정보 파급 모델링을 위한 블로그 네트워크 구성

(Construction of a Blog Network based on Information Diffusion)

임 승 환 † 김 상 욱 ††
(Seung-Hwan Lim) (Sang-Wook Kim)
강 규 황 ††† 도 영 주 †
(Kyu-Hwang Kang) (Young-Joo Do)

요 약 독립 전파 모델은 블로그 월드 내에서 정보가 파급되는 현상을 분석하기 위해서 널리 이용되는 모델이다. 본 논문에서는 블로그 월드에서의 정보 파급 분석에 독립 전파 모델을 적용하기 위하여 블로그 네트워크를 구성하는 방법을 제안한다. 블로그 네트워크의 구성을 위하여 제안된 방법은 사용자 간의 액션 이력을 분석하여 두 사용자 간의 관계를 설정하고, 두 사용자 간의 파급 확률을 계산한다. 사용자간의 파급 확률을 계산하기 위해서 사용자가 정보의 파급을 의도하고 작성한 게시글들 중에서 실제로 특정 사용자에게 파급된 게시글들의 비율을 이용한다. 실제 블로그 월드의 데이터를 이용하여 정보의 파급 현상을 분석한 결과, 제안하는 기법이 기존의 기법에 비해서 정보의 파급 현상을 충실하게 반영하고 있는 것으로 나타났다.

키워드 : 사회연결망 분석, 블로그, 데이터 마이닝, 정보 파급, 정보 파급 모델

Abstract The independent cascade model has been widely used to analyze information diffusion in the blog world. In this paper, we propose a new method to construct a blog network for applying the independent cascade model to analyzing of information diffusion in a blog world. To construct a blog network, the proposed method establishes the edge between two users and calculates diffusion probabilities between them by analyzing the activities happened between two users. To calculate diffusion probabilities, the method exploits the ratio of the number of documents actually diffused to a specific user to that of documents written for the purpose of being diffused to other blogs. The experimental result using a real world blog data demonstrates that our method reflects actual information diffusion in a blog world better than existing ones.

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

1. 서 론

온라인 사회 연결망을 이용한 대표적인 서비스로서 블로그 서비스(blog service)를 들 수 있다. 블로그(blog)는 블로그 서비스의 사용자가 자신의 생각을 온라인상에 게시할 수 있는 일종의 개인 웹사이트이다[1-4]. 사용자들은 서로 영향을 주고 받으면서, 다양한 활동을 할 수 있는데, 본 논문에서는 이렇게 형성된 블로그 사용자들 간의 온라인 사회를 블로그 월드(blog world)라고 부른다. 또한, 블로그 사용자들은 블로그 월드 내에서 서로 다양한 관계들을 설정할 수 있는데, 이를 통해서 형성된 온라인 사회 연결망을 블로그 네트워크(blog network)라고 부른다.

블로그 서비스 업체들에서는 블로그 사용자가 다른 사용자의 게시글에 관심이 있는 경우, 이에 관련된 게시글을 작성하거나, 해당 게시글을 복사해올 수 있는 기능을 제공하고 있는데, 이러한 기능을 각각 엮인글(trackback), 스크랩(scrap) 이라고 부른다. 블로그 월드 내에서는 엮인글과 스크랩을 통해서 블로그들 간에 게시글의 파급이 빈번하게 발생한다. 블로그 월드에서 발생하는 이러한 정보의 파급을 분석하는 것은 정보 파급의 예측, 이상 정보의 검출, 마케팅에의 응용, 블로그 네트워크의 활성화 등에 이용될 수 있는 매우 유용한 연구 이슈이다[5-7].

블로그 월드 내에서 이루어지고 있는 정보의 파급 현상을 분석하기 위해서, 독립 전파 모델(independent cascade model)을 이용한다[6]. 이를 위해서 블로그 월드 내에 존재하는 블로거들 간의 파급 관계들을 독립 전파 모델에 적용 가능한 블로그 네트워크의 형태로 나타내야 한다. 정보의 파급을 기반으로 하는 블로그 네트

· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학IT연구센터지원사업(IITA-2009-C1090-0902-0040)과 한국과학재단의 2009년도 특정기초연구사업(No. R01-2008-000-20872-0)의 부분적인 지원을 받아 수행되었습니다.
· 이 논문은 2009 한국컴퓨터종합학술대회에서 '정보 파급을 기반으로 하는 블로그 연결망의 구성'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 한양대학교 전자컴퓨터통신공학과
shlim@agape.hanyang.ac.kr
trinity@agape.hanyang.ac.kr
†† 종신회원 : 한양대학교 전자컴퓨터통신공학과 교수
wook@hanyang.ac.kr
††† 학생회원 : 한양대학교 전자컴퓨터통신공학과
kkhfiles@hanyang.ac.kr

논문접수 : 2009년 8월 13일
심사완료 : 2009년 9월 30일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제15권 제11호(2009.11)

워크에서 노드는 사용자로, 간선은 사용자들 간의 관계로 타나낼 수 있다. 또한, 독립 전파 모델에서는 간선마다 확률 값을 갖는데, 이는 블로그 네트워크에서 사용자들 간의 파급 확률로 나타낼 수 있다. 따라서, 독립 전파 모델을 이용하여, 블로그 월드에서 정보의 파급 현상을 분석하기 위해서는 사용자들 간에 정보의 파급이 발생할 확률 값을 계산하는 것이 필수적이다. 정보 파급의 분석 정확도를 높이기 위해서 이 값 역시 두 블로거 간의 파급 관계를 잘 표현하고 있어야 한다. 그러나 이 값으로서 실질적인 값을 부여하는 기법에 대한 연구의 성과들은 미흡한 상태이다.

따라서 본 논문에서는 블로그 월드 내에서 발생하는 정보의 파급 현상을 독립 전파 모델을 이용하여 분석 가능하도록, 정보의 파급 현상을 반영하는 블로그 네트워크를 구축하는 방안에 대하여 논의한다. 이를 위해서 블로그 월드에서 발생한 정보의 파급 기록을 토대로 사용자 간의 파급 확률을 부여하는 방안을 제안한다.

2. 관련 연구

사회 연결망에서 파급 현상을 분석하기 위한 기존의 연구들의 기본 아이디어는 다음과 같다. 연결망의 노드들은 상호간에 영향을 주고 받을 수 있으며, 이 영향에 의해서 특정 노드의 성향이 영향을 준 노드의 성향과 같아질 수 있다. 본 논문에서는 이러한 경우에 이 노드는 영향을 준 노드에 의해서 동화되었다(assimilate)고 부른다.

참고문헌 [8]에서는 선형 임계값 모델을 제안하였다. 선형 임계값 모델은 각 노드에 임계값을 부여하고, 노드 간의 관계에 가중치를 부여하여 특정 노드가 주변 노드들로부터 받은 영향의 정도(가중치)를 누적한 값이 해당 노드가 갖고 있는 임계값 이상이면, 이 노드는 영향을 미친 노드들에 의하여 동화된 것으로 간주한다. 그러나 실제 블로그 월드 내에서 게시글의 파급은 사용자간의 독립적인 관계에 의해서 이루어지는데 반해, 선형 임계값 모델은 여러 사용자들의 영향의 합에 의한 파급을 설명하기 위한 모델이므로 블로그 네트워크에 적용하기에는 적절하지 않다.

참고문헌 [6]에서는 독립 전파 모델을 제안하였다. 독립 전파 모델은 노드간의 간선에 확률을 부여하여 노드 간에 영향을 미칠 때, 이 확률에 의하여 동화 여부를 결정한다. 본 논문에서는 이 값을 노드간의 동화확률이라고 부른다. 블로그 월드에서 특정 사용자가 임의의 게시글을 파급하는 것은 자신의 이웃 사용자들에게 영향을 받아서가 아니라, 해당 게시글을 소유하고 있는 사용자에게만 영향을 받아서 이루어진 것이다. 따라서 독립 전파 모델은 블로그 월드에서 발생하는 정보의 파급 현상을 분석하기 위한 타당한 모델이라고 할 수 있다.

참고문헌 [9]에서는 일반화된 전파 모델(general cascade model)을 제안하였다. 일반화된 전파 모델은 독립 전파 모델에서 특정 노드를 동화시키기 위해서 이웃의 노드들이 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 월드에서 정보의 파급 현상을 분석하기 위해서 독립 전파 모델을 이용한다. 독립 전파 모델을 이용하여 분석을 수행하기 위해서는, 블로그 월드로 부터 블로그 네트워크를 도출하는 과정이 선행되어야 한다. 또한, 독립 전파 모델을 만족하는 블로그 네트워크의 구성을 위해서는 사용자 간의 관계에 동화 확률을 필요로 한다. 정보 파급의 분석 정확도를 높이기 위해서 이 값 역시 두 블로거 간의 파급 관계를 잘 표현하고 있어야 한다. 그러나 기존의 연구들에서는 주로 연결망 내에서의 파급 관계를 설명할 수 있는 모델을 제안하는 데에 초점을 맞추고 있었으므로, 사용자간의 동화확률로서 실질적인 값을 부여하기 위한 방안에 대한 연구는 미흡한 상태이다. 따라서 본 논문에서는 블로그 월드에서 발생하는 정보 파급 현상을 정확하게 분석할 수 있도록, 사용자간의 동화확률로서 실질적인 값을 부여하기 위한 방안에 대하여 논의한다.

3. 제안하는 방법

3.1 용어 정리

앞으로의 논의 전개를 위해서 필요한 용어 및 기호들을 정리하면 다음과 같다. U_A는 식별자가 A인 사용자를 의미한다. D_A는 U_A가 소유한 게시글들의 집합을 의미하고, $D_{A,i}$는 U_A의 i번째 게시글을 의미한다. $D_{A\rightarrow B}$는 D_A 중에서 U_B에 의해서 파급된 게시글들의 집합을 의미하고, $P_{A\rightarrow B}$는 D_A가 U_B에게 파급될 확률을 의미한다. 사용자 간의 파급 확률을 계산하기 위해서 게시글의 파급 유효도를 이용하고, 이를 score($D_{A,i}$)로 표기한다. 게시글의 파급 유효도를 계산하기 위해서 사용자가 블로그 월드 내에서 취할 수 있는 액션들을 이용하는데, 사용자 액션의 종류로는 게시글 작성(write), 조회하기(read), 덧글 남기기(comment), 엮인글 달기(trackback), 스크랩 하기(scrap)의 다섯 가지가 있으며, 이러한 액션을 각각 W, R, C, T, S로 표기한다. 또한, 각각의 액션에 중요도를 부여하기 위하여 가중치를 할당할 수 있다. 액션 W, R, C, T, S를 위한 가중치는 W_W, W_R, W_C, W_T, W_S로 표기한다.

본 논문에서는 사용자 U_A의 게시글이 사용자 U_B에게 파급될 확률 $P_{A\rightarrow B}$가 D_A중에서 U_B가 관심을 갖고 엮인

글이나 스크랩을 통해서 파급해간 게시글들의 비율과 관련되어 있다는 점에 착안하였다. 이를 통해서 도출된 사용자 간의 파급 확률 부여 방법의 기본 아이디어는 식 (1)과 같다.

$$P_{A\to B}=\frac{|D_{A\to B}|}{|D_A|} \quad (1)$$

식 (1)에서 $|D_A|$와 $|D_{A\to B}|$는 각각 D_A와 $D_{A\to B}$의 개수를 의미한다. 따라서, 식)1)에서 $P_{A\to B}$는 D_A 중에서 U_B가 관심을 갖고 파급해간 문서들 $D_{A\to B}$의 비율을 의미한다. 본 논문에서는 이후의 전개를 통해서 식 (1)의 정확도를 개선하기 위한 방안에 대해서 논의한다.

3.2 유효 게시글을 이용한 확률 부여 방법

블로그 월드의 사용자들이 게시글을 작성하는 목적은 첫째, 자신의 블로그를 방문하는 다른 사용자들에게 정보를 제공하기 위해서 둘째, 정보제공과는 상관없이 자신의 감정이나 상황을 기록으로 남기기 위해서 이렇게 두 가지로 구분할 수 있다.

따라서 본 절에서는 사용자가 다른 사용자들에게 정보를 제공하기 위한 목적으로 작성한 게시글 즉, 다른 사용자들에게 파급되기를 기대하고 작성한 게시글들이 파급될 확률을 계산하는 방안을 제안한다. 이를 위해서는 사용자가 소유하고 있는 각 게시글에 대해서 작성된 의도를 파악할 수 있어야 한다. 각 게시글의 작성 의도는 각 사용자들에게 직접 질의를 통해서 파악할 수 있지만, 모든 사용자에게 소유하고 있는 게시글들에 대한 작성 의도를 질의하는 것은 현실적으로 불가능 하다. 따라서 본 논문에서는 각 게시글이 다른 사용자들로부터 유발한 액션들을 계량화함으로써 작성 의도를 추정하고자 한다. 이는 파급을 의도하고 작성된 게시글은 그렇지 않은 게시글에 비해서 다른 사용자들에게 많은 액션들을 유발한다는 사실에 기인한 것이다.

본 논문에서는 게시글이 파급을 의도하고 작성된 정도를 파급 유효도라고 정의하고, 이를 계산하기 위해서 게시글이 다른 사용자들에게 유발한 액션들의 데이터를 이용한다. 식 (2)는 파급 유효도를 계산하는 방법을 나타낸 것이다. score($D_{A,i}$)는 문서 $D_{A,i}$의 파급 유효도를 의미한다. 이는 다른 사용자들이 $D_{A,i}$에 대해서 보인 조회하기, 덧글 남기기, 엮인글 달기, 스크랩 하기 액션의 횟수와 각 액션의 가중치의 곱을 더해서 계산한다.

$$\begin{aligned} score(D_{A,i}) &= W_R{}^{*}R\_Count(D_{A,i})+ \\ &W_C{}^{*}C\_Count(D_{A,i})+W_T{}^{*}T\_Count(D_{A,i}) \\ &+W_S{}^{*}S\_Count(D_{A,i}) \end{aligned} \quad (2)$$

사용자 U_A가 소유하고 있는 게시글들 D_A 중에서 U_A가 파급의 의도를 갖고 작성한 게시글들의 집합은 D_A 중에서 파급 유효도가 임계값 θ 이상을 만족하는 게시글들만 선정함으로써 구할 수 있다. 이를 이용하여 식 (1)을 개선하면 식 (3)과 같다. 식 (3)에서 $P_{A\to B}$는 U_A가 파급 의도를 갖고 작성된 게시글들 중에서 U_B에게 파급된 게시글들의 비율을 의미한다. $D_A{}^{*}$는 D_A 중에서 파급 의도를 갖고 작성된 게시글들의 집합을 나타낸다.

$$P_{A\to B}=\frac{|D_{A\to B}|}{|D_A{}^{*}|} \quad (3)$$

$$\text{단, } (score(D_{A,i})\geq\theta, D_{A,i}\in D_A)\Rightarrow D_{A,i}\in D_A{}^{*}$$

이 방법에서는 $D_A{}^{*}$에 포함되지 않는 게시글이 $D_{A\to B}$에 포함될 수 있다. 따라서 파급을 의도하고 작성된 게시글에 한해서만 U_B에게 정상적으로 파급된 것으로 간주해야 한다고 생각할 수 있다. 그러나 이것은 U_A가 파급을 의도하고 작성한 게시글임에도 불구하고 파급 유효도가 낮게 측정되어서 $D_A{}^{*}$에 포함되지 못했기 때문일 수 있고, U_A가 파급을 의도하지 않고 작성한 게시글임에도 불구하고 U_B가 관심을 갖고 파급을 했기 때문일 수도 있다. 후자의 경우에는 U_A가 파급을 의도하지 않고 작성된 게시글임에도 불구하고 U_B가 파급을 한다는, 두 사용자 간의 독특한 특성이 $P_{A\to B}$에 반영되어 있다고 볼 수 있다. 따라서 본 논문에서는 $P_{A\to B}$를 계산하기 위한 분자로서 식 (1)에서와 같이 $|D_{A\to B}|$를 수정 없이 이용한다.

이러한 이유로 인해서 분자 $|D_{A\to B}|$가 분모 $|D_A{}^{*}|$보다 큰 값을 가질 수 있으므로, $P_{A\to B}$가 1보다 큰 값을 확률로서 가질 수 있다. 이 경우에 $P_{A\to B}$의 값으로 최대값인 1을 부여하여, 값의 범위를 조절한다.

3.3 유효 점수를 이용한 확률 부여 방법

유효 게시글을 이용한 확률 부여 방법은 파급 유효도가 임계값 이상인 게시글들을 모두 동일하게 취급하여 게시글들의 개수의 비율을 확률 값으로 이용하였다. 그러나 본 절에서 제안하는 유효 점수를 이용한 방법은 확률의 계산을 위해서 게시글들을 동일하게 취급하지 않고, 각 게시글이 갖고 있는 파급 유효도 만큼 비중을 두어 처리한다.

이 방법은 D_A에 속한 모든 게시글들의 파급 유효도의 총합과, $D_{A\to B}$에 속한 모든 게시글들의 파급 유효도의 총합의 비율을 확률 값으로 이용한다. 이는 U_A가 파급을 의도한 전체 규모에서 실제로 U_B에게 파급된 비율을 의미한다.

각 게시글들이 갖는 파급 유효도의 값은 매우 다양하다. 또한, 많은 수의 사용자에게 파급된 게시글의 파급 유효도는 일반적인 게시글의 파급 유효도에 비해서 매우 큰 값을 갖는다. 따라서 확률 계산 단계에서 큰 파급 유효도를 갖는 게시글들이 전체 파급 유효도의 합의 대부분을 차지하게 된다. 이는 확률을 계산하는 데에 일반적인 파급 유효도를 갖는 다수의 게시글들이 무시된다는 것을 의미한다. 따라서 본 논문에서는 게시글들의 파급 유효도를 다음과 같이 정규화하여 이용한다. 먼저,

각 게시글의 파급 유효도를 계산하고, 이 값을 기준으로 상위 n% 이상의 게시글들에 대해서 최대값인 1을 부여한다. 상위 n% 중, 가장 작은 파급 유효도 값으로 나머지 게시글들의 파급 유효도를 나누어 0에서 1사이의 값을 갖도록 값의 범위를 조정한다. 그림 1은 파급 유효도 값을 정규화하는 예를 나타낸다.

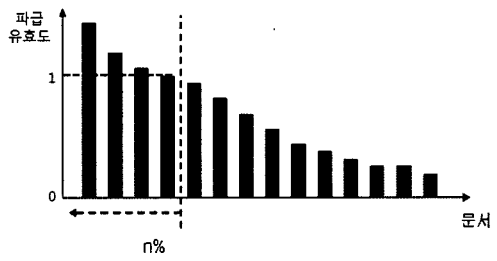


그림 1 파급 유효도 정규화의 예

유효점수를 이용하여 확률 부여하기 위해서 D_A의 모든 게시글들의 파급 유효도를 계산하여, 이를 정규화한 값들의 총합을 계산한다. 이 값은 U_A가 파급을 의도한 전체 규모라고 볼 수 있다. 이 중에서 실제로 U_B에게 파급된 비율을 계산하기 위해서, D_A에서와 마찬가지 방법으로 $D_{A\rightarrow B}$에 포함된 모든 게시글들의 파급 유효도 값들을 정규화하여 합을 구할 수 있다. 그러나 $D_{A\rightarrow B}$에 포함된 모든 게시글들은 이미 D_A에서 D_B로 파급되기 위한 충분한 조건을 만족한 것으로 볼 수 있다. 따라서 $D_{A\rightarrow B}$의 모든 게시글들에 정규화된 파급 유효도의 최대값인 1을 부여한다. 식 (4)는 유효 점수를 이용하여 확률을 부여하는 방법을 나타낸 것이다. norm_score($D_{A,i}$)는 $D_{A,i}$의 파급 유효도 값을 정규화한 값을 의미한다. $|D_{A\rightarrow B}|$는 $D_{A\rightarrow B}$의 모든 게시글들에 정규화된 파급 유효도의 값으로 1을 부여한 값을 의미한다.

$$P_{A\rightarrow B} = \frac{|D_{A\rightarrow B}|}{\sum_{D_{A,i}\in D_A} norm\_score(D_{A,i})} \quad (4)$$

이 방법에서 분자 $|D_{A\rightarrow B}|$가 분모인 D_A의 파급 유효도를 정규화한 값의 총합보다 큰 값을 가질 수 있으므로, $P_{A\rightarrow B}$가 1보다 큰 값을 확률로서 가질 수 있다. 이 경우에 유효 문서를 이용하여 확률을 부여하는 방법과 마찬가지로 $P_{A\rightarrow B}$의 값으로 최대값인 1을 부여하여, 값의 범위를 조절한다.

5. 성능 평가

5.1 실험 환경

본 논문에서는 성능 분석을 위하여 실제 블로그 월드에서 수집한 데이터를 사용하였다. 성능 분석의 비교 대상으로 선정한 블로그 네트워크 구축 기법들은 본 논문에서 제안한 유효 게시글을 이용하여 확률을 부여하는 기법 ED, 본 논문에서 제안한 유효 점수를 이용하여 확률을 부여하는 기법 ES, 모든 간선에 동일하게 1%의 확률을 부여한 기법 CST1[9], 모든 간선에 동일하게 5%의 확률을 부여한 기법 CST5의 총 네가지 이다.

본 논문에서는 위의 기법들의 성능을 측정하기 위해서, 블로그 월드에서 실제로 정보가 파급된 기록들과 각 기법을 통해서 구성한 블로그 네트워크들을 대상으로 독립 전파 모델을 수행하여 생성된 파급 기록을 비교하였다. 비교의 척도로는 정보 검색 분야에서 널리 사용되는 응답도(recall)와 정밀도(precision)를 사용하였다[10].

5.2 결과 분석

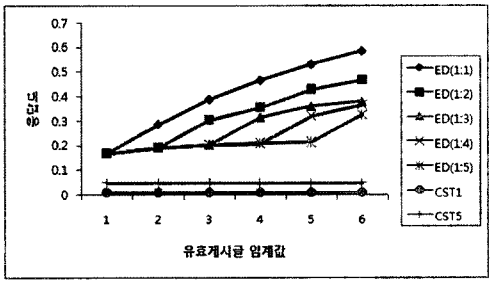
기법 ED와 기법 ES를 이용하여 정보의 파급을 분석하기 위해서는 사전에 분석가가 분석의 목적에 맞도록 기법 ED와 기법 ES의 수행에 필요한 인자들을 설정하여야 한다. 이에 본 논문에서는 기법 ED와 기법 ES의 인자 값들을 변경하면서 성능을 측정하고, 이를 기법 CST1, 기법 CST5와 성능을 비교하는 실험을 수행한다.

실험 1에서는 기법 ED의 액션 가중치 덧글:(엮인글, 스크랩)의 비율을 1:1, 1:2, 1:3, 1:4, 1:5로 변경하면서 분석을 수행하였다. 엮인글과 스크랩 액션을 함께 취급한 것은 이들이 재생산 액션으로서 정보의 파급에 유사한 의미를 갖기 때문이다. 또한, 기법 ED의 유효 문서의 판별 기준이 되는 임계값은 1부터 시작하여 1씩 증가한 값을 갖도록 설정하였다. 그림 2는 실험 1의 결과를 나타낸다. 그림 2(a), 그림 2(b)의 x축은 유효 게시글의 판별 기준이 되는 임계값을 의미하고, y축은 각각 응답도와 정밀도를 의미한다.

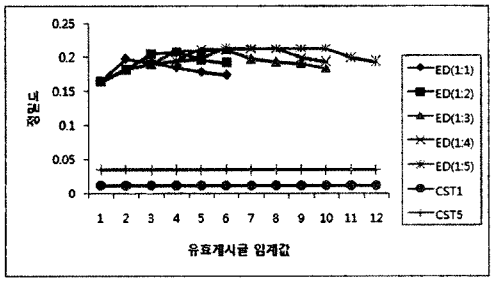
인자 값들의 변화에 따른 기법 ED의 응답도 변화를 살펴본 결과, 그림 2(a)와 같이 덧글:(엮인글,스크랩) 액션의 비율이 감소하고, 임계값이 증가할수록 응답도가 증가하는 것으로 나타났다. 이러한 결과는 다음과 같은 이유에 기인한다. 덧글:(엮인글,스크랩) 액션의 비율이 감소하고, 임계값이 증가할수록 기법 ED에 의해서 유효 게시글로 판정 받는 게시글의 수가 줄어들게 된다. 이로 인해 간선에 부여되는 파급 확률은 큰 값을 갖게 되고, 블로그 네트워크 내의 많은 사용자들을 동화시키게 된다. 따라서 실제 파급 기록에 존재하는 사용자들을 포함할 가능성이 증가하게 되므로, 높은 응답도를 나타낸다.

인자 값들의 변화에 따른 기법 ED의 정밀도 변화를 살펴본 결과, 그림 2(b)와 같이 덧글:(엮인글,스크랩) 액션 비율의 증가에 따라서 정밀도도 증가하는 경향을 보였다. 또한, 임계값의 증가에 따라서 정밀도는 증가하다가 감소하는 경향을 보이는데, 덧글:(엮인글,스크랩) 액션의 비율이 클수록 정밀도의 증가 및 감소하는 경사가

완만해지는 것으로 나타났다. 이는 덧글:(엮인글,스크랩) 액션의 비율이 커질수록 정밀도의 변화에 미치는 임계값의 영향력이 줄어들게 됨을 의미한다. 이 실험을 통해서 분석가가 액션의 가중치들과 임계값을 결정할 때, 정보 파급 분석의 정밀도를 높이기 위해서는 액션의 가중치들과 임계값을 독립적으로 결정하는 대신, 함께 고려하여 결정해야 함을 알 수 있다.



(a) 응답도



(b) 정밀도

그림 2 액션의 가중치, 임계값 변화에 따른 기법 ED의 성능 변화

실험 2에서는 기법 ES의 액션의 비율을 1:1, 1:2, 1:3, 1:4, 1:5로 변경하면서 분석을 수행하였다. 또한, 게시글의 파급 유효도 값을 정규화하기 위한 최대값 설정 비율 즉, 최대값을 부여받는 상위 n%는 1%, 5%, 10%로 설정하였다. 실험 결과, 기법 ES의 응답도, 정밀도는 액션 가중치의 비율, 최대값 설정 변화에 거의 영향을 받지 않는 것으로 나타났다.

실험 1과 실험 2의 결과, 기법 ED, 기법 ES가 기법 CST1, 기법 CST5에 비해서 높은 응답도와 정밀도를 보였다. 또한, 기법 ES는 기법 ED에 비해서 액션들의 가중치에 따라서 민감하게 성능이 변화하지 않는 것으로 나타났다. 이를 통해서 기법 ES는 분석가가 특정한 액션에 관심을 갖고 있지 않거나, 액션의 가중치에 영향을 받지 않고, 정보 파급 분석을 수행하기를 원하는 경우, 이용하기에 적당한 기법임을 알 수 있다.

6. 결 론

본 논문에서는 블로그 월드에서 발생하는 정보 파급 현상을 분석하기 위해서, 블로그 네트워크를 구성하는 방안에 대하여 논의하였다. 정보 파급 현상의 분석을 위해서 독립 전파 모델을 이용하는데, 독립 전파 모델로 분석이 가능한 블로그 네트워크를 구성하기 위해서는 블로그 사용자 간에 정보가 파급될 확률을 부여하는 것이 필수적이다. 그러나 기존의 연구들은 사회 연결망에서의 파급 현상을 설명하기 위한 모델을 제안하는 데에 초점을 맞추고 있었으므로, 사용자 간의 파급 확률로서 실질적인 값을 부여하기 위한 방안에 대한 연구는 미흡한 상태이다.

이에 본 논문에서는 두 사용자 간의 파급 확률이 한 사용자가 갖고 있는 게시글들 중에서 다른 사용자가 관심을 갖고 파급해간 게시글들의 비율과 관련이 있다는 점에 착안하여, 파급 확률로서 실질적인 값을 부여하는 방안을 제안하였다. 이를 위해서 게시글들의 파급 유효도를 계량화하는 방법을 제안하였다. 이를 이용하여, 사용자가 파급을 의도하고 작성한 게시글들과 이 사용자가 갖고 있는 게시글 중에서 특정 사용자에게 파급된 게시글들의 개수의 비율을 파급 확률로 이용하는 방안을 제안하였다. 또한, 사용자가 파급을 의도한 전체 규모 중에서 특정 사용자에게 실제로 파급된 규모의 비율을 파급 확률로 이용하는 방안을 제안하였다.

참 고 문 헌

[1] Blogger.com Co., Ltd. http://blogger.com.
[2] MySpace.com Co., Ltd. http://www.myspace.com
[3] NHN Co., Ltd. http://blog.naver.com.
[4] Cyworld Communications Co., Ltd., http://www.cyworld.com.
[5] G. Ellison, "Learning, Local Interaction, and Coordination," *Econometrica: Journal of the Econometric Society*, vol.61, no.5, pp.1047-1071, 1993.
[6] J. Goldenberg, Barak Libai, and Eitan Muller, *Talk of the network: A complex systems look at the underlying process of word-of-mouth*, Marketing Letters, 2001.
[7] WegoNet, *Brand Strategy in Communities*, (in Korean), E-Design Press, 2004.
[8] M. Granovetter, "Threshold Models of Collective Behavior," *American Journal of Sociology*, AJS, vol.86, no.6, pp.1420-1443, 1978.
[9] D. Kempe, et al., "Maximizing the Spread of Influence Through a Social Network," In *Proc. ACM Int'l. Conf. on Knowledge Discovery and Data Mining*, ACM SIGKDD, pp.137-146, 2003.
[10] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval," *ACM Press*, pp.75-84, 1999.