# 블로그 월드를 위한 커뮤니티 추출 방안

신정환, 김상욱, 윤석호
한양대학교 전자컴퓨터통신공학과
e-mail:{sin, wook, bogely}@agape.hanyang.ac.kr

# On Extracting a Community in the Blog World

Jung-Hwan Shin, Sang-Wook Kim, Seok-Ho Yoon
Department of Electronics and Computer Engineering, Hanyang University

**요 약**

블로그 월드에는 동일한 주제와 관련된 포스트들에 공통적으로 관심을 보이는 블로거들이 존재한다. 본 논문에서는 이러한 블로거들의 집합을 블로그 커뮤니티(blog community)라 정의한다. 블로그 커뮤니티는 타겟 마케팅, 양질의 정보 공유, 블로그 월드의 활성화 등 다양한 블로그 비즈니스 정책을 수립하는데 활용될 수 있다. 그러나 블로그 커뮤니티는 카페 등과 달리 멤버쉽으로 운영되는 집단이 아니기 때문에 커뮤니티에 속하는 멤버를 쉽게 파악할 수 없다. 본 논문에서는 주어진 주제와 관련된 블로그 커뮤니티를 추출하는 효과적인 방법을 제안한다. 먼저, 주어진 주제에 대한 시드 포스트들을 선택하고, 이 시드 포스트들을 통해서 주제와 관련된 블로거들을 선택한다. 다음으로, 선택된 블로거들을 통해서 주제와 관련된 포스트들을 선택한다. 위와 같은 과정을 반복해 나가면서 블로그 월드에 존재하는 주어진 주제와 관련된 모든 블로거들을 선발한다. 실제 블로그 데이터를 이용한 실험을 통하여 제안하는 방법의 우수성을 검증하였다.

## 1. 서론

블로그는 개인의 관심사에 따라 온라인에 글을 게시할 수 있는 개인 웹사이트다. 여기서 블로그에 게시된 글을 포스트라고 한다. 블로그의 주인인 블로거는 다른 블로그에 있는 포스트에 여러 가지 액션을 통해 관심을 표현한다. 액션에는 포스트에 짧은 의견을 적는 덧글(comment), 포스트를 복사해서 자신의 블로그에 가져오는 스크랩(scrap), 특정 포스트의 내용을 참고하여 그에 대한 링크를 포함하는 새로운 포스트를 작성하는 트랙백(trackback)이 있다.

블로그 월드에는 동일한 주제와 관련된 포스트들에 공통적으로 관심을 보이는 블로거들이 존재한다. 본 논문에서는 이러한 블로거들의 집합을 블로그 커뮤니티(blog community)라고 정의한다. 특정 주제와 관련된 블로그 커뮤니티의 추출은 커뮤니티의 주제와 관련된 타겟 마케팅(target marketing), 양질의 포스트 공유[1], 블로그 월드 활성화 등 다양한 블로그 비즈니스 정책을 수립하는데 이용할 수 있다.

그러나 블로그 커뮤니티는 카페 등과 달리 멤버쉽(membership)으로 운영되는 집단이 아니기 때문에 커뮤니티에 속하는 멤버를 쉽게 파악할 수 없다[2]. 본 논문에서는 블로그 월드에서 발생하는 여러 가지 액션을 활용하여 주어진 주제에 관한 블로그 커뮤니티를 추출하는 방안에 관하여 논의하고자 한다.

## 2. 관련연구

기존의 커뮤니티 추출 방안에 관한 연구는 웹 커뮤니티 추출 방안과 블로그 커뮤니티 추출 연구로 나눌 수 있다. 주어진 주제에 대하여 웹 커뮤니티를 추출하는 방법으로는 주어진 웹 페이지에 대하여 관련된 웹 페이지들을 검색하는 연구가 있다. 대표적인 방법으로 Companion 알고리즘을 이용한 방법[3], Co-citation을 이용한 방법[3], HITS[4]를 이용한 방법[5], 그리고 랜덤 워크를 이용한 방법[6] 등이 있다. 특정 주제에 대한 커뮤니티가 아니라 웹 상의 모든 커뮤니티들을 추출하는 연구도 있다. 대표적인 방법으로 (i,j)코어 알고리즘[1]을 이용한 방법과 최대 흐름-최소 분할(max flow-min cut) 이론을 이용한 방법[7]이 있다.

그러나 이러한 웹 커뮤니티 추출과 관련된 연구들은 블로그 커뮤니티 추출에 그대로 적용하기 어렵다. 그 이유는 다음과 같다. 첫째, 웹에는 웹 페이지, 한 종류의 객체만 존재하지만 블로그 월드에는 블로그와 포스트, 서로 다른 두 종류의 객체가 존재한다. 둘째, 웹에서 링크는 하이퍼링크가 유일하지만, 블로그 월드에는 액션에 따라 다양한 종류의 링크가 있다.

블로그 월드를 대상으로 하는 기존의 연구는 특정 주제와 관계없이 모든 커뮤니티들을 추출하는 연구가 있었다[2]. 이러한 연구에서 특정 주제에 관련된 커뮤니티를 찾기 위해서는 추출된 모든 커뮤니티의 내용을 직접 확인해야만

한다. 본 논문에서는 블로그 월드에서 주어진 주제와 관련된 블로그 커뮤니티를 찾는 새로운 방안을 제안하고자 한다.

### 3. 제안하는 방법

블로거는 자신이 관심 있는 주제와 관련된 포스트들에 액션을 가하는 경향을 보인다. 만일, 어떤 블로거가 주제 A에 해당되는 포스트들에 액션을 가하는 경향이 있다면, 그 블로거는 주제 A에 관심이 있는 것으로 볼 수 있다. 역으로 주제 A에 관심이 있다고 판단되는 다수의 블로거들에게 액션을 받은 포스트들은 주제 A와 관련된 내용을 포함하고 있다고 볼 수 있다. 본 논문에서는 이러한 특징을 이용하여 주어진 주제에 관련된 포스트들의 집합과 블로거들의 집합을 점진적으로 찾아냄으로써 최종적인 블로그 커뮤니티를 찾아내고자 한다.

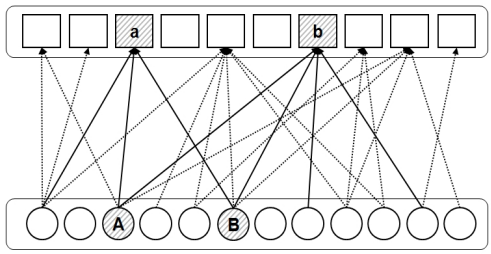
제안하는 블로그 커뮤니티 추출 방안은 다음과 같다. 먼저, 해당 주제와 관련된 소수의 시드 포스트들을 도메인 전문가를 통해 선발한다. 이러한 시드 포스트들에 대하여 기준 이상의 액션들을 가한 블로거들을 선발한다. 이와 같이, 주어진 주제에 관심을 보이는 블로거를 *커뮤니티 멤버(community member)*로 정의한다. 다시 선발된 커뮤니티 멤버들로부터 기준 이상의 액션을 받은 포스트들을 선발한다. 이와 같이, 해당 주제의 커뮤니티 멤버들이 관심을 보이는 포스트를 *커뮤니티 포스트(community post)*로 정의한다. 앞서 제안한 커뮤니티 멤버와 커뮤니티 포스트를 선발하는 과정을 더 이상 기준을 만족하는 커뮤니티 멤버가 나타나지 않을 때까지 반복한다.

그림 1은 커뮤니티 멤버와 커뮤니티 포스트 선발 과정을 나타낸 것이다. 상단의 사각형은 포스트를 의미하고, 하단의 원은 블로거를 의미한다. 원에서 사각형으로 그려진 화살표는 블로거가 포스트에 취한 액션을 나타낸다. 본 예에서는 커뮤니티 멤버와 커뮤니티 포스트의 선발 기준을 두 번 이상의 액션으로 가정한다. 그림 1의 (a)에서 a와 b는 커뮤니티 포스트들을 나타내며, A와 B는 커뮤니티 포스트들에게 기준 이상의 액션을 가하여 커뮤니티 멤버로 선발된 블로거들을 나타낸다. 그리고 A와 B를 통해 그림 1의 (b)에서와 같이 c가 새롭게 커뮤니티 포스트로 선발된다.

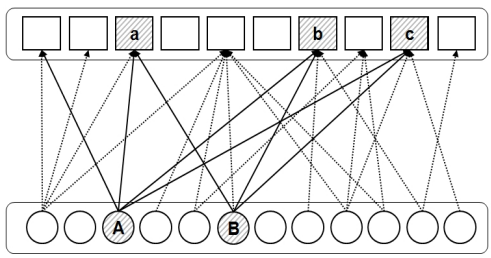
커뮤니티 멤버 선발 시 액션의 양만을 기준으로 하는 경우, 해당 커뮤니티의 선발 기준은 만족하지만 여러 주제에 관심이 있는 블로거가 커뮤니티 멤버로 선발될 우려가 있다. 여러 주제에 관심이 있는 블로거들이 커뮤니티 멤버로 선발되면 주어진 주제와 관련이 없는 블로그가 커뮤니티에 진입할 수 있다. 이러한 문제점을 해결하기 위해서 커뮤니티 멤버를 선발하는 기준으로 액션의 순도를 함께 이용한다. 액션의 순도는 다음과 같이 정의되며, 임의의 블로거가 얼마나 해당 주제에 대해서만 관심을 보이는지를 나타낸다.

$$\text{블로거 액션의 순도} = \frac{\text{커뮤니티 포스트에 가한 액션 수}}{\text{전체 액션 수}}$$

마찬가지로 커뮤니티 포스트의 선발 기준에도 액션의 순도를 함께 이용한다.




(a) 커뮤니티 멤버 선발




(b) 커뮤니티 포스트 선발

(그림 1) 블로그 커뮤니티 추출 과정

### 4. 성능 평가

본 실험의 목적은 제안하는 방법으로 추출된 블로그 커뮤니티가 얼마나 주제에 대하여 관련된 커뮤니티 멤버들을 포함하는가를 규명하는 것이다.

#### 4.1. 실험 환경

본 실험에서는 네이버 블로그 데이터를 이용한다. 개인 정보 유출을 사전에 방지하기 위해 블로거 개인 정보를 완전히 제거한 액션 통계 정보만을 사용하여 실험하였다.

전체 데이터 중 액션이 일정 이상인 250만 개의 블로그와 1,000만 개의 포스트를 이용하여 '요리, 축구, 영어, 여행, 자동차'의 5개의 주제에 대해서 커뮤니티를 추출하였다. 시드의 선발 기준은 (1)내용이 주제에 적합하며, (2)양질의 정보를 가지고 있고, (3)일정 이상의 액션수를 가진 포스트이다.

성능 평가 척도로는 아래의 식과 같이 정의되는 블로그 커뮤니티 정확도를 이용하였다.

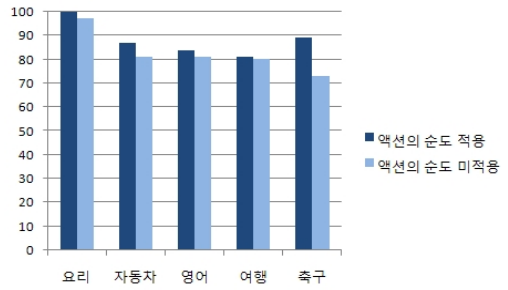
$$\text{블로그 커뮤니티 정확도} = \frac{\text{주제와 관련된 커뮤니티 멤버 수}}{\text{전체 커뮤니티 멤버 수}}$$

커뮤니티 멤버가 주어진 주제와 관련 있는지 알아보기 위하여 블로그가 해당 주제에 관한 포스트를 가지고 있는지 여부를 직접 확인하였다.

### 4.2. 실험 결과

첫 번째 실험은 제안하는 방법에서 커뮤니티 멤버와 커뮤니티 포스트의 선발 기준으로 액션의 양만을 이용한 경우와 액션의 순도를 함께 이용하는 경우에 대한 실험이다. 각 주제에 대하여 80개의 시드 포스트들을 선발하여 실험하였다.

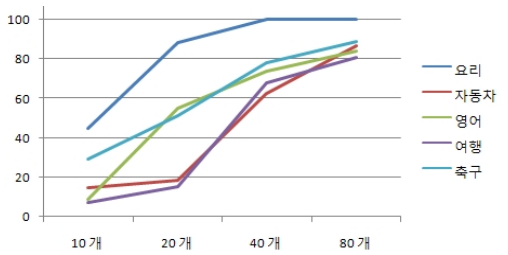
그림 2는 5개의 주제에 대하여 제안하는 방법으로 추출된 블로그 커뮤니티들의 정확도를 나타낸 것이다. 액션의 순도를 함께 적용한 경우 액션의 양만을 사용한 경우보다 모두 높은 정확도를 보였다. 요리의 경우 정확도가 99%로 가장 높게 나왔으며 나머지 경우에서는 80~90% 사이의 정확도를 보였다. 요리에 대한 커뮤니티 정확도가 비교적 높게 나온 이유는 요리 주제에 대한 질이 좋은 포스트가 많아서 블로거들이 요리에 대해 다양한 액션을 많이 했기 때문이다.




(그림 2) 블로그 커뮤니티의 정확도

두 번째 실험은 시드의 개수에 따른 블로그 커뮤니티 정확도를 평가하는 실험이다. 5개의 주제에 대하여 시드의 개수를 10개, 20개, 40개, 80개로 설정하고, 각 설정에 따른 블로그 커뮤니티의 정확도를 확인하였다.

그림 3은 실험 결과를 나타낸 것이다. 가로축은 시드의 개수를 나타내며 세로축은 블로그 커뮤니티의 정확도를 나타낸다. 실험 결과에 의하면 시드의 수가 적으면 정확도가 매우 낮고, 시드의 수가 증가할수록 정확도는 높아지나 시드의 수가 일정 이상 되면 정확도가 향상되는 정도가 상대적으로 약화되는 것으로 나타났다.




(그림 3) 시드 개수에 따른 블로그 커뮤니티 정확도

## 4. 결론

블로그 월드에는 동일한 주제와 관련된 포스트들에 공통적으로 관심을 보이는 블로거들이 존재한다. 본 논문에서는 이러한 블로거들의 집합을 블로그 커뮤니티라고 정의하였다. 특정 주제와 관련된 블로그 커뮤니티의 추출은 커뮤니티의 주제와 관련된 다양한 블로그 비즈니스 정책을 수립하는데 이용할 수 있다. 본 논문에서는 주어진 주제와 관련된 블로그 커뮤니티를 추출하는 방법을 제안하였다. 제안하는 방법은 주어진 주제에 대한 시드 포스트들을 이용하여 커뮤니티 멤버와 커뮤니티 포스트를 단계적으로 확장함으로써 최종적으로 블로그 커뮤니티를 추출한다. 실험을 통해 제안하는 방법의 우수성을 검증하였다.

## 감사의 글

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

## 참고 문헌

[1] R. Kumar et al., "Trawling the Web for Emerging Cyber-Communities," In *Proc. 8th Int'l. Conf. on World Wide We b*, WWW, pp. 1481-1493, 1999.

[2] K. Ishida, "Extracting Latent Weblog Communities: A Partitioning Algorithm for Bipartite Graphs," In *Proc. 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, WWW, 2005.

[3] J. Dean and M. R. Henzinger, "Finding Related Pages in the World Wide Web," In *Proc. 8th Int'l. Conf. on World Wide Web*, WWW, pp. 1467-1479, 1999.

[4] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," In *Proc. the ACM-SIAM Symp. on Discrete Algorithms*, SODA, pp. 604-632, 1998.

[5] D. Gibson, J. M. Kleinberg, and P. Raghavan, "Inferring Web Communities from Link Topology," In *Proc. the 9th ACM Conf. on Hypertext and Hypermedia*, Hypertext, pp. 225-234, 1998.

[6] J. Huang, T. Zhu, and D. Schuurmans, "Web Communities Identification from Random Walks," In *10th European Conf. on Principles and Practice of Knowledge Discovery in Database, PKDD*, pp. 187-198, 2006.

[7] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient Identification of Web Communities," In *Proc. the 6th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, KDD, pp. 150-160, 2000.