

블로그 연결망에서 파급력을 가진 파워 유저의 파악 기법

(Determining Diffusion Power Users in a Blog Network)

임 승 환[†] 김 상 욱^{**} 박 선 주^{***}
(Seung-Hwan Lim) (Sang-Wook Kim) (Sunju Park)

요약 블로그 내에는 다수의 사용자들에게 콘텐츠를 파급시키는 영향력 있는 사용자들이 존재한다. 본 논문에서는 이러한 사용자들을 파급 파워 유저라고 정의하고, 이러한 사용자들을 파악하는 방법에 대하여 논의한다. 본 논문에서는 블로그 연결망에서 파급 파워 유저를 파악하기 위해서 독립 전파 모델을 이용하는데, 독립 전파 모델의 수행을 위해서는 사용자들 간의 동화확률로서 실질적인 값을 부여하는 것이 필수적이다. 따라서 본 논문에서는 사용자의 재생산 파워의 개념과 이를 계량화하는 방법을 제안하고, 사용자의 콘텐츠 파워와 재생산 파워를 이용하여 사용자 간의 동화확률을 부여하는 방안을 제안한다. 끝으로, 실제 블로그 연결망에서 제안하는 기법과 기존의 기법을 이용하여 파워 유저들을 파악하는 실험을 수행하고, 결과를 비교 및 분석한다.

키워드 : 사회연결망 분석, 블로그, 데이터 마이닝, 정보파급, 파급 파워 유저

Abstract For business purposes, it is important to identify diffusion power users, a group of users who have big influence on other users in diffusing content. In this paper, we use the independent cascade model for determining diffusion power users, and to do so, we need a method for calculating the assimilation probability between users. This paper proposes the concepts of user delivery power and a way to quantifying the value of this. User delivery power is used to compute the assimilation probability with user content power. We analyze the proposed method by comparing its performance with those of existing methods through experiments using a real blog network data.

Key words : Social Network Analysis, Blog, Data Mining, Information Diffusion, Diffusion Power User

· 본 연구는 NHN(주)의 지원을 받았습니다. 그러나, 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자(들)의 것이며, 반드시 지원 회사의 입장을 대변하는 것은 아닙니다. 또한 본 연구는 2010년도 정부(교육과학기술부)의 재원에 의한 한국연구재단의 지원(Grant No. 2008-0061006)과 지식경제부 및 정보통신연구진흥원의 대학IT연구센터지원사업(IITA-2009-C1090-0902-0040), 2010년도 두뇌한국21사업의 부분적인 지원을 받아 수행되었습니다.

† 학생회원 : 한양대학교 전자컴퓨터통신공학과
shlim@agape.hanyang.ac.kr

** 종신회원 : 한양대학교 전자컴퓨터통신공학과 교수
wook@hanyang.ac.kr

*** 종신회원 : 연세대학교 경영대학 교수
boxenju@yonsei.ac.kr

논문접수 : 2009년 11월 4일

심사완료 : 2010년 1월 7일

Copyright©2010 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 자료의 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. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 컴퓨팅의 실제 및 래터 제16권 제4호(2010.4)

1. 서론

최근, 인터넷의 발달로 인해서 개인의 의사 표현과 사고를 위한 장이 오프라인에서 온라인으로 급속히 이동하고 있다. 따라서 온라인상에서 개인들이 서로 관계를 맺고 교류를 갖는 온라인 사회가 출현하였다. 이러한 온라인 사회를 대상으로 하는 서비스의 대표적인 예로서 블로그 서비스(blog service)를 들 수 있다[1-3]. 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.

블로그 연결망 데이터를 대상으로 하는 분석기법들 가운데 ‘파워유저 파악하기’가 있다[4,5]. 본 논문에서는 블로그 연결망 내의 일반 사용자들과 달리 분석가의 관심축면에서 높은 파워를 가진 사용자들을 파워 유저(power user)라고 정의한다. 이러한 사용자의 예로서 다른 사용

자들에게 콘텐츠를 파급시키는 능력이 큰 사용자를 들 수 있다. 본 논문에서는 이러한 사용자를 블로그 연결망에서의 파급 파워 유지(diffusion power user)라고 정의한다. 블로그 연결망에서 파급 파워 유저를 파악함으로써 이들을 중심으로 블로그 서비스 활성화를 위한 다양한 정책을 적용하여 비즈니스의 성공을 도모할 수 있을 것이다.

파급 파워 유저를 파악하기 위한 대표적인 방법으로서 독립 전파 모델(independent cascade model)을 이용한 방법을 들 수 있다[6]. 독립 전파 모델은 파급 파워 유저를 파악하기 위해서 사용자 간에 파급이 일어날 확률 값을 필요로 한다. 그러나 이 값으로서 실질적인 값을 부여하는 기법에 대한 연구는 미흡한 상태이다.

따라서 본 논문에서는 블로그 연결망에서 독립 전파 모델을 이용하여 파급 파워 유저를 파악하기 위해서 사용자 간에 파급이 발생할 확률로서 실질적인 값을 부여하는 방안에 대하여 논의한다. 이를 통해서 실제 응용에서 필요로 하는 파급 파워 유저들을 파악할 수 있다. 또한, 사용자 간에 파급이 발생할 확률을 이용하여 커뮤니티 식별, 사회연결망 클러스터링[7] 등의 그래프 마이닝 연산을 수행할 수 있다.

2. 콘텐츠 파워

본 장에서는 [8]에서 제안한 사용자의 콘텐츠 파워를 이용하여 파워유저를 선정하는 방안에 대하여 간략하게 설명한다.

표 1은 앞으로의 논의의 전개를 위해서 필요한 용어 및 기호들을 정리한 것이다. U_i 는 식별자가 i 인 사용자를 의미한다. D_i 는 U_i 가 소유한 게시글들의 집합을 의미하고, D_{ij} 는 U_i 의 j 번째 게시글을 의미한다. 게시글 D_{ij} 가 사용자들에게 미치는 콘텐츠 영향력을 이 게시글의 콘텐츠 파워(document content power)라고 정의하며, $DCP(D_{ij})$ 로 표기한다. 또한, 사용자 U_i 가 다른 사용자들에게 미치는 콘텐츠 영향력을 사용자의 콘텐츠 파워(user content power)라고 정의하고, $UCP(U_i)$ 로 표기한다.

만일 사용자 U_i 가 큰 콘텐츠 파워를 갖는다면, 사용자 U_i 는 양질의 콘텐츠를 소유하고 있다고 볼 수 있다. 이에 착안하여 본 논문에서는 사용자의 콘텐츠의 질의 척도로서 사용자의 콘텐츠 파워를 이용한다. 또한, 블로그 사용자들은 자신이 직접 게시글을 작성하는 것 외에도 다른 사용자들의 게시글들을 스크랩하거나 엮인글을 달 수 있는데, 본 논문에서는 사용자의 이러한 행동을 ‘콘텐츠 재생산’이라고 부른다. 사용자가 콘텐츠를 재생산하는데 적극적인 정도를 사용자의 재생산 파워(user delivery power)라고 정의하며, 사용자 U_i 의 재생산 파워를 $UDP(U_i)$ 로 나타낸다.

사용자가 블로그 서비스를 이용할 때, 취할 수 있는 액션으로는 게시글 작성(write), 조회하기(read), 덧글 남기기(comment), 스크랩하기(scrap), 엮인글 달기(link)의 다섯 가지가 있으며, 이러한 액션을 각각 W, R, C, S, L로 표기한다. 게시글의 콘텐츠 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_l 로 표기한다.

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

표 1 용어 정의

U_i : user i $D_i = \{D_{i,1}, D_{i,2}, \dots\}$: The set of documents owned by U_i D_{ij} : Document j of user i $DCP(D_{ij})$: Document Content Power of D_{ij} $UCP(U_i)$: 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 $ActionWeight = \{w_w, w_r, w_c, w_s, w_l\}$: Weights of actions

표 2 게시글의 콘텐츠 파워 계산

$DCP(D_{ij}) = w_D * D_DCP(D_{ij}) + w_I * I_DCP(D_{ij})$ $D_DCP(D_{ij}) = w_r * R_Count(D_{ij}) + w_c * C_Count(D_{ij}) + w_s * S_Count(D_{ij}) + w_l * L_Count(D_{ij})$ $I_DCP(D_{ij}) = w_D * \sum D_DCP(D_{r,j'}) + w_I * \sum I_DCP(D_{r,j'})$, where $D_{r,j'}$ represents documents reproduced from D_{ij}

사용자의 콘텐츠 파워는 사용자 블로그에 등록되어 있는 모든 게시글들의 콘텐츠 파워를 이용하여 계산한다. 게시글의 콘텐츠 파워는 블로그에 등록된 이후 노출된 시간에 비례하여 증가하는 경향이 있다. 따라서 오래 전에 등록된 게시글은 최근에 등록된 게시글에 비하여 실질적인 콘텐츠의 영향력은 작더라도 오랜 노출 시간 때문에 큰 콘텐츠 파워를 가지는 것으로 왜곡될 수 있다. 따라서 노출 시간의 역수(inverse of exposed time)

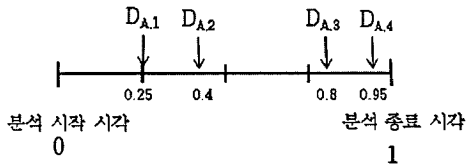
를 게시글의 콘텐츠 파워에 곱함으로써 노출 시간의 차이로 인한 왜곡을 보정한다. 식 (1)은 게시글의 콘텐츠 파워를 이용하여 사용자의 콘텐츠 파워를 계산하는 방법을 나타낸 것이다. 여기서, $IET_{D_i,j}$ 는 게시글 $D_{i,j}$ 의 상대적 노출 시간의 역수를 의미한다.

$$UCP(U_i) = \sum_j IET_{D_i,j} * DCP(D_{i,j}) \quad (1)$$

그림 1은 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.

3. 제안하는 기법

본 논문에서는 각 사용자 간의 동화확률을 계산하기 위해서 앞 장에서 언급한 사용자의 콘텐츠 파워와 사용자의 재생산 파워를 이용하는 방법을 제안한다.



Document	DCP	IET
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 1 사용자 콘텐츠 파워 계산의 예

3.1 사용자 재생산 파워

사용자가 콘텐츠를 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 워인 글 달기만을 고려하였다. 식 (2)는 사용자 U_i 의 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.

$$UDP(U_i) = w_s * S_Count(U_i) + w_l * L_Count(U_i) \quad (2)$$

3.2 동화확률 계산 기본 전략

동화확률 $P_{A \rightarrow B}$ 는 U_A 의 콘텐츠의 질을 의미하는 $UCP(U_A)$ 와 U_B 의 콘텐츠 재생산에 적극적인 정도를 의미하는 $UDP(U_B)$ 에 의하여 결정된다고 볼 수 있다. 이때, $P_{A \rightarrow B}$ 를 결정하는 데에 반영된 $UCP(U_A)$ 와 $UDP(U_B)$ 의 비율을 측정할 수 있다면, $P_{A \rightarrow B}$ 는 $UCP(U_A)$ 와 $UDP(U_B)$

에 각각의 반영비율을 곱하여 계산할 수 있을 것이다. 따라서 본 논문에서는 동화확률을 결정하는데 이용된 사용자 콘텐츠 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 $P_{A \rightarrow B}$ 를 계산하는 방법을 나타낸 것이다.

$$P_{A \rightarrow B} = w_{CP} * C_{PA} + w_{DP} * D_{PB} \quad (\text{단, } w_{CP} = 1 - w_{DP}) \quad (3)$$

여기서, $P_{A \rightarrow B}$ 는 확률값으로서 0에서 1사이의 값을 가지므로, $UCP(U_A)$ 와 $UDP(U_B)$ 를 각각 0에서 1사이로 정규화한 값인 C_{PA} 와 D_{PB} 로 변환하여 계산한다. 이를 위해서 전체 사용자들의 파워값의 분포영역을 동일한 사용자 수를 갖는 임의의 개수의 그룹으로 나누어 각 그룹 내의 사용자들이 동일한 파워 값을 갖도록 한다.

3.3 가중치 결정

이상적인 w_{CP} 와 w_{DP} 는 해당 w_{CP} 와 w_{DP} 를 이용하여 사용자 간의 동화확률을 부여하여 파급을 예측하였을 때, 가장 높은 정확도를 갖는 값이다. 파급 예측 결과의 정확도를 측정하는 방법은 3.4절에서 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[9]. 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.

그림 2는 라그랑지 보간법을 이용하여 가장 높은 정확도를 갖는 w_{CP} 와 w_{DP} 를 선택하는 방법을 나타낸 것이다. 위의 그래프는 w_{CP} 와 w_{DP} 의 분포를 나타낸다. x축은 w_{CP} 를 나타내고, y축은 w_{DP} 를 나타낸다. $w_{CP} = 1 - w_{DP}$ 이므로 (w_{CP}, w_{DP})는 대각선의 분포를 가진다. 이 예에서는 정확도의 변화를 분석하기 위해서 4개의 w_{CP} 의 값들인 $w_{CP1}, w_{CP2}, w_{CP3}, w_{CP4}$ 를 임의로 선정하였다. 아래의 그래프는 $w_{CP1}, w_{CP2}, w_{CP3}, w_{CP4}$ 를 통해서 얻어진 파급 예측 정확도를 이용하여 라그랑지 보간법으로 w_{CP} 의 변화에 따른 파급 예측 정확도의 변화를 예상한 것이다. x축은 w_{CP} 를 나타내고, y축은 w_{CP} 에 따른 파급 예측 정확도를 나타낸다. 이 예에서는 분석 결과, w_{CP} 가 w_{CP5} 일 때 가장 높은 파급 예측 정확도를 가지는 것으로 나타났다. 따라서 식 (3)의 w_{CP} 로서 w_{CP5} 를 선택하고, w_{DP} 로서 $1 - w_{CP5}$ 를 선택한다.

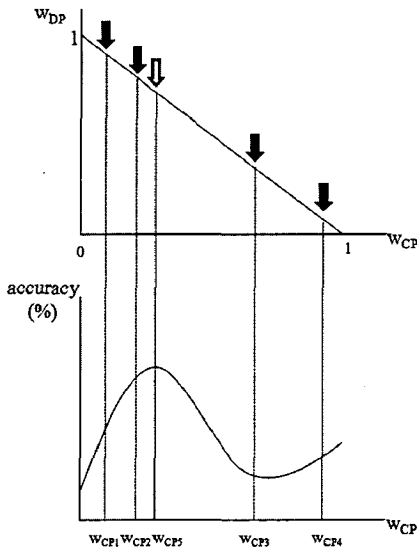


그림 2 WCP와 WDP 선택의 예

3.4 파급 예측 정확도 측정

본 논문에서는 파급 예측 정확도를 계산하기 위해서 특정 사용자가 게시글을 작성하였을 때, 이 게시글에 의하여 실제로 동화된 사용자들과 독립 전파 모델을 통해서 이 사용자가 동화시키리라고 예측된 사용자들을 비교하는 방법을 이용한다. 그림 3은 파급 예측 정확도를 측정하는 예를 나타낸 것이다. (a)는 실제로 파급된 기록(diffusion history)을 나타낸다. (b)는 (a)내에 존재하는 사용자들 간의 파급 관계들을 종합하여 블로그 연결망을 구축한 것이다. 또한, (c)는 특정 사용자가 게시글을 작성한 경우, 독립 전파 모델을 이용해서 동화될 사용자들을 예측한 것이다. 이 예에서 ICM(A)={B, E}는 UA가 게시글을 작성하였을 때, 독립 전파 모델을 통해서 UB, UE가 동화되리라고 예측되었음을 의미한다. 그러나 UA가 게시글을 작성한 경우, 실제로 이 게시글을 파급한 사용자들은 DH#1, DH#3를 통해서 UB, UC, UD, UG임을 알 수 있다.

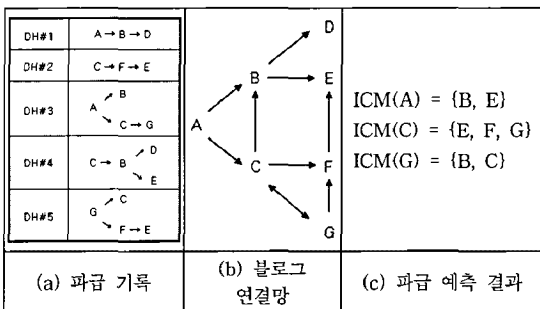


그림 3 분석 정확도 계산의 예

본 논문에서는 정보검색분야에서 사용되는 응답도(recall)와 정밀도(precision)의 개념을 이용하여 이들의 평균을 ICM(A)의 정확도로 사용한다. ICM(A)의 응답도는 $|B|/(B, C, D, G)=1/4$ 이고, 정밀도는 $|B|/(B, E)=1/2$ 이다. 따라서, ICM(A)의 정확도는 $(1/4+1/2)/2=3/8(=0.37)$ 이 된다. 동일한 방법으로 ICM(C)와 ICM(G)의 정확도를 측정하면, 각각 $7/12(=0.58)$, $5/12(0.41)$ 이 된다. WCP와 WDP에 따른 전체 파급 예측 정확도는 각 사용자의 ICM의 정확도의 합을 이용한다. 이 때, 각 사용자가 작성한 게시글이 실제로 파급된 횟수를 각 ICM의 정확도에 반영하여 가중치 평균을 계산한다. ICM(A)의 반영 비율은 사용자 UA에 의해서 실제로 파급된 횟수 즉, $|DH#1|+|DH#3|=2+3=5$ 이고, ICM(C)의 반영 비율은 $|DH#2|+|DH#4|=2+3=5$, ICM(G)의 반영 비율은 $|DH#5|=3$ 이 된다. 따라서, WCP와 WDP에 따른 파급 결과의 전체 정확도는 $(0.37*5+0.58*5+0.41*3)/(5+5+3) = 5.98/13=0.46$ 이 된다.

4. 성능 평가

4.1 실험 환경

본 연구에서는 성능 분석을 위하여 한국의 블로그 포털 사이트의 2006년 7월부터 약 7개월간 수집한 데이터를 사용하였다. 분석 기간 중에 생성된 게시글의 수는 약 100,000,000개이며, 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 이웃 관계가 설정되어 있어도 빈번한 교류를 갖지 않는 경우가 많다는 사실에 근거한 것이다.

본 실험에서는 비교 대상으로 선정된 파워 유저 선정 기법들은 다음과 같다. 사용자 간의 서로이웃 관계를 이용하여 연결 중앙성 기법인 (DEG), 게시글들의 댓글의 개수를 이용하는 기법인 Comment_Direct(C_D), 간접적으로 생산한 게시글들의 댓글의 개수를 이용하는 기법인 Comment_Indirect(C_I), 직접적으로 생산한 게시글들과 간접적으로 생산한 게시글들의 댓글의 개수를 이용하는 기법인 Comment_Total(C_T), 마찬가지로 스크랩 및 엮인글의 개수를 이용한 기법인 (SL_D), (SL_I), (SL_T), 컨텐츠 파워를 이용한 기법인 (CP_D), (CP_I), (CP_T), 사용자의 재생산 파워를 이용한 기법인 (DP). 사용자의 파급 파워를 이용한 기법은 독립 전파 모델의 수행을 위해서 필요로 하는 사용자 간의 파급확률을 부여하는 정책에 따라 나뉘며, 모든 관계의 파급확률로서 1%를 부여한 (CST_1), 모든 관계의 파급확률로서 10%를 부여한 (CST_10)[8], 본 논문에서 제안하는 기법을 이용하여 사용자 간의 파급확률을 부여하는 기법인 (DFP)

를 사용하였다.

4.2 결과 분석

본 논문에서는 시간의 흐름에 따라 각 기법을 이용하여 선정한 파워유저 집합의 생존률 변화를 비교하였다. 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 5는 실험의 결과를 나타낸 것으로서, 지면의 제약으로 인해 기법(DEG), 기법(C_T), 기법(SL_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.

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 재생산 파워를 계량화하는 방안을 제안하였고, 사용자의 재생산 파워와 사용자의 콘텐츠 파워를 이용하여 사용자 간의 동화확률로서 실질적인 값을 부여하는 방안을 제안하였다. 또한, 계산된 사용자 간의 동화확률의 정확도를 측정하는 기법을 제안하고, 이를 이용하여 가장 높은 정확도를 갖는 값을 사용자 간의 동화확률로서 결정하는 기법을 제안하였다. 제안하는 기법과 기존의 기법들을 이용하여 실제 블로그 연결망에서 파워 유저들을 파악하는 실험을 수행하고, 이 결과를 비교 및 분석하였다.

참고 문헌

- [1] Blogger.com Co., Ltd. <http://blogger.com>.
- [2] MySpace.com Co., Ltd. <http://myspace.com>.
- [3] NHN Co., Ltd. <http://blog.naver.com>.
- [4] D. Gruhl, et al., "Information Diffusion Through Blogspace," *WWW*, pp.491-501, 2004.
- [5] S. Nakajima, et al., "Discovering Important Bloggers Based on a Blog Thread Analysis," *WWW 2nd Annual Workshop on the Weblogging Ecosystem*, 2005.
- [6] J. Goldenberg, et al., Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211-223, 2001.
- [7] D. Cai et al., "Mining Hidden Community in Heterogeneous Social Networks," In *Proc. Int'l Workshop on Link Discovery: Issues, Approaches and Applications*, pp.58-65, 2005.
- [8] S. H. Lim, S. W. Kim, S. Park and J. H. Lee, "Determining Content Power Users in a Blog Network," In *Proc. 3rd ACM KDD Workshop on Social Network Mining and Analysis(SNA-KDD 2009)*, Paris, France, June. 28-July. 1, 2009.
- [9] E. Waring, *Philosophical Transactions of the Royal Society*, 69, 59-67, 1979.

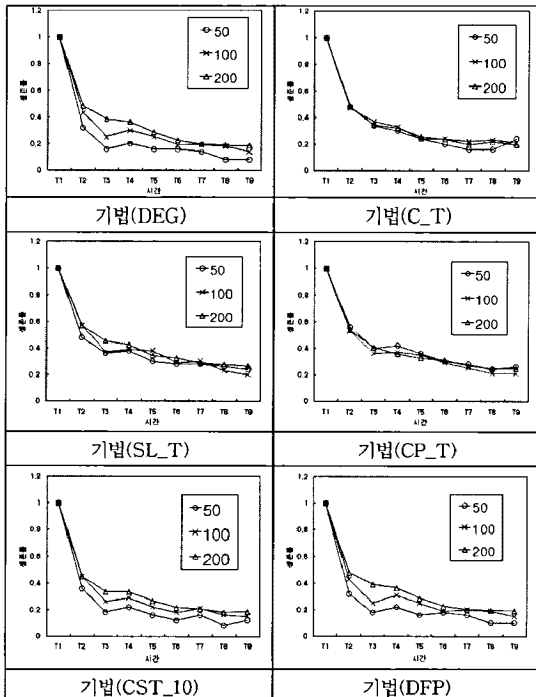


그림 4 각 기법의 파워유저 집합의 생존률 변화



임 승 환

2003년 한양대학교 전자컴퓨터공학부 졸업(학사). 2005년 한양대학교 정보통신대학원 졸업(공학석사). 2005년~현재 한양대학교 대학원 전자통신컴퓨터공학과 박사과정 재학중. 관심분야는 데이터베이스, 데이터 마이닝, 사회 연결망 분석



김 상 옥

1989년 2월 서울대학교 컴퓨터공학과(학사). 1991년 2월 한국과학기술원 전산학과(석사). 1994년 2월 한국과학기술원 전산학사(박사). 1991년 7월~1991년 8월 미국 Stanford University, Computer Science Department, 방문 연구원. 1994년 3월~1995년 2월 KAIST 정보전자연구소 전문 연구원. 1999년 8월~2000년 8월 미국 IBM T.J. Watson Research Center, Post-Doc. 1995년 3월~2003년 2월 강원대학교 정보통신공학과 부교수. 2003년 3월~현재 한양대학교 정보통신대학 정보통신학부 교수. 2009년 1월~현재 미국 Carnegie Mellon University, Visiting Scholar. 관심분야는 데이터베이스 시스템, 저장 시스템, 트랜잭션 관리, 데이터 마이닝, 멀티미디어 정보 검색, 공간 데이터베이스/GIS, 주기억장치 데이터베이스, 이동 객체 데이터베이스/텔레메틱스, 사회 연결망 분석, 웹 데이터 분석



박 선 주

1989년 서울대학교 컴퓨터공학과(학사)
1991년 서울대학교 컴퓨터공학과(석사)
1999년 U of Michigan, Ann Arbor, CSE(박사). 1999년~2005년 Rutgers University, MSIS Department(조교수)
2005년~현재 연세대학교 경영학과(부교수). 관심분야는 에이전트 시스템, 옵션, 온라인 사회연결망, 네트워크 가격정책