■ 2009년도 학생논문 경진대회 수상작

블로그 월드에서 주제 중심의 잠재적 커뮤니티 추출 방안

(Extraction of Latent Topic-based Communities in Blogspace)

신 정 환 † 윤 석 호 †† 김 상 욱 ††† 박 선 주 ††††
(Jung-Hwan Shin) (Seok-Ho Yoon) (Sang-Wook Kim) (Sunju Park)

요 약 블로그 월드에는 동일한 주제와 관련된 포스트들과 이 포스트들에 공통적으로 관심을 보이는 블로거들이 존재한다. 본 논문에서는 이러한 블로거들과 포스트들의 집합을 블로그 커뮤니티로 정의한다. 블로그 커뮤니티는 타겟 마케팅, 양질의 정보 공유, 블로그 월드의 활성화 등 다양한 블로그 비즈니스 정책을 수립하는데 활용될 수 있다. 블로그 커뮤니티는 카페 등과 달리 멤버십으로 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. 이와 같은 과정을 반복해 나가면서 블로그 월드에 존재하는 주어진 주제와 관련된 블로거들과 포스트들을 선발한다. 본 논문에서는 추출된 블로그 커뮤니티 주제의 정확도를 측정함으로써 제안하는 방법의 우수성을 검증하였다.

키워드 : 블로그월드, 블로그 커뮤니티 추출, 데이터 마이닝

Abstract In blogspace, there are posts that deal with a common topic and bloggers that are interested in these posts. In this paper, we define a blog community as a group of these bloggers and posts. With a blog community, we can establish various business policies for target marketing, sharing high quality data, and mobilizing the activities in the blogspace. Unlike internet cafes, bloggers participate in blog communities without explicit membership. So, it is not easy to identify the members of a community. In this paper, we propose an effective approach for extracting a blog community that is related to a given topic. First, we choose seed posts that is highly related to a given topic, and select bloggers that are related to the topic with the seed posts. Then, we select posts that are related to the topic with the selected bloggers. By repeating this, we find all the posts and bloggers that are members of the community related to a given topic in blogspace. We verify the superiority of the proposed approach by analyzing extracted blog communities.

Key words : Blogosphere, Extraction of Blog Communities, Data Mining

· 본 연구는 NHN(주)의 지원을 받았습니다. 그러나, 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자(들)의 것이며, 반드시 지원회사의 입장을 대변하는 것은 아닙니다. 또한, 본 연구는 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 부분적인 지원을 받았습니다(No. R01-2008-000-20872-0).

† 정 회 원 : 휴어테크 코딩인스펙터팀 사원
sin@agape.hanyang.ac.kr
†† 학생회원 : 한양대학교 전자컴퓨터통신공학과
bogely@agape.hanyang.ac.kr
††† 종신회원 : 한양대학교 전자컴퓨터통신공학과 교수
wook@hanyang.ac.kr
†††† 종신회원 : 연세대학교 경영학부 교수
boxenju@yonsei.ac.kr

논문접수 : 2009년 6월 12일
심사완료 : 2009년 11월 11일

정보과학회논문지: 데이타베이스 제37권 제1호(2010.2)

1. 서 론

커뮤니티(community)란 사교성, 정보, 소속감, 사회적 정체성 등을 바탕으로 사람들 사이에서 발생하는 연결망(network)이다[1]. 고전적인 의미의 커뮤니티는 주로 가까운 곳에 거주하는 이웃들 사이의 관계로 제한되었었지만, 20세기 들어 교통수단의 발달과 전화와 인터넷을 비롯한 각종 통신 매체의 발달로 이러한 공간적인 제약이 거의 사라졌다. 최근에는 온라인에서도 커뮤니티를 쉽게 찾아볼 수 있다[1]. 관심사가 비슷한 사람들끼리 모여 운영되는 온라인 카페, 친구나 동창들을 서로 연결해 주는 블로그 월드(blogosphere), 포털 사이트가 주제별로 운영하는 게시판, 마케팅이나 고객 관리의 차원에서 특정 상품을 주제로 운영되는 기업 사이트 등이 그 대표적인 예이다.

대부분의 온라인 커뮤니티는 멤버 가입에 의한 멤버십을 통하여 운영되므로 커뮤니티의 바운더리(boundary)가 명시적으로 드러난다[2]. 그러나 참여의사를 특별히 밝히지 않았다 하더라도 공통적인 주제에 대해 흥미를 가지는 사람들의 집합을 하나의 커뮤니티로 간주할 수도 있다. 본 논문에서는 이러한 커뮤니티를 잠재적 커뮤니티(latent community)라 정의한다. 잠재적 커뮤니티는 멤버십으로 운영되는 것이 아니기 때문에 커뮤니티에 속하는 멤버들을 찾아내거나 바운더리를 정하는 데 어려움이 있다[3].

블로그 월드에서도 이러한 잠재적인 온라인 커뮤니티의 예를 찾아볼 수 있다. 블로거들은 포스트들에게 트랙백(trackback)이나 덧글(comment)등의 액션을 통하여 관심을 표현한다. 따라서 공통적인 주제에 대한 내용을 담고 있는 포스트들과 이러한 포스트들에 액션을 취함으로써 관심을 표현한 블로거들의 집합을 잠재적인 커뮤니티로 간주할 수 있다. 본 논문에서는 이러한 블로그 월드 내에서의 잠재적 커뮤니티를 블로그 커뮤니티(blog community)라 정의하고, 블로그 커뮤니티에 속하는 블로거들과 포스트들을 각각 커뮤니티 멤버(community member)와 커뮤니티 포스트(community post)라 정의한다.

본 논문에서는 주어진 주제와 관련된 잠재적인 블로그 커뮤니티를 추출하는 새로운 방법을 제안한다. 특정 주제에 대한 잠재적인 온라인 커뮤니티 추출에 관한 기존 연구는 주로 웹 커뮤니티를 대상으로 진행되었다[4-7]. 웹 커뮤니티는 임의의 주제에 대하여 연관된(related) 웹페이지들의 집합을 말하며, 웹 커뮤니티를 찾기 위하여 하이퍼링크 정보를 이용한다. 그러나 웹에서는 웹페이지라는 한 종류의 객체 타입만 존재하지만 블로그 월드에서는 블로거와 포스트라는 두 종류의 객체 타입이 존재하며, 이 두 객체 타입들 사이에 서로 다른 종류의 액션으로 인한 상이한 링크 타입들이 존재한다. 따라서 하이퍼링크만을 대상으로 하는 기존의 웹 커뮤니티 추출 방법을 블로그 월드에 직접 적용하기에는 어려움이 있다.

본 논문에서는 주어진 주제에 대하여 관심을 가지는 커뮤니티 멤버들의 집합과 이 주제에 대한 내용을 담고 있는 커뮤니티 포스트들의 집합을 점진적으로 확대해나가는 과정을 반복함으로써 블로그 커뮤니티를 추출하는 방법을 제안한다. 우선 주어진 주제에 대한 내용을 담고 있는 소수의 시드 포스트들을 선발한다. 그 다음, 이 시드 포스트들에 대하여 일정 기준 이상의 액션을 가한 블로거들을 커뮤니티 멤버로 선발한다. 이렇게 선발된 커뮤니티 멤버들로부터 일정 기준 이상의 액션을 받은 포스트들을 새로운 커뮤니티 포스트를 추가 선발한다. 이러한 과정을 반복하여 미리 예측된 커뮤니티 크기에 도달할 때까지 커뮤니티 멤버들과 커뮤니티 포스트들을 점진적으로 확장해 나간다. 이와 더불어 추출된 커뮤니티의 질을 향상시키는 다양한 방법에 대하여 논의한다. 실제 블로그 데이터를 이용한 다양한 실험을 통하여 제안된 방법의 우수성을 검증한다.

이러한 온라인 커뮤니티를 추출하는 것은 다음과 같은 세 가지 이유 때문에 중요하다. 첫째로 사용자가 원하는 유용한 정보를 쉽고 빠르게 제공해줄 수 있다. 둘째로 온라인 사회에 진화에 대한 통찰력 있는 지식을 제공해 줄 수 있다. 셋째로 온라인 상에 존재하는 포탈 회사들이 이러한 커뮤니티를 추출함으로써 타겟 광고를 할 수 있다[3].

본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구와 관련된 기존의 연구들을 기술한다. 제 3장에서는 주제별 잠재적 커뮤니티를 추출하기 위한 제안된 방법의 기본 전략, 알고리즘, 그리고 알고리즘의 개선 방안 등에 관하여 상세히 다룬다. 제 4장에서는 다양한 실험을 통해서 제안하는 방안이 블로그 커뮤니티를 올바르게 추출하는지 검증한다. 제 5장에서는 결론을 제시한다.

2. 관련 연구

온라인에서 링크 정보를 이용하여 잠재적인 커뮤니티를 추출하는 연구들은 주로 웹페이지들을 대상으로 이루어졌으며, 최근에는 블로그 사용자의 수가 증가함에 따라 블로그 커뮤니티를 추출하는 연구들이 진행되고 있다. 제 2.1절에서는 기존의 웹 커뮤니티 추출 연구들에 대하여 간략히 설명하고, 제 2.2절에서는 블로그 커뮤니티 추출 연구들에 대하여 논의한다.

2.1 웹 커뮤니티 추출

기존의 웹 커뮤니티 추출 방법들은 웹 페이지들 사이

의 하이퍼링크로 형성된 위상 구조를 분석하여 커뮤니티를 추출한다. 이들은 웹 커뮤니티를 '커뮤니티의 외부보다 내부에서 더 많은 하이퍼링크를 가지는 웹 페이지들의 집합'으로 정의[8-10]하고, 하이퍼링크가 밀집된 정도를 기준으로 전체 연결망을 분할하여 커뮤니티를 추출한다[2,8-12]. 웹 커뮤니티 추출에 관한 기존의 연구는 웹의 전체 연결망에 존재하는 '모든' 커뮤니티들을 추출하는 연구와 사용자의 관심 대상이 되는 '특정' 커뮤니티 하나를 추출하는 연구로 구분될 수 있다.

모든 커뮤니티들을 추출하는 연구 중에서 참고 문헌 [3]는 웹 커뮤니티의 핵심(core)을 완전 부분 그래프(complete sub-graph)로 보고 전체 연결망에서 완전 부분 그래프를 추출한 후 HITS 알고리즘[11]을 이용하여 웹 커뮤니티를 추출하는 방법을 제안하였다. 특정 커뮤니티 하나를 추출하는 연구들 중 참고문헌 [6]에서는 웹 커뮤니티를 '특정 주제와 관련된(related) 웹 페이지들의 집합'으로 정의하였다. 참고문헌 [4-7]에서는 특정 주제의 내용을 담은 웹 페이지를 질의로 제시하였을 때, 이 웹 페이지와 밀접한 관련이 있는 웹 페이지들을 커뮤니티로 추출하는 방법을 제안하였다. 질의 웹 페이지를 이용한 대표적인 웹 커뮤니티 추출 방안으로는 Companion 알고리즘과 Co-citation 알고리즘이 있다[5]. Companion 알고리즘은 질의 웹 페이지의 부모들(parents), 자식들(children), 그리고 형제들(siblings)로 이루어진 인접(vicinity) 그래프를 추출한다. 추출된 인접 그래프를 대상으로 HITS 알고리즘을 적용하여 질의 웹 페이지와 추출된 노드들 사이의 관련된 정도를 계산한다. 이는 질의 웹 페이지를 공통으로 참조하거나, 동일한 웹 페이지들을 질의 웹 페이지와 공통으로 참조하는 웹 페이지들을 질의 웹 페이지와 주제가 유사한 웹 페이지들로 보고, 이 웹 페이지들 중에서 권위가 높은 웹 페이지를 찾는 방법이다. Co-citation 알고리즘은 링크 정보를 이용하여 두 노드의 유사도를 측정하는 대표적 척도인 Co-citation[13]을 이용하여 주어진 웹페이지와 유사하다고 판단되는 웹페이지들을 커뮤니티로 추출하는 방법을 제안하였다[5]. 참고문헌 [8,9,12]에서는 최대 흐름(maximum flow) 알고리즘[14]을 이용하여 웹 상에서 정보의 흐름이 최대가 되도록 하는 서로 긴밀하게 연결된 노드들의 부분 그래프를 추출하는 방법을 제안하였다. 이는 커뮤니티 내에서 멤버간의 상호 의사소통이 활발하게 이루어진다는 개념을 최대 흐름 알고리즘을 이용해서 웹 환경에 적용한 것이다.

기존의 웹 커뮤니티 추출에 관한 연구들에서는 웹 페이지라는 하나의 객체 타입과 하이퍼링크라는 하나의 링크 타입만이 존재한다. 그러나 본 논문에서 추출하고자 하는 블로그 커뮤니티는 블로거와 포스트라는 두 종류의 객체 타입이 존재한다. 또한 블로거와 포스트 사이에는 하이퍼링크 외의 여러 종류의 링크 타입으로 표현될 수 있는 다양한 액션들이 존재한다. 따라서 기존의 웹 커뮤니티 추출 방법을 블로그 커뮤니티에 그대로 적용할 수는 없다.

2.2 블로그 커뮤니티 추출

기존의 블로그 커뮤니티에 관한 연구들은 주로 기존의 웹 커뮤니티 추출 방법을 이용하여 블로그 월드에서 커뮤니티들을 찾는 연구들이었다. 참고문헌 [15]에서는 웹 커뮤니티의 핵심인 완전 서브 그래프를 추출하는 방법을 블로그 환경에 적용할 수 있도록 확장하였다. 참고문헌 [16]에서는 블로그와 웹페이지를 이분 그래프로 나타내고 서로 관련성이 약한 블로그 쌍의 연결을 제거하여 블로그 커뮤니티를 추출하는 WP(weakest pair) 알고리즘을 제안하였다.

최근에는 커뮤니티의 본질적인 의미를 블로그 커뮤니티에 적용하는 연구들이 진행되고 있다[17-19]. 해당 연구들은 현실 세계에서 커뮤니티란 멤버들간의 상호 교류(communication)로 형성된 집단[1]이라는 개념을 블로그 월드에 적용하여 두 블로거 사이에 존재하는 교류의 정도를 정량적으로 분석하여 블로그 커뮤니티를 추출하였다. 참고문헌 [19]에서는 주어진 주제에 대한 커뮤니티를 찾기 위하여 시드 블로거로부터 n단계 안으로 연결된 모든 블로거들을 해당 주제와 관련 있는 후보 블로거들로 가정한다. 후보 블로거들이 선택되면 후보 블로거들 사이의 교류 정도를 액션의 수로 평가하여 가중치 그래프를 생성하고, 생성된 가중치 그래프에 agglomerative clustering의 방식인 Islands Partitioning 알고리즘을 이용해서 그래프를 분할하여 블로그 커뮤니티를 찾는다.

본 논문에서는 특정 주제와 관련이 있는 블로그 커뮤니티를 추출하고자 한다. 따라서 전체 연결망에 존재하는 모든 커뮤니티를 추출하는 WP(weakest pair) 알고리즘은 이러한 목적에 적합하지 않다. 물론, 기존의 블로그 커뮤니티 추출 방법으로 모든 블로그 커뮤니티를 찾은 후에 도메인 전문가를 통해서 추출된 블로그 커뮤니티들의 주제를 차례로 파악한 후에 주어진 주제와 일치하는 블로그 커뮤니티를 선택할 수도 있다. 그러나 이러한 방법은 블로그 월드의 규모를 고려했을 때 적절하지 않은 방법이다. Islands Partitioning 알고리즘을 이용한 커뮤니티 추출 방법은 시드 블로거를 이용하여 커뮤니티를 추출하기 때문에 특정 주제와 관련이 있는 블로그 커뮤니티를 추출할 수 있다. 그러나 이러한 방법은 단순히 액션을 수를 이용해서 커뮤니티를 추출하기 때문에 해당 주제와 관련이 적거나 없는 블로거들이 커뮤니티에 포함될 우려가 있다. 따라서 본 논문에서는 주어

진 주제와 관련이 있는 블로그 커뮤니티를 추출하는 새로운 방법을 제안하고자 한다.

3. 주제별 커뮤니티 추출 방법

3.1 문제 정의

본 논문에서는 블로그 월드를 그림 1과 같이 이분 그래프(bipartite graph)로 표현한다. 상단의 사각형으로 표현된 노드들은 포스트들을 나타내고, 하단의 원으로 표현된 노드들은 블로거를 나타낸다. 화살표로 표시된 에지들은 블로거가 포스트에 가한 액션(댓글, 트랙백, 스크랩 등)을 나타낸다. 기존의 연구에서와 같이 커뮤니티를 긴밀하게 상호 연결된 서브 그래프라고 정의한다면[20-22], 그림 1의 점선 내의 색칠된 다각형에 포함된 노드들은 다른 노드들에 비해 밀집되게 연결되어 있으며, 이러한 집단을 커뮤니티로 간주할 수 있다.

그러나 위상 구조적으로 밀집하게 연결된 그림 1의 커뮤니티를 '특정 주제'에 대한 블로그 커뮤니티라 확신할 수는 없다. 특정 주제에 대한 블로그 커뮤니티는 해당 주제와 관련된 노드들로만 밀집되게 연결되어 있어야 하기 때문이다. 예를 들어, P5는 커뮤니티 내에서는 3개의 에지를 갖고 있지만, 커뮤니티 밖으로부터 그보다 더 많은 에지를 갖고 있다. 따라서 커뮤니티 멤버가 아닌 블로거들로부터 관심을 더 많이 받은 P5는 해당 주제에 대한 노드가 아닐 가능성이 높다. 또한, 커뮤니티 멤버로 인정받지 못한 B3는 액션의 대부분을 커뮤니티에 속하는 포스트들에게 가한 것으로 보아 해당 주제에 대한 관심이 높은 노드이다. 그림 2는 위의 사항을 반영하여 새롭게 추출한 '특정 주제에 대한' 커뮤니티의 예를 보여준다.

주어진 그래프에서 위상 구조적으로 서로 밀집되게 연결된 노드들의 집합은 기존의 방법을 통해 쉽게 찾아낼 수 있다. 그러나 위의 예에서 본 바와 같이, 단지 서로 밀집되게 연결되었다고 해서 동일한 주제와 관련된 노드들이라고 단정지을 수는 없다. 본 논문에서는 각 노드들이 주어진 주제에 대하여 관련이 있는지를 평가하여, 주제와 관련된 노드들만으로 구성된 블로그 커뮤니티를 추출하는 새로운 방법을 제안한다.

3.2 블로그 커뮤니티 추출 알고리즘

제안하는 방법의 기본 아이디어는 신뢰할 수 있는 커

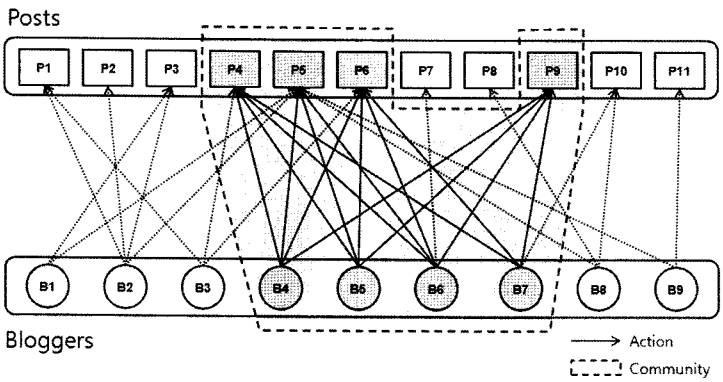


그림 1 이분 그래프로 표현된 블로그 월드에서 위상구조적으로 긴밀하게 연결된 커뮤니티의 예

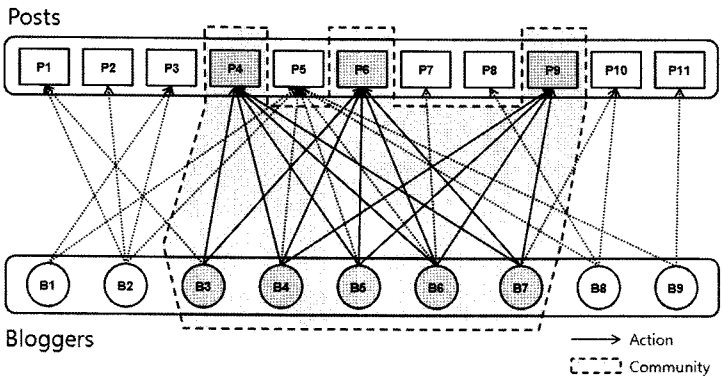


그림 2 커뮤니티의 내부 에지와 외부 에지를 동시에 고려하여 선택한 커뮤니티의 예

뮤니티 멤버들과 포스트들의 수를 단계별로 확장해 나가는 것이다. 주제 A에 관심을 가지는 블로거에게 액션을 받은 포스트는 주제 A에 대한 내용을 담고 있을 가능성이 높다. 따라서 주제 A에 관심을 가진 다수의 블로거들에게서 공통적으로 액션을 받은 포스트는 주제 A에 대한 내용을 담고 있는 포스트로 간주할 수 있다. 역으로, 주제 A에 대한 내용을 담고 있는 포스트에 액션을 가한 블로거는 주제 A에 관심이 있을 가능성이 높다. 따라서 주제 A에 대한 내용을 담고 있는 다수의 포스트들에게 액션을 가한 블로거는 주제 A에 관심이 있는 블로거로 간주할 수 있다. 제안하는 방법은 주제 A에 속하는 것이 확실히 판명된 소수의 포스트들을 시드로 하여 커뮤니티 멤버와 커뮤니티 포스트들을 점진적으로 확장해 나감으로써 블로그 커뮤니티를 찾아내고자 한다.

3.2.1 시드 포스트 선발

추출되는 블로그 커뮤니티의 정확도는 주어진 시드 포스트들에 의해 크게 좌우되므로 좋은 시드의 선발은 매우 중요하다. 따라서 본 논문에서는 커뮤니티 추출과정의 자동화를 지향함에도 불구하고 시드 포스트들은 도메인 전문가를 통해 선발한다.

시드 포스트는 다음 가이드라인에 따라 선발된다. 첫째, 관심 주제에 적합한 포스트여야 한다. 시드들의 주제는 블로그 커뮤니티의 주제를 결정하기 때문에 주제에 정확히 부합하는 포스트를 시드로 해야 정확한 블로그 커뮤니티를 추출할 수 있다. 둘째, 액션수가 일정 이상인 인기 있는 포스트이어야 한다. 시드 포스트들이 받은 액션의 수가 적으면 첫 단계의 확장 과정이 진행되지 않거나 진행되더라도 정확도가 떨어질 가능성이 높기 때문이다. 셋째, 주제에 대한 유용한 정보를 담고 있는 포스트여야 한다. 포스트가 유용한 정보를 담고 있어야 포스트가 받은 액션들을 신뢰할 수 있기 때문이다.

3.2.2 커뮤니티 멤버 및 커뮤니티 포스트의 단계별 확장

커뮤니티 멤버들과 커뮤니티 포스트들을 단계별로 확장해 나가는 방법을 설명하기 위한 기호들을 다음과 같이 정의한다.

각 단계에서는 우선 커뮤니티 멤버와 커뮤니티 포스트 후보를 선발한 후에 이들 중 검증 과정을 통과한 것들만을 최종적으로 커뮤니티 멤버와 커뮤니티 포스트로 선발한다. i 단계까지의 확장으로 선발된 커뮤니티 멤버의 집합을 CM(i)로, i 단계까지의 확장으로 선발된 커뮤니티 포스트의 집합을 CP(i)로 정의하며, i 단계에서 커뮤니티 멤버 후보로 선발된 블로거들의 집합을 CM(i)'로, i 단계에서 커뮤니티 포스트 후보로 선발된 포스트들의 집합을 CP(i)'로 정의한다. i 단계에서 커뮤니티 멤버 선발을 위한 기준으로 사용되는 임계값을 e_CM(i)로, i 단계에서 커뮤니티 포스트 선발 기준으로 사용되는 임계값을 e_CP(i)로 정의한다. 추출과정은 커뮤니티 포스트가 일정한 수만큼 구해지면 전체 커뮤니티 추출과정이 종료되며, 이 종료 기준이 되는 포스트 수의 하한과 상한을 LB, UB로 정의한다.

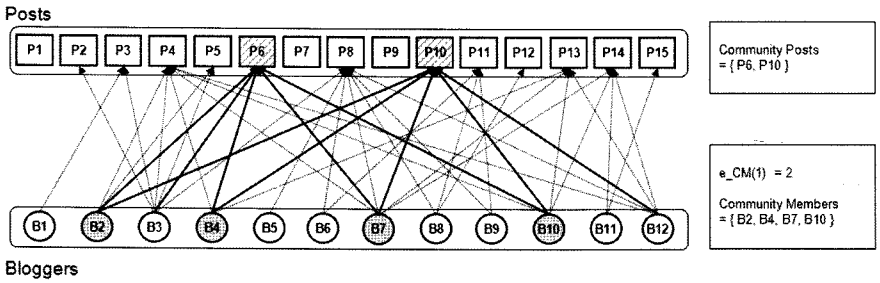
당분간 e_CM(1), LB, UB를 위한 값은 주어진다고 가정하고 논의를 전개한다. 이들의 값을 자동으로 설정하는 방법은 제 3.2.4절에서 자세히 논의한다. 단계별 확장은 다음과 같이 진행된다. 먼저, CP(0)은 도메인 전문가를 통해 선발된 시드 포스트들로 설정하고, CM(0)는 공집합으로 한다. CP(0)에 한번이라도 액션을 가했던 블로거는 1단계 커뮤니티 멤버 후보 CM(1)'가 되며, 그 중 e_CM(1) 이상의 액션을 가한 블로거들만을 CM(1)으로 추가 선발한다. CM(1)에게서 한번이라도 액션을 받은 포스트는 1단계의 커뮤니티 포스트 후보 CP(1)'가 되며, 그 중 e_CP(1) 이상의 액션을 받은 포스트들만을 CP(1)로 선발한다. 즉, 각 단계 i에서는 직전 단계 i-1에서 구한 CP(i-1)에게 e_CM(i) 이상의 액션을 취한 블로거를 CM(i)로 선발하고, 다시 이 CM(i)로부터 e_CP(i) 이상의 액션을 받은 포스트들을 CP(i)로 선발한다. 이러한 방법을 반복하여 커뮤니티를 점진적으로 확장해 나간다. 단계를 거듭할수록 커뮤니티에 속하는 블로거와 포스트가 증가하면서 전체적인 액션의 수도 함께 증가하기 때문에 커뮤니티 선발의 기준이 되는 다음 단계를 위한 임계값도 단계별로 증가시켜 나간다. 단계별 임계값의 자동 설정 방법은 제 3.2.3절에서 자세히 설명한다.

그림 3은 기존에 선발된 커뮤니티 멤버와 커뮤니티 포스트로부터 새로운 커뮤니티 멤버와 커뮤니티 포스트를 선발하는 과정의 예를 보여준다. e_CM(1)과 e_CP(1)은 각각 2와 3으로 설정되어 있다고 가정하고, 임계값들은 각 단계마다 1씩 증가한다고 가정하자. 실선 화살표는 커뮤니티 멤버 선발에 유효한 액션들을 표시하며, 점선 화살표는 그 외의 액션들을 표시한다.

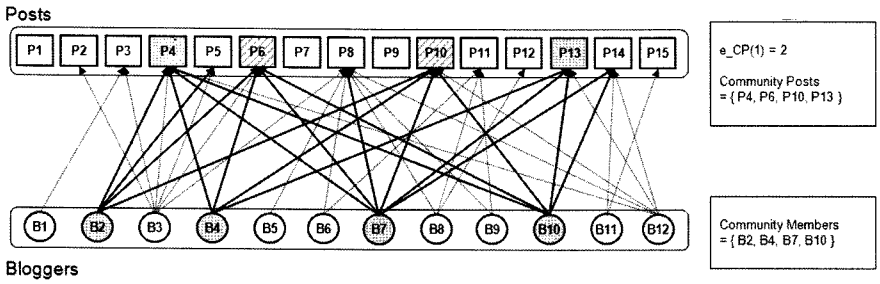
그림 3(a)에서 빗금으로 표시된 {P6, P10}은 시드 포스트들이며, 따라서 CP(0)가 된다. 시드 포스트 {P6, P10}에 액션을 가한 {B2, B3, B4, B6, B7, B10, B12}은 CM(1)'이 된다. 이 중에 색칠된 {B2, B4, B7, B10}들이 e_CM(1)(=2) 이상을 만족시켜 CM(1)으로 선발된다. 그림 5(b)는 그림 5(a)에서 선발된 CM(1)으로부터 CP(1)을 선발하는 과정을 보여준다. CM(1)으로부터 e-CP(1)(=3) 이상의 액션을 받은 {P4, P13}이 CP(1)에 추가된다. 마찬가지로, 그림 5(c)에서 CP(1)에 e-CM(2)(=3) 이상의 액션을 가한 {B12}가 CM(2)에 추가된다.

3.2.3 단계별 임계값 조정

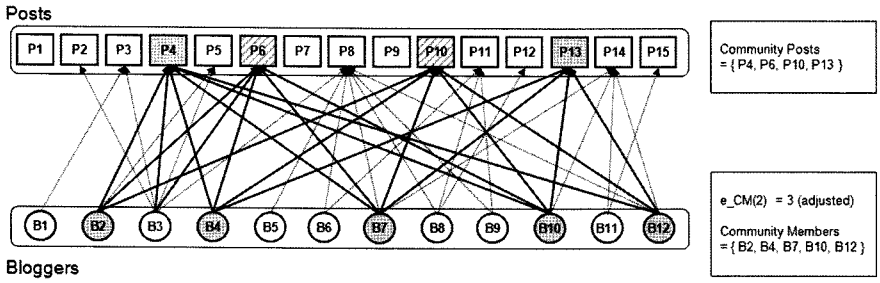
확장 단계를 거듭할수록 확장 과정에 참여하는 블로거와 포스트의 수가 증가하기 때문에, 확장 단계가 증가



(a) 시드 포스트 {P6, P10}으로부터 커뮤니티 멤버 {B2, B4, B7, B10}을 선발



(b) 커뮤니티 멤버 {B2, B4, B7, B10}으로부터 커뮤니티 포스트 {P4, P13}을 선발



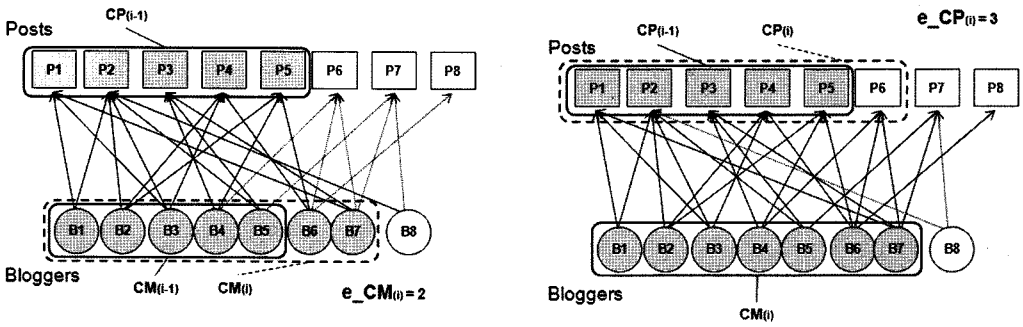
(c) 커뮤니티 포스트 {P4, P6, P10, P13}으로부터 커뮤니티 멤버 {B12}를 선발

그림 3 블로그 커뮤니티 추출 과정의 예

함에 따라 e_CM(i)와 e_CP(i)도 함께 증가시켜야 한다. 본 절에서는 i-1 단계에서 선발된 커뮤니티 멤버와 커뮤니티 포스트에 대한 정보를 활용하여 i 단계의 임계값을 합리적으로 조정하는 방법을 제안한다.

CM(i-1) 내의 각 커뮤니티 멤버가 CP(i-1)내의 포스트들에게 가한 액션 수를 파악하여 이들 중 최소값을 e_CM(i)로 조정하는 방법을 채택한다. 이는 i 단계에서 CM(i)에 포함될 새로운 커뮤니티 멤버들의 선발 요건을 CM(i)에 자동적으로 포함될 기존 커뮤니티 멤버들의 최소값으로 하겠다는 것을 의미한다. 마찬가지로, CP(i-1) 내의 각 커뮤니티 포스트가 CM(i)내의 멤버들로부터 받은 액션 수를 파악하여 이들 중 최소값을 e_CP(i)로 조정한다.

그림 4는 임계값을 자동으로 조정하는 예를 보여주고 있다. 그림 4(a)는 CP(i-1)가 선발된 상태에서 e_CM(i)을 조정하여 CM(i)을 선발하는 과정을 보여주고, 그림 4(b)에서는 CM(i)가 선발된 상태에서 e_CP(i)을 자동으로 설정하여 CP(i)을 선발하는 과정을 보여준다. 그림 4(a)에서 CM(i-1)의 각 커뮤니티 멤버는 적어도 두 번 이상 CP(i-1)의 커뮤니티 포스트들에게 액션을 가했으므로 e_CM(i)은 2로 설정된다. 따라서 이를 기준으로 CP(i-1)에 속한 커뮤니티 포스트들에게 두 번 이상의 액션을 가한 블로거{B6, B7}를 CM(i)에 추가한다. 그림 6(b)에서 CP(i-1)의 각 커뮤니티 포스트는 CM(i)의 커뮤니티 멤버들로부터 적어도 세 번 이상의 액션을 받으므로 e_CP(i)은 3으로 설정된다. 따라서 이를 기준으로 커뮤니티 멤버들에게 세 번 이상의 액션을 받은 포스트 {P6}을 CP(i)에 추가한다. 이러한 방식을 통하여 각 단



(a) 커뮤니티 멤버 선발 임계값 조정 (b) 커뮤니티 포스트 선발 임계값 조정

그림 4 임계값 조정 과정의 예

계에서의 e_CM(i), e_CP(i)를 위한 합리적인 값을 자동적으로 설정할 수 있다.

3.2.4 커뮤니티 멤버 선발을 위한 초기 임계값과 종료 기준의 설정

확장 과정의 임계값 e_CM(i)와 e_CP(i)은 i-1단계에서 선발된 커뮤니티 멤버들과 커뮤니티 포스트들의 액션을 고려하여 자동으로 조정되지만, e_CM(1) 선발 과정에서는 이전 단계가 존재하지 않으므로 동일한 방식의 임계값 자동 설정이 불가능하다. CM(i)와 CP(i)는 모두 초기 커뮤니티 멤버들을 기반으로 확장되므로, 초기 커뮤니티 멤버 선발 임계값 e_CM(1)은 선발된 블로그 커뮤니티의 크기와 질에 큰 영향을 미친다. e_CM(1)이 너무 작게 설정되면 초기에 너무 많은 블로거들과 포스트들이 블로그 커뮤니티에 진입하여 주어진 주제와 다른 주제가 섞일 위험이 높아진다. e_CM(1)이 너무 크게 설정되면 초기에 커뮤니티 멤버들을 확보하지 못하여 원하는 크기의 잠재적 커뮤니티를 추출하지 못한 상태에서 확장이 중단될 수 있다. 따라서 합리적으로 e_CM(1)을 설정하는 방안이 필요하다.

제안된 방법을 이용하여 주어진 주제에 해당되는 커뮤니티 멤버와 포스트를 올바르게 추출하기 위해서 확장 과정의 종료 기준이 필요하다. 본 논문에서 해결하고자 하는 문제는 주어진 주제에 해당되는 커뮤니티 멤버와 포스트를 모두 찾는 것이다. 따라서 확장 과정의 종료 기준은 추출하는 커뮤니티 멤버와 포스트의 수가 실제 블로그 월드에 존재하는 주어진 주제에 해당되는 블로거와 포스트들의 수와 일치 여부이다. 블로그 월드에서 주어진 주제에 해당되는 블로거들의 수는 일반적으로 파악하기 힘들다. 이는 블로거들이 여러 가지 주제에 관심을 동시에 가지기 때문이다. 그러나 주어진 주제에 해당되는 포스트들의 수는 비록 대략적이기는 하나 주어진 주제에 해당되는 볼로거들의 수를 파악하는 것보다 쉽다. 이는 블로그 월드를 관리하는 기업이 블로거들에게 포스트들을 추천하고자 전문가들을 통해 일정 수의 포스트들의 주제를 파악하고 있기 때문이다. 따라서 샘플링과 같은 방법을 통해서 블로그 월드 내에 있는 주어진 주제에 해당되는 포스트들의 수를 대략적으로 파악할 수 있다. 본 논문에서는 랜덤 샘플링을 통해 전체 포스트들에 대한 해당 주제 포스트들의 비율을 구하고 전체 포스트의 수에 그 비율을 곱해서 주어진 주제에 해당된다고 예상되는 포스트들의 수를 구한다. 그러나 볼르그 월드에서 주어진 주제에 해당된다고 예상되는 포스트들은 커뮤니티 포스트에 속한 포스트들과 같이 해당 주제와 관련 있다고 확신할 수 없다. 따라서 해당 주제에 관심을 가지는 커뮤니티 멤버에 의해서 한 번이라도 액션을 받은 커뮤니티 포스트 후보의 수와 주어진 주제에 해당된다고 예상되는 포스트들의 수를 비교한다.

그러나 임의의 확장 단계에서 커뮤니티 포스트의 수가 블로그 월드 내에 있는 주어진 주제에 해당되는 포스트들의 수와 일치할 때 알고리즘을 종료하는 것은 문제가 있다. 이는 포스트 커뮤니티 추출 시 주어진 주제에 해당된다고 판단되는 포스트가 커뮤니티 포스트에 차례로 한 개씩 들어오는 것이 아니라 한 번에 들어오기 때문이다. 따라서 종료 기준을 포스트의 일정 수에 맞추는 것이 아니라 일정 구간으로 두는 것이 좋은 방법이다. 단계별 확장 과정 중에서 커뮤니티 포스트의 수가 미리 정해진 일정 구간에 속하게 되면 종료하고자 한다. 본 논문에서는 일정 구간의 상한(UB)와 하한(LB)를 다음과 같이 설정한다. UB는 실제 해당 주제에 대한 포스트들의 수를 x라 하였을 때 (x+0.1x)로 설정하고 LB는 (x-0.1x)로 설정한다.

본 논문에서는 확장 단계의 종료 구간을 정했기 때문에 피드백(feedback)을 활용하여 e_CM(1) 값을 자동 설정하는 방법을 취한다. 구체적인 방법은 그림 5와 같다. 우선, e_CM(1)을 최대 가능한 값인 시드의 개수로 설정한다(줄 7). 이 경우 모든 시드 포스트들에 액션을

```
1  function ExtractBlogCommunityUsingFeedback() returns blog_community
2
3          /* Initialize */
4          UB ← (x+0.1x); //the estimated number of posts is x
5          LB ← (x-0.1x);
6          △ ← number of seed posts;
7          e_CM(1) ← number of seed posts;
8          blog_community ← ∅ // community member U community post
9          /* Adjust the initial threshold for selecting CM using feedback */
10         while flag is not 0
11                 flag ← ExtractBlogCommunity(e_CM(1), LB , UB, blog_community );
12                 △ ← △ / 2;
13
14                 if flag = -1 then e_CM ← e_CM - △;
15                 else if flag = 1 then e_CM ← e_CM + △;
16         end while
17
18         return blog_community;
19 end
```

그림 5 피드백을 이용한 초기 임계값 설정 과정

취한 블로거들의 수가 매우 적기 때문에 추천된 주제 관련 포스트 수에 이르지 못한 상태에서 확장이 종료되는 상황이 발생할 수 있다. 이와 같이 추정된 주제 관련 포스트 수에 이르지 못한 상태에서 확장이 종료된다는 피드백을 flag를 통하여 받으면(줄 11, flag=-1) e_CM(1)을 위한 현재 값을 직전의 값과 차이의 절반만큼 낮추고(줄 14) 단계별 확장 과정을 다시 시작한다. 역으로, 확장 과정에 의하여 추정된 주제 관련 포스트 수를 초과한다는 피드백을 flag를 통하여 받으면(줄 11, flag=1) 현재의 값을 직전의 값과 차이의 절반만큼 높이고(줄 15) 단계별 확장 과정을 다시 시작한다. 이와 같이 e_CM(1)을 피드백을 통해 자동적으로 설정함으로써, 최종 커뮤니티 멤버가 해당 주제라고 추천하는 포스트의 수(커뮤니티 포스트 후보의 수)가 주제와 관련된 포스트 개수의 추정치에 근접할 때 알고리즘이 종료될 수 있도록 종료 기준을 정할 수 있다. 그림 6은 블로그 커뮤니티 추출 알고리즘을 보여준다.

3.3 정확도 향상을 위한 기법들

하나의 블로그는 여러 다양한 주제에 관심을 가질 수 있다. 확장 단계에서 다양한 주제에 대해 관심을 가지는 블로거들이 커뮤니티에 진입되면, 주어진 주제와 관련이 없는 다른 주제에 관한 내용을 담은 포스트들이 블로그 커뮤니티에 진입될 수 있다. 마찬가지로, 블로그 메인 페이지에 소개되거나 최신 뉴스 정보를 담은 포스트가 많은 액션을 받게 됨으로써 커뮤니티 포스트로 선발되면, 주어진 주제에 관심이 없는 블로거들이 커뮤니티 멤버로 진입할 수 있다. 이 경우 단계별 확장 과정에서 해당 주제와 관련이 없는 다른 주제들로 커뮤니티가 바람직하지 않은 방향으로 확산될 우려가 있다. 본 절에서는 이러한 문제들을 해결하여 제안하는 방법의 정확도를 향상시키기 위한 기법들을 제안한다.

3.3.1 폴더 정보의 활용

하나의 블로그는 여러 주제에 관심을 가질 수 있다. 대부분의 블로거들은 자신의 관심 포스트들을 폴더 단위로 나누어 정리하며, 일반적으로 하나의 폴더는 하나의 주제와 연관된 포스트들을 저장한다. 이러한 특성을 활용하여 블로거 대신 폴더를 단위로 추출 알고리즘을 적용하면, 추출된 블로그 커뮤니티의 정확도를 높일 수 있다.

폴더를 하나의 블로거로 간주하는 것은 기존의 블로거-포스트 관계를 폴더-포스트 관계로 보는 것을 의미한다. 이를 위해 폴더를 구분하지 않은 블로거는 추출대상에서 제외시킨다. 폴더 구분이 없는 경우는 하나의 폴더에 여러 주제의 포스트들이 존재하여 확장 과정에 주어진 주제와 관련 없는 블로거나 포스트의 진입을 유발하기 때문이며, 실제로 폴더를 구분하지 않는 블로거는 매우 드문 것으로 나타났다.

본 논문에서는 비록 폴더를 커뮤니티 멤버로 추출하지만 필요에 따라서 폴더들을 소유한 각각의 블로거들을 해당 주제의 커뮤니티 멤버로 볼 수 있다. 일반적으로 블로거들은 여러 주제에 관심을 가지고 있기 때문에 블로거들이 단일 주제의 커뮤니티에만 속한다고 말하는 것은 자연스럽지 못하다. 따라서 폴더-포스트 관계를 이용해서 커뮤니티 멤버를 추출하면 블로거들이 여러 주제의 커뮤니티에 속하게 되기 때문에 보다 자연스러운 커뮤니티 추출이 가능해 진다.

3.3.2 액션의 순도의 활용

확장 단계에서 액션의 양만을 기준으로 커뮤니티 멤버와 커뮤니티 포스트를 선발하는 경우, 두 개 이상의

```
1  function ExtractBlogCommunity(e_CM(1), LB , UB, blog_community ) returns flag
2
3     /* Initialize */
4     CM(0) ← ∅;
5     CP(0) ← {seed posts}; /* refer to §3.2.1 */
6     i ←1;
7
8     /* Iteratively extract CM(i) and CP(i): refer to §3.2.2 */
9     loop do
10        /* Adjust e_CM(i) & select CM(i) */
11        if i > 1 then adjust e_CM(i);                    /* refer to §3.2.3 */
12        for each blogger b not in CM(i-1) do
13                if (total number of actions from b to CP(i-1)) ≧ e_CM(i)
14                        then add b to CM(i);
15
16        /* Adjust e_CP(i) & select CP(i) */
17        adjust e_CP(i);                                  /* refer to §3.3.3 */
18        for each post p not in CP(i-1) do
19                if (total number of actions from CM(i) to p) ≧ e_CP(i)
20                then add p to CP(i);
21
22        /* Check the termination criteria and set flag: refer to §3.3.4 */
23        if (CP(i)' = CP(i-1)') and (number of CP(i)') < LB then return -1;
24        else if (number of CP(i)') is between LB and UB
25                 then blog_community ← CP(i) ∪ CM(i) return 0;
26        else if (number of CP(i)') > UB then return 1;
27
28        i ←i +1;
29    end loop
30
31 end
```

그림 6 블로그 커뮤니티 추출 알고리즘

주제와 연관된 포스트들에 관심을 가지는 블로거들 또는 주제와 관련 없이 일반적으로 인기가 많은 포스트들이 선발될 우려가 있다. 이러한 문제점을 해결하기 위해서 커뮤니티 멤버나 커뮤니티 포스트를 선발하는 기준으로 기존의 액션 수뿐만 아니라 액션의 순도를 추가적으로 이용하는 기법을 제안한다. 커뮤니티 멤버 선발 시 액션의 순도는 임의의 블로거가 얼마나 해당 주제에 대해서만 관심을 가지는지를 나타내며, 아래와 같이 정의된다.

$$\text{블로그 액션의 순도} = \frac{\text{블로그가 커뮤니티 포스트에 가한 액션수}}{\text{블로그의 전체 액션수}}$$

마찬가지로, 커뮤니티 포스트의 선발 시에 포스트에 대한 액션의 순도는 다음과 같이 정의된다.

$$\text{포스트 액션의 순도} = \frac{\text{포스트가 커뮤니티 멤버에게 받은 액션수}}{\text{포스트가 받은 전체 액션수}}$$

액션의 순도에 대한 단계별 임계값은 액션의 수에 대한 임계값과 마찬가지 방법으로 피드백을 활용하여 자동 설정하는 방법을 취한다.

3.3.3 인기 포스트 제한

인기 포스트란 대중적으로 인기가 많아 액션수가 높은 포스트를 뜻한다. 예를 들어, 블로그 포털의 메인 페이지에 소개된 포스트는 평소에 해당 주제에 관심을 두지 않던 블로그거들도 액션을 가하는 경우가 많다. 이러한 인기 포스트가 단계별 확장 과정에 시드로 포함될 경우 주어진 주제에 해당되는 커뮤니티 멤버와 포스트를 올바르게 추출하기 어렵게 된다.

인기 포스트를 단계별 확장 과정에 포함되지 않도록 하기 위해서는 어떤 포스트를 인기 포스트로 인정할 것인지에 대한 기준이 필요하다. 본 논문에서는 블로그 포털의 메인에 등록되었던 포스트들에게 가해졌던 평균 액션수를 인기 포스트의 기준으로 설정한다. 그리고 이 이상이 되는 것을 단계별 확장 과정에서 커뮤니티 포스트에 포함시키지 않도록 제외시킨다. 그러나 최종 커뮤니티 포스트에는 주어진 주제에 적합한 인기 포스트가 포함되어야 한다. 따라서 단계별 확장에서 커뮤니티 포스트로 선택된 인기 포스트들을 일시적으로 포함시키지 않았다가 최종 커뮤니티 포스트에는 포함시킨다.

4. 성능 평가

본 장에서는 제안된 방법에 대한 정확도 평가로서 추출된 블로그 커뮤니티가 주제와 관련된 커뮤니티 멤버들과 커뮤니티 포스트들을 얼마나 포함하는가를 실험을 통해 검증한다. 제 4.1절에서는 실험 환경과 성능 평가 척도를 설명한다. 제 4.2절에서는 제안된 방법의 정확도를 다양한 주제를 대상으로 측정한다. 또한, 제안된 방법의 정확도를 수작업으로 분류된 커뮤니티의 정확도와 비교한다. 제 4.3절에서는 제안된 방법에서 사용된 기법들이 정확도에 미치는 영향을 측정한다.

4.1 실험 환경

실험을 위해 국내 블로그 서비스 중 하나인 네이버 블로그(blog.naver.com)에서 2006년 4월부터 수개월간 수집하여 익명으로 처리한 데이터를 사용하였다. 데이터에서 액션이 매우 적은 포스트는 단계별 확장 방안에서 거의 점수를 받을 확률이 없다. 따라서 본 실험에서는 포스트 중 받은 액션의 수가 3 이하인 포스트를 미리 제거하고 남은 포스트들과 이 들에 액션을 가한 블로거들을 대상으로 하였다.

주제에 따른 블로그 커뮤니티의 정확도를 측정하기 위해서 본 논문에서는 4개의 주제 '요리, 자동차, 영어, 축구'에 대하여 도메인 전문가의 도움을 받아 각각 80개의 시드 포스트를 선발하였다. 도메인 전문가는 제 3.2.1절의 시드 선발 기준에서 언급한 주제 적합성, 정보의 양과 질, 인기도를 기준으로 시드 포스트를 선발했다. 전체 포스트 개수 중 각 주제에 해당하는 포스트 개수의 비율을 네이버(사)의 통계 자료를 통해 구하여 추출 알고리즘의 종료기준인 실제 주제 관련 포스트 수의 예측치를 계산할 때 사용하였다.

제안된 알고리즘의 성능을 평가하는 척도로는 아래와 같이 추출된 커뮤니티의 전체 멤버 중 해당 주제에 관심을 가지는 커뮤니티 멤버가 얼마나 되는지를 평가하는 커뮤니티 멤버의 정확도와, 총 커뮤니티 포스트 중 해당 주제에 대한 커뮤니티 포스트가 얼마나 되는지를 평가하는 커뮤니티 포스트의 정확도를 이용하였다. 커뮤니티 멤버와 커뮤니티 포스트의 주제 관련 여부는 사람이 직접 평가하였다.

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

$$\text{커뮤니티 포스트의 정확도} = \frac{\text{주제와 관련된 커뮤니티 포스트 수}}{\text{커뮤니티 포스트 수}}$$

4.2 추출된 블로그 커뮤니티의 정확도 분석

제안하는 방법의 우수성을 보이기 위해 세 가지 비교 실험을 한다. 첫째, 다양한 주제가 주어졌을 때 제안하는 방법을 통해서 추출된 커뮤니티의 정확도를 측정한다. 둘째, 수작업으로 분류된 결과와 제안된 방법을 비교한다. 셋째, HITS를 기반으로 한 커뮤니티를 추출하는 방법[5]으로 블로그 커뮤니티를 추출한 결과와 제안된 방법으로 커뮤니티를 추출한 결과의 정확도를 비교한다.

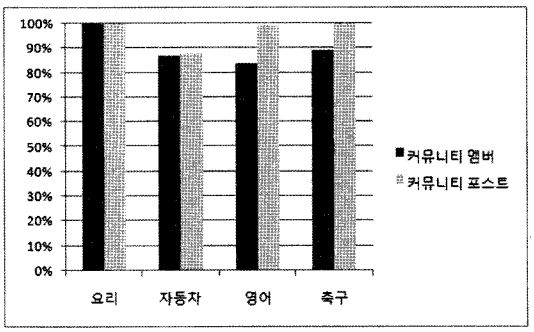
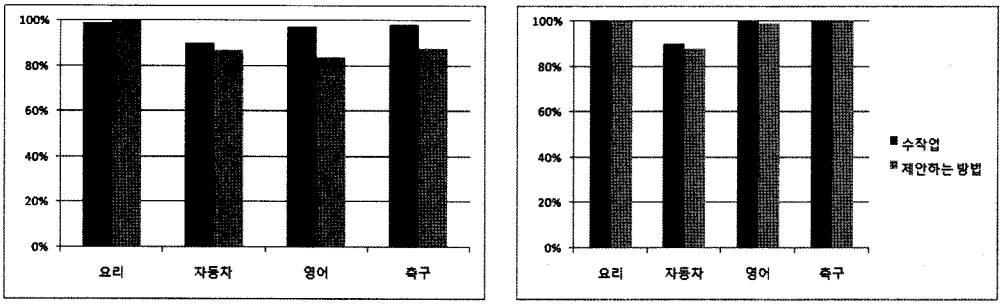


그림 7 제안하는 방법을 이용하여 추출한 블로그 커뮤니티의 정확도

그림 7은 제안하는 방법을 이용하여 각 주제별 커뮤니티를 추출한 커뮤니티 멤버와 커뮤니티 포스트의 정확도이다. 주제에 따른 각 방법들의 정확도를 살펴보면 액션 수가 높은 요리, 영어, 축구 주제에 대해서 높은 정확도를 보이는 반면, 액션 수가 적은 자동차 주제에 대해서는 정확도가 낮게 측정된다. 각 방법들은 액션을 기준으로 시드와 관련된 블로거와 포스트를 추출하기 때문에 액션의 수가 충분히 많지 않으면 노이즈에 영향을 쉽게 받아서 정확도가 낮아진다.

커뮤니티 멤버와 커뮤니티 포스트와의 정확도를 살펴보면 전반적으로 커뮤니티 포스트의 정확도가 커뮤니티 멤버에 비해 높게 측정된다. 이는 추출을 위한 분석 데이터 수집 시점과 정확도 판정 시점이 다르기 때문이다. 포스트는 시간이 지나도 내용이 거의 변하지 않지만 폴더는 시간이 지남에 따라 다른 주제의 포스트가 섞이는 경향을 보인다. 본 실험에서는 2년 전에 수집된 데이터를 대상으로 커뮤니티 추출을 하였으나, 정확도 판정은 현재 운용 중인 블로그에 직접 방문하여 수행하였다. 확인 결과, 주어진 주제라고 판단되는 폴더 내에서 주어진 주제와 관련이 없는 포스트들이 존재했으며, 폴더의 제목 역시 바뀌어 있었다. 또한, 일반적으로 포스트는 하나의 주제에 해당되지만, 하나의 폴더에는 여러 주제의 포스트가 섞일 수 있는 여지가 크다.

본 논문에서는 블로그 서비스 제공 회사에서 블로거와 폴더를 분류하기 위해 시행한 수작업 분류의 결과와 제안하는 방법의 결과를 비교하였다. 수작업 분류 결과의 정확도는 데이터 제공 업체의 메인 페이지에 카테고리 별로 분류된 블로거와 포스트를 랜덤 샘플링으로 추출하여 그 정확도를 측정하였다. 그림 8은 수작업 분류



(a) 커뮤니티 멤버 (b) 커뮤니티 포스트

그림 8 수작업 분류 체계와의 정확도 비교

체계와 정확도를 비교한 결과를 보여준다. 수작업 분류 체계의 정확도에 비하여 제안하는 방법의 정확도가 평균적으로 약 5% 정도 낮게 나타났다. 즉, 제안하는 방법을 사용한다면, 적은 노력으로 수작업에 근접하는 높은 정확도를 얻을 수 있음을 보여준다.

4.3 제안된 방법에 대한 심층 분석

앞 절에서 본 바와 같이 제안된 방법은 주제 관련 커뮤니티를 높은 정확도로 찾아낼 수 있다. 본 절에서는 제안된 방법에 대한 심층 분석을 하고자 한다. 첫째, 제안된 방법의 정확도를 단계별로 확인한다. 둘째, 본 논문에서 제안한 정확도 향상 기법들이 정확도에 미치는 영향을 평가한다. 셋째, 시드의 개수가 정확도에 미치는 영향을 살펴본다.

표 1은 주제별 블로그 커뮤니티의 단계별 추출 과정을 보여준다. 축구는 5단계, 영어와 요리는 4단계, 자동차는 3단계에서 확장이 종료되었다. 새롭게 추가되는 커뮤니티 멤버와 커뮤니티 포스트의 수는 단계별로 증가하며, 그 정확도가 지속적으로 유지됨을 볼 수 있다. 이것은 제안된 방법의 확장 과정이 신뢰할 수 있다는 것을 의미한다.

정확도 향상을 위해 제안된 기법들이 정확도에 미치는 영향을 평가하였다. 그림 9는 각 기법의 사용 여부에 따른 정확도의 변화를 보여준다. 그림 9에서 알 수 있듯이 첫째로 각 기법을 사용한 경우의 정확도가 사용하지 않은 경우의 정확도 보다 높다. 둘째로 폴더를 이용한 기법이 정확도 향상에 가장 큰 영향을 주었다. 셋째로 모든 기법들을 적용한 경우의 정확도가 가장 높다. 모든 기법을 적용한 결과가 가장 높은 정확도를 보였다.

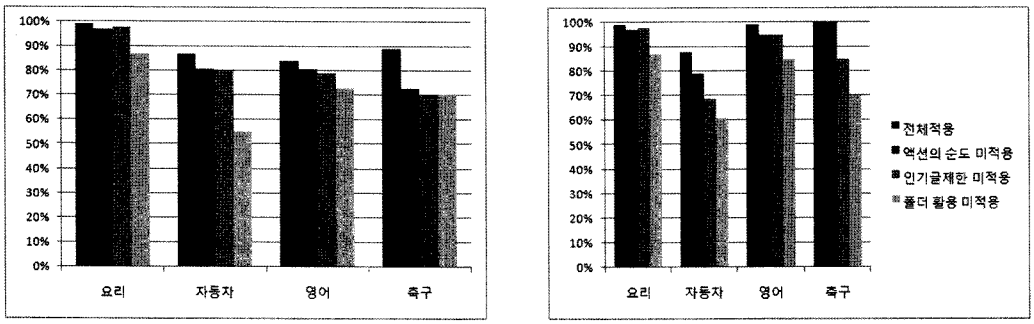
시드의 개수에 따른 정확도 평가는 적어도 몇 개의 시드가 있어야 정확도가 보장되는지, 시드가 많으면 많을수록 정확도는 높아지는지 알아보고자 하였다. 5개의 주제에 대하여 시드의 개수를 10개, 20개, 40개, 80개로 설정하고, 각 설정에 따른 커뮤니티 멤버의 정확도와 커뮤니티 포스트의 정확도를 측정하였다. 그림 10은 실험 결과를 나타낸 것이다. 실험 결과에 의하면 시드의 수가 적으면 정확도가 매우 낮고, 시드의 수가 증가할수록 정확도는 높아지나 시드의 수가 일정 이상 되면 정확도가 향상되는 정도가 상대적으로 약화되는 것으로 나타났다.

시드의 수가 많아지면 많아질수록 블로그 커뮤니티에 다른 주제의 포스트나 블로거가 들어오는 것을 방지되

표 1 주제별 블로그 커뮤니티 추출 단계에서의 정확도

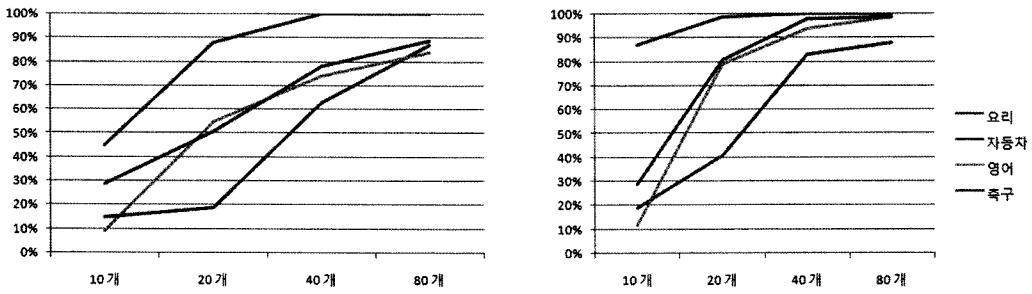
(개수(정확도))

	요리		자동차		영어		축구	
	커뮤니티 블로그	커뮤니티 포스트	커뮤니티 블로그	커뮤니티 포스트	커뮤니티 블로그	커뮤니티 포스트	커뮤니티 블로그	커뮤니티 포스트
1단계	433 (100%)	522 (100%)	66 (88%)	150 (94%)	197 (91%)	255 (100%)	45 (90%)	129 (97%)
2단계	1674 (100%)	1503 (100%)	103 (84%)	210 (91%)	654 (86%)	333 (100%)	79 (89%)	232 (99%)
3단계	3153 (100%)	2195 (100%)	110 (87%)	231 (88%)	711 (84%)	344 (100%)	143 (88%)	310 (99%)
4단계	3445 (100%)	2425 (100%)	-	-	713 (84%)	346 (99%)	287 (87%)	344 (99%)
5단계	-	-	-	-	-	-	323 (88%)	356 (99%)



(a) 커뮤니티 멤버 (b) 커뮤니티 포스트

그림 9 정확도 향상 기법들의 사용여부에 따른 정확도의 변화

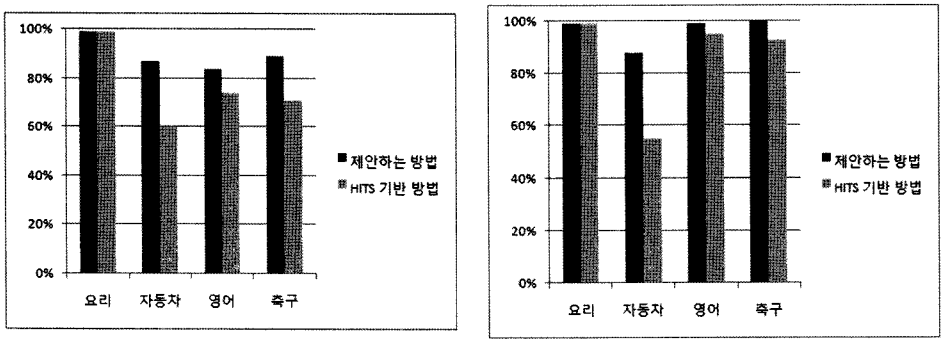


(a) 커뮤니티 멤버 (b) 커뮤니티 포스트

그림 10 시드의 개수에 따른 정확도의 변화

기 때문에, 충분히 많은 수의 시드 수를 확보하여야 한다. 그러나 전체 블로그 커뮤니티의 수에 비해서 실험을 통해서 살펴본 시드의 수는 대단히 적기 때문에 충분한 수의 시드 수를 전문가가 직접 선별하는 것은 비용이 많이 드는 작업이 아니다.

그림 11은 제안하는 방법과 HITS 기반 방법을 각각 이용하여 주제별 커뮤니티를 추출하였을 때 커뮤니티 멤버와 커뮤니티 포스트의 정확도이다. 제안하는 방법과 HITS 기반 방법의 정확도를 비교하여 살펴보면, 전반적으로 HITS 기반 방법에 비해 제안하는 방법의 정확도가 높게 나타난다. HITS 기반 방법은 대략적인 방법으로 시드와 연결된 후보 블로거와 포스트들을 추출한 후에 HITS 알고리즘을 이용해서 커뮤니티를 추출한다. 이러한 방법은 단지 시드 포스트와 링크로 연결되어 있는 블로거와 포스트를 주어진 주제와 같은 주제라고 판단하고 있기 때문에 정확한 블로그 커뮤니티 추출이 어렵다. 반면에 본 논문에서 제안하는 방법은 커뮤니티 멤버와 포스트가 확실하다고 판단되는 블로거와 포스트를



(a) 커뮤니티 멤버 (b) 커뮤니티 포스트

그림 11 제안하는 방법과 HITS 기반 방법의 정확도 비교

포함시켜나가기 때문에 다른 주제의 블로거와 포스트가 들어가는 것을 미연에 방지한다.

5. 결 론

블로그 월드에는 공통적인 주제의 내용을 담고 있는 포스트들의 집합과 이러한 포스트들의 집합에 관심을 표현한 블로거들의 집합이 존재한다. 본 논문에서는 이러한 잠재적 커뮤니티를 블로그 커뮤니티로 정의하고, 블로그 월드에서 주어진 주제와 관련된 잠재적 블로그 커뮤니티를 추출하는 방법에 대해서 논의하였다.

제안하는 방안은 주어진 주제를 담고 있는 소수의 핵심 포스트들을 시드로 정하고, 시드에 일정 이상의 액션을 가하는 블로거들과 그러한 블로거들에 의해 일정 이상의 액션을 받는 포스트들을 예측한 블로그 커뮤니티의 크기가 될 때까지 확대해서 주어진 주제와 관련된 최종적인 블로그 커뮤니티를 찾는다. 제안하는 방안은 특정 주제의 내용을 담고 있는 포스트들에 공통적으로 관심을 보이는 블로거들은 주어진 주제에 관심 있는 블로거 일 가능성이 높고, 이러한 블로거들이 공통적으로 관심을 보이는 포스트들이 주어진 주제의 내용을 담고 있을 가능성이 높다는 사실에 기반을 둔다. 본 논문에서는 제안하는 방안의 세부 기법들로 시드 포스트 추출 기준, 단계별 커뮤니티 멤버 및 커뮤니티 포스트 확장 방법, 단계별 임계값 조정 방법, 그리고 커뮤니티 멤버 선발 임계값의 초기값 설정 방법을 제안하였다. 또한, 제안하는 방안의 정확도 향상을 위해 액션의 순도 임계값, 폴더 정보의 활용, 그리고 인기 포스트 삭제에 대해서 논의하였다. 본 논문에서는 제안하는 방안의 우수성을 검증하기 위해서 다양한 실험을 수행하였다. 실험 결과, 제안하는 방법이 다양한 주제에 대해서 약 85%의 높은 정확도를 보였고 HITS 기반 방법보다 정확도가 높았으며 수작업으로 분류한 결과와 비교해도 정확도에 큰 차이가 없었다. 또한, 단계별 정확도 측정과 정확도 향상 기법이 정확도에 미치는 영향 그리고 시드의 개수에 따른 정확도 측정을 통해서 제안하는 방법을 심층 분석하였다.

향후 연구로서 포스트에 포함된 '내용'과 사용자의 액션을 함께 분석함으로써 추출된 블로그 커뮤니티의 정확도를 높이는 방안과 주제별 커뮤니티를 활용한 포스트 랭크의 정확도를 높이는 방안 등을 고려하고 있다.

참 고 문 헌

[1] B. Wellman, "Community: From Neighborhood to Network," *Communications of the ACM*, vol.48, no.10, pp.53-55, 2005.

[2] Y. R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng, "Discovery of Blog Communities based on Mutual Awareness," In *Proceedings of the 3rd Annual Workshop on the Weblogging Ecosysytem*, 2006.

[3] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trwaling the Web for Emerging Cyber-Communities," In *Proceedings of the 8th International Conference on World Wide Web*, pp. 1481-1493, 1999.

[4] P. A. Chirita, D. Olmedilla and W. Nejdl, "Finding Related Pages Using the Link Structure of the WWW," In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 632-635, 2004.

[5] J. Dean, and M. R. Henzinger, "Finding Related Pages in the World Wide Web," In *Proceedings of the 8th International Conference on World Wide Web*, pp.1467-1479, 1999.

[6] T. Murata, "Discovery of Web Communities Based on the Co-occurrence of References," In *Proceedings of the 3th International Conference on Discovery Science*, LNAI 1967, pp.65-75, 2000.

[7] T. Murata, "Discovery of Web Communities from Positive and Negative Examples," In *Proceedings of the 6th International Conference on Discovery Science*, LNAI 2843, pp.365-372, 2003.

[8] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient Identification of Web Communities," In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.150-160, 2000.

[9] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-Organization of the Web and Identification of Communities," *IEEE Computer*, vol. 35, no.3, pp.66-71, 2002.

[10] G. Greco, S. Greco and E. Zumpano, "Web Communities: Models and Algorithms," *World Wide Web: Internet and Web Information Systems*, pp. 59-82, 2004.

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

[12] N. Imafuhi and M. Kitsuregawa, "Effects of Maximum Flow Algorithm on Identifying Web Community," In *Proceedings of the 4th International Workshop on Web information and Data Management*, pp.43-48, 2002.

[13] H. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents," *Journal of the American Society for Information Science*, vol.24, no.4, pp.265-269, 1973.

[14] L. R. Ford and D. R. Fulkerson, "Maximal Flow through a Network," *Canadian Journal of Mathematics*, pp.399-404, 1956.

[15] R. Kumar, j. Novak, P. Raghavan, and A.

Tomkins, "On the Bursty Evolution of Blogspace," In *Proceedings of the 12th International Conference on World Wide Web*, pp.568-576, 2003.
[16] K. Ishida, "Extracting Latent Weblog Communities - A Partitioning Algorithm for Bipartite Graphs," In *Proceedings of the Second Annual Workshop on the Weblogging Ecosystem: Aggregation*, Analysis and Dynamics, 2005.
[17] A. Chin and M. Chignell, "A Social Hypertext Model for Finding Community in Blogs," In *Proceedings of the 7th Conference on Hypertext and Hypermedia*, pp.11-22, 2006.
[18] Y. R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng, "Blog Community Discovery and Evolution Based on Mutual Awareness Expansion," *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp.48-56, 2007.
[19] Y. Zhou and J. Davis, "Discovering Web Communities in the Blogspace," In *Proceedings of the 40th Annual Hawaiian International Conference on System Science (HICSS)*, 2007.
[20] I. S. Dhillon, "Co-clustering Documents and Words using Bipartite Spectral Graph Partitioning," In *Proceedings of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.269-274, 2001.
[21] T. Hu, H. Xiong, and S. Y. Sung, "Co-Preserving Patterns in Bipartite Partitioning for Topic Identification," In *Proceedings of 7th SIAM International Conference on Data Mining*, pp.509-514, 2007.
[22] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Neighborhood Formation and Anomaly Detection in Bipartite Graphs," In *Proceedings of the 5th IEEE International Conference on Data Mining*, pp.418-425, 2005.

신 정 환

2007년 2월 한양대학교 컴퓨터공학과 졸업(학사). 2009년 2월 한양대학교 전자컴퓨터통신대학원 졸업(공학석사). 2009년 2월~현재 슈어소프트테크 공통기술팀 연구원으로 재직중. 관심분야는 데이타마이닝, 사회연결망 분석, 소프트웨어 테스팅, 정적 분석 기법

윤 석 호

2005년 성결대학교 컴퓨터공학과 졸업(학사). 2007년 한양대학교 정보통신대학원 졸업(공학석사). 2007년~현재 한양대학교 대학원 전자통신컴퓨터공학과 박사과정 재학중. 관심분야는 사회연결망분석, 인터넷 포탈 데이타 분석, e-비지니스, 데이타 마이닝

김 상 욱

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. 관심분야는 데이타베이스 시스템, 저장 시스템, 트랜잭션 관리, 데이타 마이닝, 멀티미디어 정보 검색, 공간 데이타베이스/GIS, 주기억장치 데이타베이스, 이동 객체 데이타베이스/텔레매틱스, 사회 연결망 분석, 웹 데이타 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, 옥션, 온라인 사회 연결망, 네트워크 가격정책