논문 2009-46CI-3-6

블로그 공간에서의 링크 기반 클러스터링 방안

(Link-Based Clustering in Blogosphere)

송 석 순*, 윤 석 호*, 김 상 욱**

(Suk-Soon Song, Seok-Ho Yoon, and Sang-Wook Kim)

요 약

본 논문에서는 블로그 공간에 존재하는 블로거와 포스트들을 클러스터링하고자 한다. 먼저 블로그 공간의 블로거와 포스트들을 각각 하나의 타입으로, 블로거와 포스트 사이의 액션을 링크로 사상한다. 다음으로, 블로그 공간의 클러스터링을 위하여 블로그 환경에 가장 적합하고 효율적인 링크 기반 클러스터링 방법인 LinkClus를 선택한다. 정확한 클러스터링을 위하여 두 가지 방법을 제시한다. 첫 번째는 클러스터의 대상을 여러 주제에 관심을 가지는 블로거 대신 하나의 주제만을 나타내는 폴더로 한다. 두 번째는 노이즈의 발생 가능성을 높이는 링크가 아주 적은 블로거와 포스트를 클러스터링 과정에서 제외시킨다. 실험을 통하여 제안하는 방안을 이용한 클러스터링 결과가 내용적으로도 유사한지 검증한다.

Abstract

This paper addresses clustering of blogs and posts in blogosphere. First, we model blogosphere as a social network where blogs and posts correspond to nodes and interactions on posts by blogs corresponds to links. Next, for clustering in blogosphere, we employ LinkClus, a link based algorithm that finds clusters of nodes in a network effectively and efficiently. For more accurate clustering, we propose two refinements: (1) change of granularity from blogs to folders, and (2) removal of blogs and posts being highly likely to incur noises. Finally, we verify the effectiveness of the proposed approach by showing how the posts and blogs in the same cluster are similar to one another in terms of their contents.

Keywords : 데이터 마이닝, 링크 기반 클러스터링, 블로그 공간

Ⅰ. 서 론

최근 들어, 블로그(blog) 사용의 활성화로 인해 사용자들이 관심을 가지는 주제가 매우 다양해졌다. 또한 이러한 사용자들의 증가로 인해 게시글(post)들의 주제 역시 다양해졌다. 이에 따라 블로그에는 유사한 주제에 관심을 가지는 블로그 사용자들과 유사한 주제를 나타내는 게시글들이 존재하여 이를 클러스터링할 수 있다. 블로그 사용자들과 게시글들을 클러스터링하게 되면 이를 이용하여 다양한 응용분야에 활용할 수 있다[1]. 예를 들어 유사한 주제에 관심을 가지는 사용자들에 대한 타겟 마케팅, 유사한 주제의 게시글들을 이용한 게시글 분류기, 유사한 주제에 관심을 가지는 사용자들과 유사한 주제의 게시글들의 관계를 이용한 추천 시스템 등에 활용할 수 있다[2]. 따라서 유사한 주제에 관심을 가지는 블로그 사용자들과 유사한 주제를 나타내는 게시글들을 클러스터링하는 것은 매우 중요하다.

블로그 사용자들은 관심 있는 게시글에 스크랩, 덧글, 엮인글 등의 액션을 취할 수 있다. 이로 인해 유사한 주제에 관심을 가지는 사용자들은 공통된 게시글들에 액션을 취했을 것이고, 유사한 주제를 나타내는 게시글들

* 정회원, ** 평생회원, 한양대학교 전자컴퓨터통신공학과
(Department of Electronics and Computer Engineering, Hanyang University)

※ 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.
※ 이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국과학재단의 부분적인 지원을 받아 수행된 연구임.(No. R01-2008-000-20872-0)
접수일자: 2009년3월27일, 수정완료일: 2009년5월4일

은 공통된 사용자들에게 액션을 받았을 것이다. 따라서 블로그 사용자들이 게시글에 취한 액션을 이용하여 클러스터링하면 유사한 주제에 관심 있는 사용자들과 유사한 주제를 나타내는 게시글을 찾을 수 있다.

블로그 사용자들과 게시글들은 서로 다른 타입의 객체들로 표현할 수 있고 사용자와 게시글 사이의 액션은 링크로 표현할 수 있다. 이는 본 논문에서 하고자 하는 액션을 이용한 클러스터링이 링크 기반 클러스터링으로 해결될 수 있다는 것을 의미한다. 링크 기반 클러스터링이란 객체들 간에 존재하는 링크 정보만을 가지고 객체들을 클러스터링하는 방법이다[3].

기존에 대표적인 링크 기반 클러스터링 방법에는 Co-Citation[4], Bibliographic Coupling[5], SimRank[6], ReCoM[7], LinkClus[3] 등이 있다. Co-Citation과 Bibliographic Coupling은 두 객체의 유사도를 두 객체가 직접적으로 연결되어 있는 객체들을 이용하여 계산한다[4~5]. SimRank는 두 객체의 유사도를 두 객체가 가리키는 모든 가능한 객체 쌍들의 유사도의 평균을 이용하여 재귀적으로 계산한다[6]. ReCoM은 같은 타입 객체들 간의 링크와 서로 다른 타입 객체들 간의 링크를 동시에 이용하여 객체들을 클러스터링한다[7]. LinkClus는 SimRank의 개념을 그대로 이용하여 유사도를 계산한다. 그러나 객체 쌍들의 유사도를 계층구조를 이용하여 계산하기 때문에 SimRank보다 성능 측면에서 더 우수하다[3]. 본 논문에서는 블로그 공간에 가장 적합하다고 판단된 LinkClus를 이용하여 블로그 공간을 클러스터링하고자 한다.

본 논문에서는 블로그 공간을 클러스터링하기 위해 블로그 공간을 이분 그래프(bipartite graph)로 모델링한다. 모델링한 이분 그래프는 LinkClus의 클러스터링 환경으로 쉽게 사상할 수 있다. 또한, 클러스터링의 정확도를 높이기 위하여 두 가지 방법을 제시한다. 첫 번째는 여러 가지 주제에 관심을 가지는 블로거 대신에 하나의 주제에 관심을 가지는 폴더를 이용하는 방법이다. 두 번째는 링크가 적어 노이즈로 간주되는 객체를 제거하는 방법이다.

본 논문에서는 제안하는 방법으로 블로그 공간을 클러스터링한 결과의 정확도를 판정한다. 제안한 방법을 모두 적용하였을 때 클러스터링의 정확도는 90.7%로 높게 측정되었다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 본 연구의 문제정의를 하고, Ⅲ장에서는 본 연구와 관련된 기존의 연구들을 기술한다. Ⅳ장에서는 LinkClus의 개념적 설명과 처리 과정에 대해 설명한다. Ⅴ장에서는 LinkClus의 블로그 적용 방안에 대해 다루고, Ⅵ장에서는 블로그 공간에서 LinkClus를 이용한 클러스터링 실험 결과를 보인다. Ⅶ장에서는 결론 및 향후 연구를 제시한다.

Ⅱ. 문제정의

블로그는 사용자가 자신의 글을 온라인상에 저장할 수 있는 일종의 개인 홈페이지이다[8]. 블로거는 블로그를 사용하는 사용자를 의미하며 포스트는 블로거가 작성하여 블로그에 저장한 글을 의미한다. 블로거들은 관심 있는 포스트에 대해 스크랩, 덧글, 엮인글 등의 액션을 취할 수 있다. 본 논문에서는 블로거들과 포스트들 그리고 블로거와 포스트 사이에 존재하는 액션들로 구성된 공간을 블로그 공간이라 한다[9]. 그림 1은 블로그 공간의 간단한 예를 나타낸다. 원은 블로그를 나타내며 블로그 안의 사각형은 포스트를 나타낸다. 화살표는 블로거가 포스트에 취할 수 있는 대표적 액션인 스크랩, 덧글 등을 나타낸다. 최근 들어 블로그를 이용하는 사용자들이 증가하면서 블로그 공간을 대상으로 한 다양한 연구가 진행되고 있다[10~14].

블로그 공간에는 다양한 주제에 관심을 가지는 블로거들과 다양한 주제를 나타내는 포스트들이 있다. 이로 인해 유사한 주제에 관심을 가지는 블로거들과 유사한 주제를 나타내는 포스트들을 클러스터링 하는 것은 다양한 응용분야에 활용할 수 있기 때문에 매우 중요하다. 예를 들어, 유사한 사용자들에 대한 타겟 마케팅, 유사한 포스트들을 이용한 포스트 분류기, 유사한 사용자들과 유사한 포스트들의 관계를 이용한 추천 시스템 등에 활용할 수 있다.

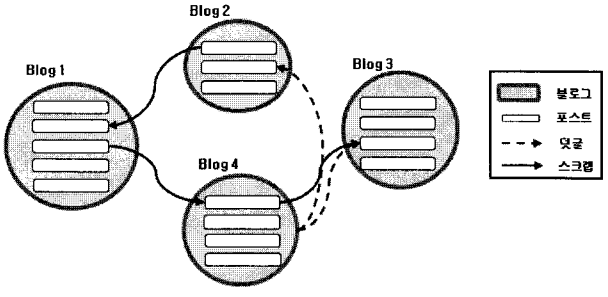


그림 1. 블로그 공간의 예
Fig. 1. Blogosphere: an example.

블로그 공간을 클러스터링 하기 위해서는 블로거가 포스트에 취한 액션을 이용해야 한다. 블로거는 관심 있는 포스트에 스크랩, 덧글 등의 액션을 취한다. 따라서 유사한 주제에 관심을 가진 블로거들은 공통된 포스트들에게 액션을 취했을 것이고 유사한 주제의 포스트들은 공통된 블로거들에게 액션을 받았을 것이다. 이렇게 액션을 기반으로 클러스터링한 블로거들은 유사한 행동패턴을 보였기 때문에 공통된 주제에 관심을 가질 것으로 기대한다. 또한 액션을 기반으로 클러스터링한 포스트들은 유사한 블로거들이 공통으로 관심을 가지기 때문에 동일한 주제의 내용을 담고 있을 것으로 기대한다. 본 논문에서는 블로거와 포스트 사이의 액션을 이용하여 블로그 공간을 클러스링하고자 한다.

블로그 공간의 블로거들과 포스트들은 서로 다른 타입의 객체들로 표현하고 블로거와 포스트 사이의 액션은 링크로 표현할 수 있다. 이는 해결하고자 하는 문제를 링크 기반 클러스터링 문제로 변환할 수 있다는 것을 의미한다. 본 논문에서는 링크 기반 클러스터링 방법을 이용하여 블로그 공간을 클러스터링하고자 한다.

III. 관련 연구

기존에 대표적인 링크 기반 클러스터링 방법에는 Co-Citation, Bibliographic Coupling, SimRank, ReCoM, LinkClus 등이 있다[2~6]. Co-Citation은 두 객체의 유사도를 두 객체가 공통적으로 가리키는 객체들의 수가 얼마나 많은가로 계산한다. 반대로 Bibliographic Coupling은 두 객체의 유사도를 두 객체를 동시에 가리키는 객체들의 수가 얼마나 많은가로 계산한다. 그림 2는 서로 다른 객체 타입인 블로거와 포스트 그리고 그 사이에 존재하는 링크를 나타낸 것이다. 그림 2에서 Co-Citation의 경우 블로거 B1과 B2의 유사도를 계산할 때 두 블로거가 공통적으로 가리키는 객체의 수를 이용하여 계산한다. Bibliographic Coupling은 B1과 B2의 유사도를 계산할 때 두 저자를 공통적으로 가리키는 객체의 수를 이용하여 계산한다. 두 방법은 객체들이 직접적으로 가리키는 객체들만을 이용하여 유사도를 계산하기 때문에 정확한 유사도를 계산하기 어렵다[6].

SimRank는 두 객체의 유사도를 두 객체가 가리키는 모든 가능한 객체 쌍들의 유사도의 평균을 이용하여 재귀적으로 계산한다. 그림 2에서 SimRank의 경우 B1과

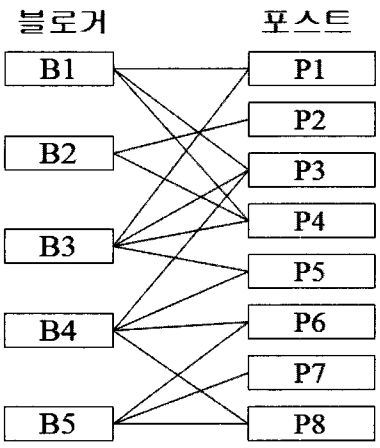


그림 2. 블로거, 포스트, 액션의 예
Fig. 2. Bloggers, posts, and a actions between them.

B2의 유사도는 두 블로거가 가리키는 객체 쌍들의 유사도 평균으로 계산된다. 따라서 B1이 가리키는 3개의 포스트와 B2가 가리키는 2개의 포스트로 만든 6개의 포스트 쌍들의 유사도 평균으로 계산된다. 여기서 포스트 쌍들의 유사도는 포스트가 가리키는 블로거 쌍들의 유사도 평균으로 재귀적으로 계산된다. 따라서 객체들이 직접적으로 가리키는 객체들뿐만 아니라 간접적으로 가리키는 객체들까지도 고려하여 유사도를 계산하기 때문에 보다 정확한 유사도를 구할 수 있다. 그러나 모든 객체 쌍들 간의 유사도를 구해야 하기 때문에 성능상의 문제점이 있다.

ReCoM은 같은 타입 객체들 간의 링크와 서로 다른 타입 객체들 간의 링크를 동시에 이용하여 클러스터링한다. ReCoM은 같은 타입 객체들 간의 링크를 이용하여 클러스터링을 한 다음 다른 타입 객체들 간의 링크를 이용하여 클러스터링의 정확도를 향상시킨다. 이 과정에서 모든 객체 쌍들 간의 유사도를 구하지 않고 클러스터 간의 유사도만을 계산하기 때문에 SimRank와 비교하여 성능측면에서는 우수하나 정확도가 낮다.

LinkClus는 SimRank의 개념을 그대로 이용하여 유사도를 계산하기 때문에 SimRank 수준의 정확한 유사도를 구할 수 있다. 그러나 계산과정에서 두 객체가 가리키는 모든 가능한 객체 쌍들의 유사도를 계층구조를 이용해서 계산하기 때문에 성능측면에서 더 효율적이다.

블로그 공간은 규모가 매우 방대하다. 따라서 블로그 공간의 블로거들과 포스트들을 클러스터링하기 위해서는 성능측면에서 우수한 ReCoM 또는 LinkClus가 적합하다. 그러나 블로그 공간의 특성상 ReCoM을 사용하기는 어렵다. ReCoM은 클러스터링을 위하여 같은 타입

객체들 간의 링크가 필요하지만 블로그 공간에서는 포스트와 포스트 사이의 링크는 거의 존재하지 않는다. 또한 정확도 측면에서도 LinkClus가 모든 객체들 간의 유사도를 계산하기 때문에 ReCoM보다 더 정확한 결과를 보인다. 본 논문에서는 정확도가 높고 성능이 가장 우수하며 블로그 공간의 구조에 가장 적합한 LinkClus를 선택하여 블로그 공간의 블로거들과 포스트들을 클러스터링하고자 한다.

Ⅳ. LinkClus

LinkClus는 SimRank와 같이 두 객체의 유사도를 두 객체가 가리키는 모든 가능한 객체 쌍들의 유사도의 평균을 이용하여 계산한다[6]. 그러나 LinkClus는 모든 객체간의 유사도를 계산하는 SimRank의 방법을 개선하기 위하여 계층적으로 객체간의 유사도를 표현한 SimTree 구조를 제안했다.

그림 3은 SimTree구조를 나타낸다. 각각의 SimTree는 하나의 객체 타입을 나타내며 SimTree간의 점선은 서로 다른 타입 객체들 간의 링크를 나타낸다. 말단노드들은 각 타입의 객체들을 의미하며 비 말단노드들은 유사한 노드들의 집합인 클러스터를 의미한다. SimTree는 계층구조로 되어 있기 때문에 어떤 레벨에 있는 노드를 클러스터로 간주할 것인가에 따라서 클러스터의 수를 결정할 수 있다. 본 논문에서는 레벨 1 노드들을 하나의 클러스터로 사용한다.

SimTree는 같은 부모에 속한 노드들 간의 유사도만을 저장한다. 같은 부모에 속하지 않은 노드들 간의 유사도는 SimTree에 저장하지 않고 해당 객체들의 조상들 사이의 유사도 이용해서 계산된다. 따라서 모든 노드 쌍들 간의 유사도를 계산하지 않고 SimTree의 계층구조를 이용해서 계산하기 때문에 성능측면에서 효율적이다.

LinkClus가 최초의 SimTree를 구축할 때에는 객체들 간의 유사도가 계산되어져 있지 않은 상태이다. 따라서 LinkClus는 최초의 SimTree를 구축하기 위해 직접적으로 연관되어 있는 노드들을 그룹화한다. 이러한 초기 SimTree구축을 위해 LinkClus는 빈발패턴을 이용하여 직접적으로 연관된 노드들을 찾는다[1, 15~16]. 이렇게 찾은 연관 노드들을 이용하여 그룹화한다. 한 번의 그룹화 과정이 끝나면 같은 그룹에 속하게 된 노드들은 동일한 부모 노드를 가지게 된다. 이 부모 노드는 자식

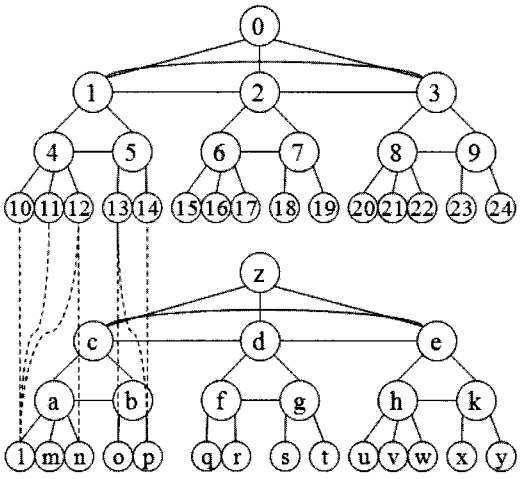


그림 3. SimTree의 구조[12]
Fig. 3. A SimTree structure.

노드들의 링크를 모두 다 가지고 있는 그룹의 대표 노드이다. 다음 그룹화 단계에서는 이 부모 노드를 대상으로 그룹화가 이루어진다. 이러한 그룹화 과정은 부모 노드의 수가 특정 수 이하가 될 때까지 반복적으로 수행된다.

각 타입의 초기 SimTree를 구축하고 나면 SimTree내에 있는 객체들 간의 유사도를 Co-Citation과 같은 방법으로 다른 타입 SimTree와의 링크를 이용하여 계산한다. 이는 노드들이 간접적으로 가리키는 노드들 간의 유사도는 반영이 되지 않은 상태이다. 따라서 LinkClus는 간접적으로 가리키는 노드들 간의 유사도를 반영하기 위해 각 SimTee내에 있는 노드들 간의 유사도를 다른 타입의 SimTree내에 있는 노드들 간의 유사도를 참조하여 갱신한다. 이렇게 갱신된 유사도를 이용하여 SimTree내의 노드들은 더 유사한 부모노드에 포함되도록 위치를 이동한다. 그림 5에서 만약 노드 1이 부모노드 a에 대한 유사도보다 부모노드의 형제노드인 b에 대한 유사도가 더 크다면 b의 자식노드에 포함되도록 위치를 이동한다. 이러한 유사도 계산과정과 SimTree의 구조변경과정을 반복함으로써 최종적인 SimTree가 구해진다.

Ⅴ. LinkClus의 블로그 공간 적용 방안

1. 블로그 공간 모델링

본 논문에서는 블로그 공간을 이분 그래프로 표현한다. 서로 다른 타입의 블로거들과 포스트들을 서로 다른 노드들의 집합으로 표현하고 블로거와 포스트 사이

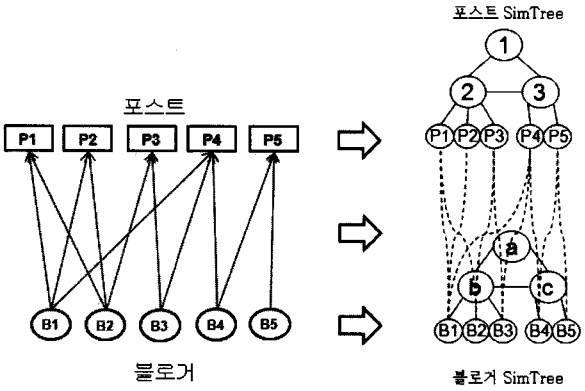


그림 4. 이분그래프를 SimTree로 사상한 예
Fig. 4. Modeling of a bipartite-graph using SimTrees.

의 액션을 링크로 표현한다.

이분 그래프로 표현된 블로그 공간은 LinkClus의 SimTree로 쉽게 사상할 수 있다. 즉 이분 그래프의 두 노드 집합을 두 개의 SimTree로 사상하고 두 노드 집합 사이의 링크를 두 SimTree사이의 링크로 사상할 수 있다. 그림 4는 블로그 공간을 표현한 이분 그래프를 LinkClus의 SimTree로 사상한 예이다. 서로 다른 객체 타입인 포스트와 블로거가 두 개의 SimTree로 사상되었고 블로거와 포스트 사이의 액션이 SimTree사이의 액션으로 사상되었다.

2. 폴더 이용

일반적으로 블로거는 여러 주제에 관심을 보인다. 즉, 블로거와 링크로 연결되어 있는 모든 포스트들이 동일한 주제를 나타내지 않는다. 따라서 블로거와 링크로 연결되어 있는 모든 포스트들을 그대로 이용하여 블로거를 클러스터링하게 되면 정확한 클러스터링 결과를 얻을 수 없다.

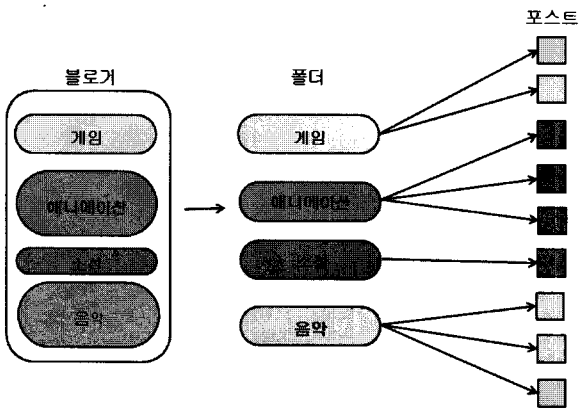


그림 5. 블로거를 폴더로 세분화한 예
Fig. 5. Refining blogs with folders.

블로그 공간 내에는 블로거들이 주제에 따라 포스트들을 분류해놓은 폴더가 존재한다. 본 논문에서는 정확한 클러스터링을 위하여 블로거-포스트 관계가 아닌 폴더-포스트 관계를 이용하고자 한다. 그림 5은 하나의 블로거를 여러 개의 폴더로 세분화한 예이다. 그림 5의 원은 블로거를 나타내며, 원안의 둥근 사각형은 폴더를 나타낸다. 사각형은 포스트를 나타내며, 화살표는 블로거가 포스트에 취한 액션을 나타낸다. 그림 5와 같이 하나의 블로거는 게임, 에니메이션, 소설, 음악 등의 다양한 주제에 관심을 가질 수 있기 때문에 블로거 대신에 세분화한 폴더를 클러스터링하여 정확도를 높이고자 한다.

3. 링크가 적은 객체 제거

블로그 공간에는 링크가 아주 적은 블로거와 포스트들이 많이 존재한다. 이러한 포스트나 블로거들은 두 객체간의 유사도 계산 시 링크가 적기 때문에 실제로 유사하지 않으면서도 유사도가 아주 높게 계산되어질 수 있다. 예를 들어 링크가 하나씩 밖에 없는 두 블로거가 우연히 동일한 포스트에게 액션을 취했다면 LinkClus의 유사도 계산방식에 의해 두 블로거의 유사도는 1로 매우 높게 계산된다. 그러나 이러한 경우는 두 블로거의 링크가 적기 때문에 높게 계산된 유사도를 신뢰할 수가 없다. 따라서 본 논문에서는 링크가 k이하인 블로거와 포스트를 노이즈로 간주하여 클러스터링 과정에서 제외시킨다. 본 논문에서는 링크가 1이하인 블로거와 포스트를 노이즈로 간주한다.

링크가 1인 블로거와 포스트를 제외시키기 위해 각 타입에서 링크가 1인 객체들을 반복적으로 제외시킨다. 그림 6은 이러한 노이즈 객체를 제외시키는 과정을 나타낸 그림이다[17]. 먼저, 링크가 1인 포스트 P1, P2, P7 세 개의 객체를 제거한다. 세 개의 객체를 제외시키고 나면 블로거 B2 역시 링크가 1이 된다. 따라서 B2 객체 역시 클러스터링 과정에서 제외시킨다.

VI. 실 험

1. 실험환경

블로그 공간에 적용한 LinkClus는 액션을 기반으로 블로거들과 포스트들을 클러스터링하였다. 따라서 같은

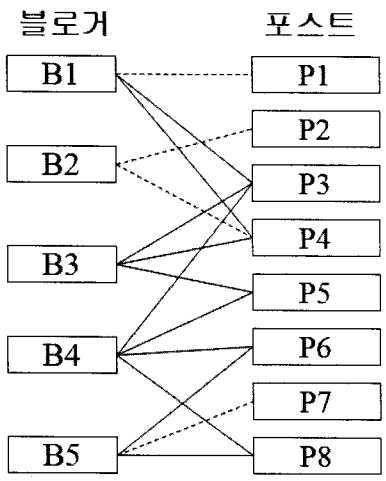


그림 6. 링크가 1인 객체의 제거
Fig. 6. Removal of objects having only one link.

포스트 클러스터에 속한 포스트들의 내용이 동일하지 않을 수 있다. 그러나 유사한 블로거들이 공통으로 액션을 취한 포스트들이기 때문에 내용적으로도 유사할 것임을 기대한다. 따라서 본 논문에서는 실험을 통하여 LinkClus로 클러스터링된 포스트들의 주제가 얼마나 동일한지 알아보고자 한다.

실험을 위해 국내 블로그 서비스 중 하나인 네이버 블로그에서 2006년 4월부터 수개월간 수집하여 익명으로 처리한 데이터를 사용하였다. 데이터의 크기는 폴더 240,000개, 포스트 150,000개, 링크 800,000개의 데이터를 사용하였으며 링크가 1인 객체들은 모두 제거하고 남은 데이터들을 대상으로 클러스터링 하였다.

2. 태그를 이용한 정확도 측정 방법

클러스터링된 포스트들의 주제가 동일한가를 정확히 확인하기 위해서는 포스트들의 내용을 사람이 일일이 확인해야 한다. 그러나 블로그 공간의 규모를 고려했을 때 사람이 직접 내용을 확인하는 방법은 실질적으로 불가능하다.

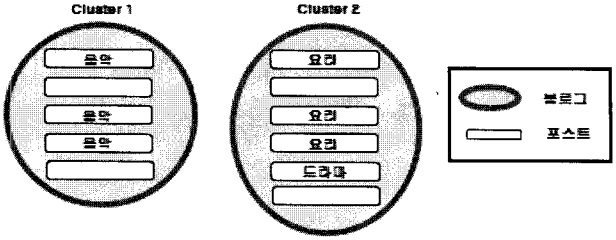


그림 7. 클러스터링 결과의 예
Fig. 7. A clustering result.

블로그 공간에는 태그 데이터가 존재한다. 태그는 블로거가 포스트를 작성할 시 포스트의 주제어라고 판단하여 작성한 데이터이다. 만약 두 포스트들의 태그가 일치한다면 두 포스트들의 주제가 동일하다는 것을 의미한다. 따라서 본 논문에서는 같은 클러스터에 있는 포스트들의 주제가 얼마나 동일한지를 측정하기 위해서 태그를 이용한다. 다음의 식이 태그를 이용한 클러스터링의 정확도를 나타낸다.

$$\text{클러스터링의 정확도} = \frac{\text{태그가 일치하는 포스트쌍의 수}}{\text{태그가 있는 포스트쌍의 수}}$$

그림 7은 클러스터링 결과의 예이다. 그림 7에서 원은 클러스터를 의미하며 원안에 속한 사각형은 포스트를 의미한다. 사각형 안에 있는 글은 포스트의 태그를 의미하며 글이 없는 사각형은 태그가 없는 포스트를 의미한다. 그림 7에서 1번 클러스터에 속한 태그가 있는 포스트 쌍의 수는 3개이며 그 중 일치하는 포스트쌍의 수는 3개이다. 2번 클러스터에 속한 태그가 있는 포스트쌍의 수는 6개이며 그 중 일치하는 포스트쌍의 수는 4개이다. 클러스터링의 결과가 이와 같을 때 클러스터링의 정확도는 (3+4)/(3+6) 으로 78%이다.

3. 정확도 실험 결과

표 1은 LinkClus를 이용하여 블로그 공간의 포스트들을 클러스터링한 결과의 정확도를 나타낸다. 본 논문에서 제안한 모든 방법을 이용하여 클러스터링한 결과의 정확도는 90.7%이다.

일반적으로 같은 주제의 포스트들은 같은 태그를 가진다. 그러나 태그는 블로거가 임의로 정한 주제어이기 때문에 같은 주제의 포스트들에 다른 태그가 붙어있을 수 있다. 예를 들어 음악이라는 동일한 주제의 두 포스트 중 하나는 음악이라는 태그를 가지지만 다른 하나는 노래라는 태그를 가질 수 있다. 그림 8은 이러한 태그

표 1. 블로그 공간의 포스트들을 클러스터링한 결과의 정확도
Table 1. Accuracy on post clustering.

방안 / 측도	블로거 이용	폴더 이용	폴더이용 + 링크 1이하 제거
정확도	74.3%	84.5%	90.7%

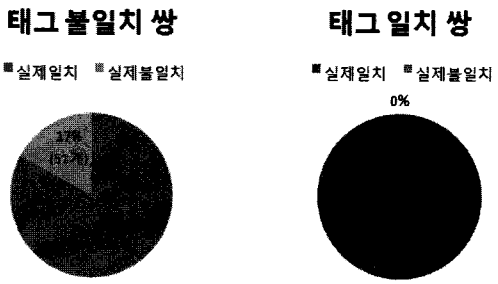


그림 8. 태그를 이용한 정확도 측정방법의 실제 정확도 조사 결과

Fig. 8. In-depth examination on accuracy using tags.

를 이용한 정확도 측정방법이 실제 정확도보다 낮음을 보여준다. 임의로 선택한 태그가 일치하는 300개 포스트쌍의 경우 100% 실제 내용이 일치했다. 또한 임의로 선택한 태그가 일치하지 않는 300개의 포스트쌍 중에서도 83%가 실제 내용이 일치했다. 따라서 실제 클러스터링의 정확도는 측정된 결과보다 더 높을 것으로 판단된다. 따라서 블로그 공간의 포스트들을 액션을 기반으로 클러스터링한 결과가 내용적으로도 유사함을 확인할 수 있다.

Ⅶ. 결론 및 향후 연구

본 논문에서는 블로그 공간에서의 링크 기반 클러스터링 방법에 대해 연구하였다. 블로거와 포스트들을 클러스터링하기 위하여 기존의 링크 기반 클러스터링 방법 중에서 LinkClus가 블로그 공간에 가장 적합하다는 것을 보였다. 또한 LinkClus를 블로그 공간에 적용하는 방안을 논의했다. LinkClus는 서로 다른 타입의 객체를 객체와 객체사이에 존재하는 링크를 통하여 클러스터링한다. 따라서 블로그 공간에 LinkClus를 적용하기 위해서 블로거와 포스트를 각각 하나의 타입으로 사상했고 블로거와 포스트 사이의 액션을 링크로 사상했다. 또한 정확한 클러스터링를 위하여 두 가지 방법을 제시했다. 첫 번째 방법은 여러 주제에 관심을 가지는 블로거 대신 하나의 주제만을 나타내는 폴더 이용하는 방법이고, 두 번째 방법은 노이즈인 적은 링크를 가진 블로거와 포스트를 제거하는 방법이다. 제안한 방안으로 블로그 공간의 포스트들을 클러스터링한 결과가 내용상으로도 주제가 일치하는지 실험을 통하여 검증했다.

블로그 공간을 클러스터링한 결과를 다양한 응용 분야에 활용할 수 있다. 공통된 주제에 관심을 보이는 블로거 클러스터는 타겟 마케팅에 이용될 수 있으며, 공통된 주제를 담고 있는 포스트 클러스터는 포스트 추천 시스템에 이용될 수 있다.

참 고 문 헌

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[2] S. Gardner, *Buzz Marketing With Blogs for Dummies*, John Wiley & Sons Inc, 2005.

[3] X. Yin, J. Han, and P. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links," *In Proc. Int'l. Conf. on Very Large Data Bases*, pp. 427-438, 2006.

[4] 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.

[5] M. Kessler, "Bibliographic Coupling Between Scientific Papers," *Journal of the American Documentation*, Vol. 14, No. 1, pp. 10-25, 1963.

[6] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity," *In Proc. Int'l. Conf. on Special Interest Group on Knowledge Discovery and Data*, pp. 538-543, 2002.

[7] J. Wang et al., "ReCoM: Reinforcement Clustering of Multi-type Interrelated Data Objects," *In Proc. Int'l. Conf. on Special Interest Group on Information Retrieval*, pp. 274-281, 2003.

[8] NHN(주), http://blog.naver.com, 2009.

[9] Wikipedia, blog, http://en.wikipedia.org/wiki/Blog, 2009.

[10] S. Herring et al., Conversations in the Blogosphere: An Analysis "From the Bottom Up," *In Proc. of the 38th Annual Hawaii Int'l. Conf. on System Scicences*, pp. 107b, 2005.

[11] Y. Lin, "Blog Community Discovery and Evolution based on Mutual Awareness Expansion," *In Proc. Int'l. Conf. on Web Intelligence*, pp. 48-56, 2007.

[12] K. Fujimura, T. Inoue, and M. Sugisaki, "The Eigenrumor Algorithm for Ranking Blogs," *In Proc. Int'l. Conf. on World Wide Web*, 2005.

[13] D. Gruhl et al., "Information Diffusion Through Blogspace" *In Proc. Int'l. Conf. on World Wide Web*, pp. 491-501, 2004.

[14] 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.

[15] J. Wang and J. Han, "CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets," *In Proc. Int'l. Conf. on Special Interest Group on Knowledge Discovery and Data,* pp. 236-245, 2003.
[16] N. Pasquier et al., "Discovering Frequent Closed Itemsets for Association Rules," *In Proc. Int'l. Conf. on Database Theory,* pp. 398-416, 1999.
[17] R. Kumar et al., "Trawling the Web for Emerging Cyber-Communities," *In Proc. Int'l. Conf. on World Wide Web,* pp. 1481-1493, 1999.

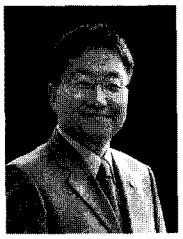
저 자 소 개

송 석 순(정회원)
2008년 한양대학교 정보통신학부 졸업 (학사).
2008년~현재 한양대학교 대학원 전자통신컴퓨터공학과 석사과정 재학중

<주관심분야 : 사회연결망분석, 인터넷 포탈 데이타 분석, e-비지니스, 데이타 마이닝>

윤 석 호(정회원)
2005년 성결대학교 컴퓨터공학과 졸업 (학사).
2007년 한양대학교 정보통신 대학원 졸업 (공학석사).
2007년~현재 한양대학교 대학원 전자통신컴퓨터공학과 박사과정 재학 중.

<주관심분야 : 사회연결망분석, 인터넷 포탈 데이타 분석, e-비지니스, 데이타 마이닝>



김 상 욱(평생회원)
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

<주관심분야 : 데이터베이스 시스템, 저장 시스템, 트랜잭션 관리, 데이터 마이닝, 멀티미디어 정보 검색, 공간 데이터베이스/GIS, 주기억장치 데이터베이스, 이동 객체 데이터베이스/텔레매틱스, 사회 연결망 분석, 웹 데이터 분석>