U_i: user I
$D_i = \{D_{i,1}, D_{i,2}, ...\}$: A collection of documents owned by U_i
$D_{i,j}$: Document j of user i
$DCP(D_{i,j})$: Document Content Power of $D_{i,j}$
$UCP(U_i)$: User Content Power of U_i
$UDP(U_i)$: User Delivery Power of U_i
ActionType = {W, R, C, S, L}: Activities of a user in a blog network
ActionWeight = $\{w_w, w_r, w_c, w_s, w_l\}$: Weights of actions

3.2. 컨텐츠 파워

본 논문에서는 사용자의 컨텐츠의 질의 척도로서 사용자의 컨텐츠 파워를 사용한다. 본 절에서는 참고문헌 [21]에서 제안된 사용자의 컨텐츠 파워를 측정하는 방안을 소개한다.

3.2.1. 게시글의 컨텐츠 파워

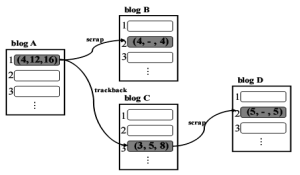
특정 게시글에 대하여 다른 사용자들이 액션을 보인다는 것은 이 게시글로 인하여 영향을 받았음을 의미한다. 이와 같은 사실에 착안하여 본 연구에서는 게시글의 컨텐츠 파워를 계량화하기 위해서 각 게시글에 대한 사용자들의 액션의 가중치와 빈도를 곱하는 방법을 사용한다.

각 게시글은 처음으로 작성된 블로그 내에서 다른 사용자들에게 직접적으로 영향을 미칠 수 있고, 스크랩 하기나 엮인글 달기를 통해서 전파된 다른 블로그 내에서 다른 사용자들에게 간접적으로 영향을 미칠 수 있다. 본 연구에서는 전자를 게시글의 직접적인 영향, 후자를 게시글의 간접적인 영향이라고 부르며, 이를 계량화한 값을 각각 해당 게시글의 직접 컨텐츠 파워(direct contents power), 간접 컨텐츠 파워(indirect contents power)라고 정의한다. 각 게시글의 컨텐츠 파워를 해당 컨텐츠의 직접 컨텐츠 파워와 간접 컨텐츠 파워의 합으로 계산한다. 여기서, 직접 컨텐츠 파워와 간접 컨텐츠 파워의 반영 비율은 응용에 따라 각각에 대한 가중치 w_D와 w_I를 부여하여 조절 할 수 있다. 표 2는 게시글의 컨텐츠 파워의 계량화 방법을 정리한 것이다.

<표 2> 게시글의 컨텐츠 파워 계산

$DCP(D_{i,j}) = w_D * D\_DCP(D_{i,j}) + w_I * I\_DCP(D_{i,j})$
$D\_DCP(D_{i,j})$
$= w_r * R\_Count(D_{i,j}) + w_c * C\_Count(D_{i,j}) + w_s * S\_Count(D_{i,j}) + w_l * L\_Count(D_{i,j})$
$I\_DCP(D_{i,j}) = w_D * \sum D\_DCP(D_{r,j'}) + w_I * \sum I\_DCP(D_{r,j'})$, where $D_{r,j'}$ represents documents reproduced from $D_{i,j}$

그림 1은 게시글들 간의 파급 관계를 이용하여 게시글의 컨텐츠 파워를 계산하는 과정의 예를 보인 것이다. 여기서는 직접 컨텐츠 파워와 간접 컨텐츠 파워의 가중치를 각각 동일하게 1로 설정한 경우를 대상으로 하였다. 각 게시글 안의 값들은 <직접 컨텐츠 파워, 간접 컨텐츠 파워, 전체 컨텐츠 파워> 를 의미한다. $D_{D,2}$는 이후에 파급된 기록이 없으므로 간접 컨텐츠 파워가 0이며, 전체 컨텐츠 파워는 직접 컨텐츠 파워인 5가 된다. 따라서 $D_{C,3}$의 간접 컨텐츠 파워는 5가 되고, 이 값에 직접 컨텐츠 파워인 3을 더하여 $D_{C,3}$의 전체 컨텐츠 파워는 8이 된다. 또한, 게시글의 원본이 되는 $D_{A,1}$의 간접 컨텐츠 파워는 $D_{C,3}$의 전체 컨텐츠 파워인 8과 $D_{B,2}$의 전체 컨텐츠 파워인 4를 합한 12가 되고, 여기에 직접 컨텐츠 파워인 4를 더한 16이 된다.



(그림 1) 게시글의 컨텐츠 파워 계산

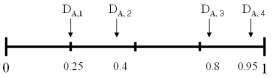
3.2.2. 사용자 컨텐츠 파워

사용자의 컨텐츠 파워는 사용자 블로그에 등록되어 있는 모든 게시글들의 컨텐츠 파워를 이용하여 계산한다. 게시글의 컨텐츠 파워는 블로그에 등록된 이후 노출된 시간에 비례하여 증가하는 경향이 있다. 따라서 오래전에 등록된 게시글은 최근에 등록된 게시글에 비하여 실질적인 컨텐츠의 영향력은 작더라도 오랜 노출 시간(exposed time) 때문에 큰 컨텐츠 파워를 가지는 것으로 왜곡될 수 있다. 본 연구에서는 이러한 문제를 해결하기 위해서 각 게시글의 컨텐츠 파워에 등록 이후의 노출 시간을 반영하는 방법을 사용한다. 즉, 전체 블로그 데이터의 분석 기간을 1로 간주하고, 해당 게시글이 등록된 시각을 참조하여 해당 게시글의 상대적 노출 시간을 계산한다. 이 노출 시간의 역수를 게시글의 컨텐츠 파워에 곱함으로써 노출 시간의 차이로 인한 왜곡을 보정한다.

(식 1)은 게시글의 컨텐츠 파워를 이용하여 사용자의 컨텐츠 파워를 계산하는 방법을 나타낸 것이다. 여기서, $ET_{Di,j}$는 게시글 $D_{i,j}$의 상대적 노출 시간의 역수를 의미한다.

$$UCP(U_i) = \sum_j ET_{Di,j} * DCP(D_{i,j}) \qquad \text{(식 1)}$$

그림 2는 (식 1)을 이용하여 $UCP(U_A)$를 계산하는 과정을 나타낸 것이다. 이 예에서 $UCP(U_A) = 50*0.25 + 35*0.4 + 35*0.8 + 10*0.95 = 64$가 된다.



Document	DCP	Inverse of Relative Exposure Time
$D_{A,1}$	50	0.25
$D_{A,2}$	35	0.4
$D_{A,3}$	35	0.8
$D_{A,4}$	10	0.95

(그림 2) 사용자 컨텐츠 파워 계산의 예

3.3. 사용자 재생산 파워

사용자가 컨텐츠를 재생산하는데 적극적인 정도는 사용자가 취한 재생산 액션들의 기록을 이용하여 계량화할 수 있다. 재생산 액션의 종류는 블로그 서비스 환경과 분석가에 따라서 다양하게 결정할 수 있다. 본 논문에서는 사용자의 재생산 액션으로서 스크랩하기와 엮인글 달기를 고려하였다. (식 2)는 사용자 U_i의 재생산 파워를 $UDP(U_i)$의 계량화 방법을 나타낸 것이다. 사용자의 재생산 파워를 계량화하기 위해서 재생산 액션들의 빈도수와 각 재생산 액션들의 가중치의 곱을 더하는 방법을 이용하였다.

$$UDP(U_i) = w_s * S\_Count(U_i) + w_l * L\_Count(U_i) \qquad \text{(식 2)}$$

3.4. 사용자간의 동화확률 측정

U_A의 컨텐츠가 U_B에게 파급될 확률 $P_{A->B}$는 U_A의 컨텐츠의 질과 U_B의 컨텐츠 재생산에 적극적인 정도에 영향을 받는다. 따라서 본 논문에서는 동화확률 $P_{A->B}$를 계산하기 위해서 U_A의 컨텐츠 파워 $UCP(U_A)$와 U_B의 재생산 파워 $UDP(U_B)$를 이용하는 방법을 제안한다.

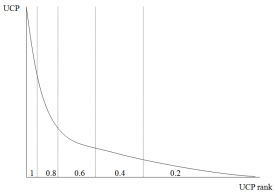
3.4.1. 사용자간의 동화확률

동화확률 $P_{A->B}$는 $UCP(U_A)$와 $UDP(U_B)$에 의하여 결정된다. 이 때, $P_{A->B}$를 결정하는데 $UCP(U_A)$와 $UDP(U_B)$이 반영된 비율을 측정할 수 있다면, $P_{A->B}$는 $UCP(U_A)$와 $UDP(U_B)$에 각각의 반영비율을 곱하여 계산할 수 있을 것이다. 따라서 본 논문에서는 동화확률을 결정하는데 이용된 사용자 컨텐츠 파워와 사용자 재생산 파워의 비율을 측정하는 방법을 제안하고, 측정된 비율을 이용하여 동화확률을 계산하는 방법을 제안한다. (식 3)은 w_{CP}와 w_{DP}를 이용하여 $P_{A->B}$를 계산하는 방법을 나타낸 것이다.

$$P_{A->B} = w_{CP} * CP_A + w_{DP} * DP_B \text{ (단, } w_{CP}=1-w_{DP}) \text{ (식 3)}$$

여기서, $P_{A->B}$는 확률값으로서 0에서 1사이의 값을 가지므로, $UCP(U_A)$와 $UDP(U_B)$를 각각 0에서 1사이로 정규화한 값인 CP_A와 DP_B로 변환하여 계산한다. 사용자의 컨텐츠 파워를 정규화하기 위해서 전체 사용자들의 컨텐츠 파워값의 분포영역을 동일한 사용자 수를 갖는 임의의 개수의 그룹으로 나누어 각 그룹 내의 사용자들이 동일한 CP 값을 갖도록 하였다. 그림 3은 사용자의 컨텐츠 파워를 정규화하는 방법을 나타낸 것이다. 이 예에서 그룹의 수는 5개로 설정하였다. 그룹의 개수는 분석가가 분석 대상인 블로그 연결망의 특성을 고려하여 결정한다. 각 그룹안의 숫자는 각 그룹에 포함된 사용자들의 컨텐츠 파워를 정규화한 값을 나타낸다.

사용자의 재생산 파워도 사용자의 컨텐츠 파워와 동일한 방법으로 정규화한다.



(그림 3) 컨텐츠 파워 정규화의 예

3.4.2. 가중치 결정

이상적인 w_{CP}와 w_{DP}는 해당 w_{CP}와 w_{DP}를 이용하여 사용자간의 동화확률을 부여하여 파급을 예측하였을 때, 가장 높은 정확도를 갖는 값이다. 파급 예측 결과의 정확도를 측정하는 방법은 3.4.3절에서 논의한다. 가장 높은 정확도를 갖는 w_{CP}와 w_{DP}를 찾기 위해서 가능한 모든 w_{CP}와 w_{DP}에 대해서 사용자의 동화확률을 부여하고 파급을 예측하여 정확도를 비교하는 것은 비현실적인 방법이다.

따라서, 본 논문에서는 라그랑지 보간법(lagrange interpolation)을 이용하여 가장 높은 정확도를 갖는 w_{CP}와 w_{DP}의 값을 측정하는 방법을 제안한다. 라그랑지 보간법은 좌표평면 위에 주어진 k개의 점을 모두 지나는 다항식을 찾는 기법이다[22]. 따라서 k개의 w_{CP}와 w_{DP}를 통해서 파급 정확도들을 측정하고, 이 값들을 토대로 라그랑지 보간법을 이용하여 w_{CP}와 w_{DP}의 변화에 따른 파급 예측 정확도의 변화를 예상하여 가장 높은 정확도를 갖는 w_{CP}와 w_{DP}를 선택한다.

4. 블로그 연결망 분석

4.1. 실험 환경

본 연구에서는 성능 분석을 위하여 한국의 블로그 포털 사이트로부터 2006년 7월부터 약 7개월간 수집한 데이터를 사용하였다. 분석 기간 중에 생성된 게시글의 수는 약 100,000,000개 이며, 블로그 연결망의 구성을 위해서 블로그들 간의 관계로서 블로그 간의 이웃 여부가 아니라, 블로그 간에 교류의 빈번한 정도가 일정 이상인 경우를 관계로 설정하였다. 이는 블로그 간에 이웃관계가 설정되어 있어도 빈번한 교류를 갖지 않는 경우가 많다는 사실에 근거한 것이다.

본 실험에서는 파워 유저 선정 기법간의 성능 비교를 위해서 사용자의 위상 구조적인 특성만을 이용하는 기법과 사용자간의 관계의 정보를 분석한 기법들을 선정하였다. 본 논문에서 비교하는 사용자의 관계의 정도를 분석한 기법들은 덧글의 개수를 이용한 기법, 게시글들의 스크랩 및 엮인글의 개수를 이용한 기법, 컨텐츠 파워를 이용한 기법, 재생산 파워를 이용한 기법, 파급 파워를 이용하여 파워 유저를 선정하는 기법이며, 각 기법들에 대한 설명은 다음과 같다.

사용자의 위상 구조적인 특성만을 이용한 기법으로는 사용자간의 서로이웃 관계를 이용하여 연결 중앙성 기법인 Degree(DEG)를 사용하였다. 인접 중앙성[23]과 사이 중앙성[24]을 이용하는 기법은 정점이 수천만에 이르는 실제 블로그 연결망으로의 적용이 현실적으로 불가능하므로 본 실험에서의 평가 대상에서 제외하였다.

덧글의 개수를 이용하는 기법으로는 직접적으로 생산한 게시글들의 덧글의 개수를 이용하는 기법인 Comment Direct(C_D), 간접적으로 생산한 게시글들의 덧글의 개수를 이용하는 기법인 Comment Indirect(C_I), 직접적으로 생산한 게시글들과 간접적으로 생산한 게시글들의 덧글의 개수를 모두 이용하는 기법인 Comment_Total(C_T)를 사용하였다.

스크랩 및 엮인글의 개수를 이용한 기법으로는 직접적으로 생산한 게시글들의 스크랩 및 엮인글의 개수를 이용한 기법인 Trackback & Scrap Direct(T&S_D), 간접적으로 생산한 게시글들의 스크랩 및 엮인글의 개수를 이용한 기법인 Trackback & Scrap Indirect(T&S_I), 직접적으로 생산한 게시글들과 간접적으로 생산한 게시글들의 스크랩 및 엮인글의 개수를 이용한 기법인 Trackback & Scrap Total(T&S_T)를 사용하였다.

사용자의 컨텐츠 파워를 이용한 기법으로는 직접적으로 생산한 게시글들만을 대상으로 계산한 컨텐츠 파워를 이용하는 기법인 ContentsPower Direct(CP_D), 간접적으로 생산한 게시글들만을 대상으로 계산한 컨텐츠 파워를 이용하는 기법인 ContentsPower Indirect(CP_I), 직접적으로 생산한 게시글들과 간접적으로 생산한 게시글들을 대상으로 계산한 컨텐츠 파워를 이용하는 기법인 ContentsPower Total(CP_T)을 사용하였다.

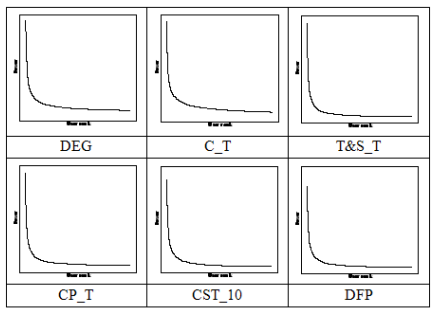
사용자의 재생산 파워를 이용한 기법은 Delivery Power(DP)로 나타내고, 스크랩 및 엮인글의 개수를 이용한 기법으로는 직접적으로 생산한 게시글들의 스크랩 및 엮인글의 개수를 이용한 기법인 Trackback & Scrap Direct(T&S_D), 간접적으로 생산한 게시글들의 스크랩 및 엮인글의 개수를 이용한 기법인 Trackback & Scrap Indirect(T&S_I), 직접적으로 생산한 게시글들과 간접적으로 생산한 게시글들의 스크랩 및 엮인글의 개수를 이용한 기법인 Trackback & Scrap Total(T&S_T)를 사용하였다.

사용자의 파급 파워를 이용한 기법은 독립 전파 모델의 수행을 위해서 필요로하는 사용자간의 파급확률을 부여하는 정책에 따라 나뉘며, 모든 관계의 파급확률로서 1%를 부여한 Constant1(CST_1), 모든 관계의 파급확률로서 10%를 부여한 Constant10(CST_10), 본 논문에서 제안하는 기법을 이용하여 사용자간의 파급확률을 부여하는 기법인 DiffusionPower(DFP)를 사용하였다.

기법 (CP_D), 기법 (CP_I), 기법 (CP_T), 기법 (DP), 기법 (DFP)에서 사용한 사용자 액션의 가중치는 해당 도메인전문가의 의견을 반영하여 게시글 작성 7, 조회하기 1, 덧글 남기기 3, 스크랩 하기 7, 엮인글 달기 7로 설정하였다. 또한, 기법 (CP_T)와 기법 (DFP)에서 사용하는 사용자의 컨텐츠 파워로서 직접 컨텐츠 파워와 간접 컨텐츠 파워의 반영 비율은 각각 동일하게 1로 설정하였다.

4.2. 결과 분석

실험 1에서는 각 기법들을 이용하여 사용자들의 파워를 계산하고, 이 값들의 분포를 분석하였다. 그림 4는 각 기법을 이용하여 측정한 사용자들의 파워 분포를 나타낸 것이다. 그림의 식별을 가능하게 하기 위해서 지나치게 크거나 작은 파워를 갖는 사용자들을 제외한 20위부터 10,000위에 해당하는 사용자들만을 나타내었다. 또한, 지면의 제약으로 인해서 기법(DEG), 기법(C_T), 기법(T&S_T), 기법(CP_T), 기법(CST_10), 기법(DFP)만을 제시하였다. x축은 파워를 기준으로 정렬한 사용자들을 의미하고, y축은 해당 사용자의 파워를 의미한다.



(그림 4) 각 기법을 이용하여 측정한 사용자들의 파워 분포

측정한 사용자들의 파워 분포는 모두 멱함수를 따르는 것으로 나타났다. 이는 블로그 연결망에는 매우 큰 파워를 갖는 소수의 사용자들이 작은 파워를 갖는 다수의 사용자들에게 거의 일방적으로 영향을 미치고 있다는 것을 의미한다. 이러한 결과는 블로그 연결망의 활성화를 위한 비즈니스 정책을 수행할 때, 모든 사용자를 대상으로 하지 않고, 영향력있는 소수의 파워 유저만을 대상으로 하기위한 근거자료가 될 수 있다.

실험 2에서는 각 기법들을 이용하여 30명의 파워 유저 집합을 선정하고 이를 비교하였다. 파워 유저 집합을 선정하기 위해서 기법(CST_1), 기법(CST_10), 기법(DFP)는 hill climbing 기법을 이용하여 파워 유저들을 선정하였고, 그 외의 기법들은 실험 1에서 계산한 사용자들의 파워를 이용하여 큰 파워를 갖는 사용자의 순으로 선정하였다. 그림 5는 실험 2의 결과를 보인 것이다. 각 기법을 이용하여 선정한 파워 유저들을 파워를 기준으로 정렬하였다. 회색바탕의 사용자는 기법(DEG)를 통해서 선정한 파워유저에 포함된 사용자 중에서 다른 기법을 통해서 선정한 파워유저 집합에도 포함된 사용자를 의미한다.

rank	DEG	Comment Count			Scrap & Link Count			Contents Power			ICM	
		C_D	C_I	C_T	SL_D	SL_I	SL_T	CP_D	CP_I	CP_T	CST	DP
1	21608	1362815	1283112	1362815	307819	521114	307819	307819	205649	307819	227694	227694
2	1212679	954535	660655	954535	227694	579543	227694	227694	521114	227694	307819	307819
3	1223870	1049339	38291	1049339	360763	228000	360763	360763	1066638	521114	976756	976756
4	307819	512942	481166	512942	976756	1283112	976756	957182	609211	360763	1153439	1153439
5	173326	1056536	1009253	1056536	240513	481166	957182	240513	38291	957182	1007110	1007110
6	571678	999317	938203	999317	957182	33801	240513	976756	228000	205649	392003	966173
7	373712	1188823	579543	1188823	1007110	938203	1007110	1007110	156438	976756	782455	392003
8	360430	307819	156438	307819	1026046	205649	1326046	275612	579543	240513	393149	561856
9	720669	1341008	269898	1341008	275612	1139729	275612	1329246	1283112	1007110	407690	407690
10	1030692	523315	971401	523315	1254859	530553	1254859	1254859	29155	1153439	13794	999317
11	221719	1007110	916298	1007110	205856	1032290	205856	153439	481166	275612	966173	393149
12	227694	1154029	120838	1154029	1153439	957182	579543	1050536	330592	1329246	240513	13794
13	1269627	948535	637961	948535	782455	538898	1153439	954535	1139729	1254859	213324	48952
14	495395	532957	305166	532957	1041565	708370	782455	1214458	1191046	1050536	561856	240513
15	191454	1120363	521114	1120363	392003	506733	1041565	84535	591032	1214458	84463	417359
16	1139525	957182	274164	957182	1214458	57297	392003	1120363	660655	954535	1214458	17470
17	1263255	1373133	1119856	1373133	407690	591032	1214458	1041565	1336501	579543	999317	90824
18	1244929	919925	1256516	719984	191454	1009053	407690	782455	1223256	999317	275612	213324
19	108546	1182178	1108529	318925	493135	513037	191454	208456	1384606	84535	17470	84463
20	127904	1214458	358220	1182178	1120363	29155	493135	999317	938203	1120363	49552	1279433
21	1050536	417359	162190	1295555	84535	802601	1120363	493135	975704	782455	417359	545303
22	351335	1295555	1151234	1214458	140607	51090	140607	191454	51090	1041565	1279433	275612
23	985608	15167	14938	417359	1050536	1336501	84535	140607	1009253	493135	360763	733732
24	629130	223291	1069062	15167	540950	1053668	84463	417359	33801	1409037	480038	782455
25	608703	719984	33801	223291	393149	1136222	1050536	371450	1360719	191454	547062	1214458
26	709532	1153439	485842	1010929	337888	745863	340950	1362915	1354168	208456	84535	208456
27	561782	1010929	705290	1153439	84463	764408	393149	337888	745863	84463	371450	480038
28	433300	360763	84463	994963	371450	409057	337888	1049339	269898	417359	1406071	360763
29	639	994963	975704	360763	561856	132075	1051059	1341008	1032290	371450	208456	1287441
30	336602	69859	361657	69859	547062	611325	547062	84463	513037	966173	99824	371450

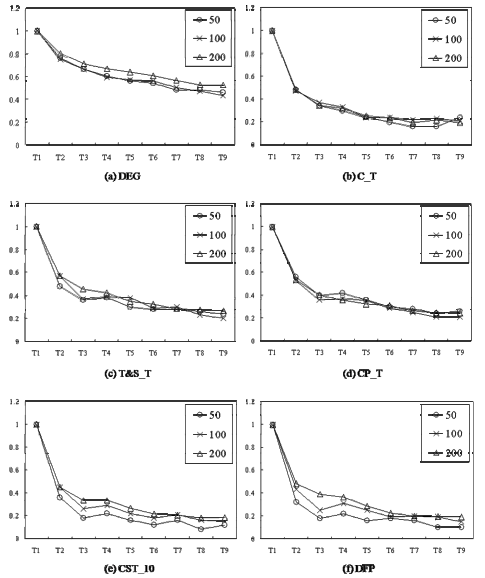
(그림 5) 각 기법에 따른 파워 유저 집합의 결과

실험 결과, 기법(DEG)에 의하여 파워유저로서 선정된 사용자 중에서 사용

자 307819, 사용자 227694, 사용자 191454, 사용자 1050536만이 다른 기법에 의하여 선정된 파워유저에 포함되었다. 이러한 사실은 사용자간의 관계의 정보를 분석하여 파워유저를 선정하는 기법들의 결과와 사용자의 위상구조만으로 파워유저를 선정하는 기법의 결과는 매우 다르다는 것을 의미한다. 따라서 파워유저 선정을 위해서는 선정하고자 하는 파워유저의 특성에 맞는 기법을 사용해야함을 알 수 있다.

실험 3에서는 시간의 흐름에 따라 각 기법을 이용하여 선정한 파워유저 집합의 생존률 변화를 비교한다. 이를 위해서 전체 분석 구간을 25일 간격으로 총 9개의 구간으로 나누어 실험을 수행하였다. 파워유저 집합의 생존률은 분석구간 1에서의 파워유저 집합이 이후의 분석구간에서 선정한 파워유저 집합에 포함되는 정도를 의미한다. 이러한 분석을 통해서 특정 시점에서 각 기법을 이용하여 선정한 파워유저들이 향후에도 파워유저로서 선정될 가능성을 예상할 수 있으며, 파워유저를 대상으로 비즈니스 정책을 수행할 때, 파워유저 집합을 선정한 시점부터 향후 어느 시점까지를 파워유저로서 인정할 것인지를 결정하는데 도움을 줄 수 있다.

그림 6은 실험 3의 결과를 나타낸 것이다. 그림 4에서와 마찬가지로 기법(DEG), 기법(C_T), 기법(T&S_T), 기법(CP_T), 기법(CST_10), 기법(DFP)만을 제시하였다. 선정한 파워유저 집합의 크기는 50명, 100명, 200명으로 설정하였다. x축은 분석구간 1과 비교 대상이 되는 분석구간을 나타내고, y축은 생존률로서 분석구간 1과 각 분석구간과의 파워유저 집합의 일치 정도를 의미한다.



(그림 6) 각 기법을 이용하여 측정한 사용자들의 파워 분포

실험 결과, 기법과 파워유저 집합의 크기에 관계없이 시간의 흐름에 따라 생존률은 감소하였으며, 대체적으로 파워유저 집합의 크기가 클수록 높은 생존률을 보였다. 이러한 결과는 블로그 연결망 내에는 큰 파워를 갖는 사용자들 간의 순위 변동이 빈번하게 발생하고 있다는 것을 의미한다. 따라서 이 결과는 파워 유저들을 대상으로 하는 비즈니스 정책을 수행하기 위해서 해당 파워 유저들의 파워유저로서의 인정기간을 고려하는 데에 사용될 수 있다. 예를 들어, 기법(DFP)를 이용하여 200명의 파워유저들을 선정하여 생존률이 0.4 이상인 경우에만 비즈니스 정책을 수행하고자 한다면, 이 파워유저들은 75일의 기간 동안에만 파워유저로서 인정해야 함을 알 수 있다.

5. 결론

본 논문에서는 블로그 연결망에서 파급 파워 유저를 파악하는 방안에 대하여 논의하였다. 이를 위해서 각 사용자의 컨텐츠 파워와 재생산 파워를 계량화하는 방안을 제안하였고, 이를 이용하여 사용자간의 동화확률로서 실질적인 값을 부여하는 방안을 제안하였다.

본 논문의 주요 공헌은 다음과 같다. 첫째, 사용자간의 동화확률을 측정하기 위하여 사용자의 컨텐츠 파워와 재생산 파워의 개념을 제안하였다. 둘째, 블로그 연결망에서 게시글들의 실질적인 영향력을 분석함으로써 각 사용자의 컨텐츠 파워를 계량화하는 방안을 제안하였다. 셋째, 사용자 컨텐츠 파워의 정확도를 향상시키기 위하여 게시글들의 영향력을 노출 시간에 따라 보정하는 방안을 제안하였다. 넷째, 사용자 재생산 파워를 계량화하는 방안을 제안하였다. 다섯째, 계산된 사용자간의 동화확률의 정확도를 측정하는 기법을 제안하고, 이를 이용하여 가장 높은 정확도를 갖는 값을 사용자간의 동화확률로서 결정하는 기법을 제안하였다. 여섯째, 제안하는 기법과 기존의 기법들을 이용하여 실제 블로그 연결망에서 파워 유저들을 파악하는 실험을 수행하고, 이 결과를 비교 및 분석하였다.

본 논문에서는 블로그 연결망에서 실질적인 사용자간의 동화확률을 부여하여 파급 파워 유저를 분석하는 데에 사용함을 보였다. 사용자간의 동화확률은 파급 파워 유저 파악 이외에도 다양한 그래프 마이닝 연산에 이용될 수 있다. 따라서 향후 연구과제로서 블로그 연결망에 부여한 사용자간의 파급확률을 이용하여 커뮤니티 식별, 클러스터링 등을 수행하기 위한 연구를 추진 중이다.

6. 참고문헌

[1] (주) SK Communications, http://www.cyworld.com.

[2] (주) NHN, http://blog.naver.com.

[3] S. Wasserman and K. Faust, *Social Network Analysis : Methods and Applications*, Cambridge University Press, 1994.

[4] Albert. H. Jeong and A. Barabasi, "Diameter of the World Wide Web," *Nature*, Vol. 401, pp. 130-131, 1999.

[5] Jeong et al., "The Large-Scale Organization of Metabolic Networks," *Nature*, Vol. 407, pp. 651-654, 2000.

[6] Milgram, "The Small World Problem," *Physiology Today*, Vol. 2, pp. 60-67, 1967.

[7] Redner, "How Popular Is Your Paper?," *European Physics Journal B*, Vol. 4, No. 2, pp. 131-134, 1998.

[8] Watts and S. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature*, Vol. 393, pp. 440-442, 1998.

[9] Jacob Goldenberg, Barak Libai, and Eitan Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, Vol. 12(3), pp. 211-223, 2001.

[10] M. Girvan and M. Newman, "Community Structure in Social and Biological Networks," *National Academic Science*, Vol. 99, No. 12, pp. 7821-7826, 2002.

[11] A. Chin and M. Chignell, "A Social Hypertext Model for Finding Community in Blogs," In *Proc. Int'l. Conf. on Hypertext and Hypermedia*, pp. 11-22, 2006.

[12] J. Brown and P. Reinegen, "Social Ties and Word-of-Mouth Referral Behavior," *Journal of Consumer Research*, Vol. 1, No. 3, pp. 350-362, 1987.

[13] J. Coleman, H. Menzel, and E. Katz, *Medical Innovations: A Diffusion Study*, Bobbs Merrill, 1966.

[14] Domingos, P. and M. Richrdson, "Mining the Network Value of Customers," In *Proc. ACM Int'l. Conf. on Knowledge Discovery and Data*, SIGKDD, pp. 57-66, 2001.

[15] Xiaodan Song, Yun Chi, Koji Hino and Belle L. Tseng, "Information Flow Modeling based on Diffusion Rate for Prediction and Ranking," In *Proc. Int'l. Conf. on World Wide Web*, WWW, pp. 191-200, 2007.

[16] David Kempe, Jon Kleinberg, and Eva Tardos, "Maximizing the spread of influence through a social network," In *Proc. ACM Int'l. Conf. on Knowledge Discovery and Data*, SIGKDD, pp. 137-146, 2003.

[17] A. Bavelas, "Communication patterns in task-oriented groups," *Journal of the Acoustical Society of America*, Vol. 22, pp. 271-282, 1950.

[18] C. Proctor and C. Loomis, "Analysis of sociometric data, Research Methods in Social Relations," pp. 561-586, 1951.

[19] A. Shimbel, "Structural parameters of communication networks," *Bulletin of Mathematical Biophysics*, Vol. 15, pp. 501-507, 1953.

[20] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, Vol. 83, No. 6, pp. 1420-1443, 1978.

[21] 김 형준, 임 승환, 김 상욱, 박 선주, "블로그 연결망에서 컨텐츠 파워 유저의 파악 방안," *한국정보처리학회 춘계학술발표대회 논문집*, Vol. 14, No. 1, pp. 67-68, 2007

[22] Waring, E. Philos. Trans. 69, 59-67, 1979.

[23] A. Bavelas, "Communication patterns in task-oriented groups," Journal of the Acoustical Society of America, Vol. 22, pp. 271-282, 1950.

[24] A. Shimbel, "Structural parameters of communication networks," Bulletin of Mathematical Biophysics, Vol. 15, pp. 501-507, 1953.

6. 감사의 글

본 연구는 NHN(주)의 지원을 받았습니다. 또한, 지식경제부 및 정보통신연구진흥원의 대학IT연구센터 지원사업 (IITA-2008-C1090-0801-0040)의 부분적인 지원을 받았습니다.