블로그 공간에서의 링크 기반 클러스터링 방안

송석순, 윤석호, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: gagler@agape.hanyang.ac.kr

Link-Based Clustering in Blogosphere

Suk-Soon Song, Seok-Ho Yoon, Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 블로그 공간에 존재하는 블로거와 포스트들을 링크 기반 클러스터링을 통해 클러스터링하고자 한다. 먼저 기존 링크 기반 클러스터링 방안 중에서 블로거와 포스트들을 클러스터링하는데 가장 적합한 LinkClus를 선택한다. LinkClus를 블로그 공간에 적용하기 위해서 블로거와 포스트를 각각 하나의 타입으로, 블로거와 포스트 사이의 액션을 링크로 사상한다. 정확한 클러스터링을 위하여 클러스터의 대상을 여러 주제에 관심을 가지는 블로거 대신 하나의 주제만을 나타내는 폴더로 한다. 또한 노이즈의 발생 가능성을 높이는 링크가 아주 적은 블로거와 포스트를 클러스터링 과정에서 제외시킨다. 실험을 통하여 제안하는 방안을 이용한 클러스터링 결과가 내용적으로도 유사한지 검증한다.

1. 서론

블로그는 사용자가 자신의 글을 온라인상에 저장할 수 있는 일종의 개인 홈페이지이다. 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[1]. 최근 들어 블로그를 이용하는 사용자들이 증가하면서 블로그 공간을 대상으로 한 다양한 연구가 진행되고 있다[2][3][4].

블로그 공간에는 공통된 포스트에 액션을 취하는 블로거들과 공통된 블로거들에게 액션을 받는 포스트들이 존재한다. 이러한 블로거들과 포스트들은 유사한 액션 패턴을 가지고 있기 때문에 액션을 기반으로 클러스터링 할 수 있다. 이렇게 클러스터링된 블로거와 포스트 클러스터들은 다양하게 활용될 수 있다. 첫째, 블로거 클러스터는 유사한 블로거들에게 공통적으로 관심을 가지는 포스트를 추천하는 추천시스템에 활용될 수 있다. 둘째, 포스트 클러스터는 유사한 포스트들을 미리 정의된 카테고리에 분류하는 포스트 분류기에 활용 가능하다.

블로거들과 포스트들은 서로 다른 타입의 객체들로 표현할 수 있고 블로거와 포스트 사이의 액션은 링크로 표현할 수 있다. 이는 본 논문에서 해결하고자 하는 문제를 링크 기반 클러스터링 문제로 변환할 수 있다는 것을 의미한다. 링크 기반 클러스터링이란 객체들 간에 존재하는 링크 정보만을 가지고 객체들을 클러스터링하는 방법이다.

기존에 대표적인 링크 기반 클러스터링 방법에는 Co-Citation[5], Bibliographic Coupling[6], SimRank[7], ReCoM[8], LinkClus[9] 등이 있다. 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. 두 방법은 객체들이 직접적으로 가리키는 객체들만을 이용하여 유사도를 계산하기 때문에 정확한 유사도를 계산하기 어렵다[7]. SimRank는 두 객체의 유사도를 두 객체가 가리키는 모든 가능한 객체 쌍들의 유사도의 평균을 이용하여 재귀적으로 계산한다. 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[9]. 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 성능적으로 가장 효율적이고 정확도 측면에서도 우수하며 블로

그 공간의 구조에 가장 적합한 LinkClus를 선택하여 블로그 공간의 블로거들과 포스트들을 클러스터링하고자 한다.

2. LinkClus

LinkClus는 두 객체의 유사도를 두 객체가 가리키는 모든 가능한 객체 쌍들의 유사도의 평균을 이용하여 계산한다[7]. LinkClus는 모든 객체간의 유사도를 계산하는 기존 방법을 개선하기 위하여 계층적으로 객체간의 유사도를 표현한 SimTree 구조를 제안했다.

그림 1은 SimTree구조를 나타내며 SimTree는 객체 타입 수만큼 생성한다. 즉 그림 1은 서로 다른 두 개의 객체 타입에 대응하는 두 개의 SimTree를 의미한다. SimTree간의 점선은 서로 다른 타입의 객체들 간의 링크를 의미한다. 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. 따라서 본 논문에서는 레벨 1 노드들을 하나의 클러스터로 정의한다.

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.

LinkClus는 초기에 SimTree를 구축하기 위해 빈발패턴(frequent pattern)[10]을 이용한다. 같은 빈발패턴에 속하게 되는 객체들은 공통된 링크를 많이 가지는 객체들이므로 이 노드들을 유사한 노드들로 보고 같은 부모노드에 속하게 한다. 초기 SimTree내에 있는 객체들 간의 유사도는 Co-Citation과 같은 방법으로 계산된다. 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. 또한 갱신된 유사도를 이용하여 SimTree는 노드와 클러스터의 위치를 이동시켜 구조를 변경한다. 이러한 과정을 반복함으로써 최종적인 SimTree가 구해진다.

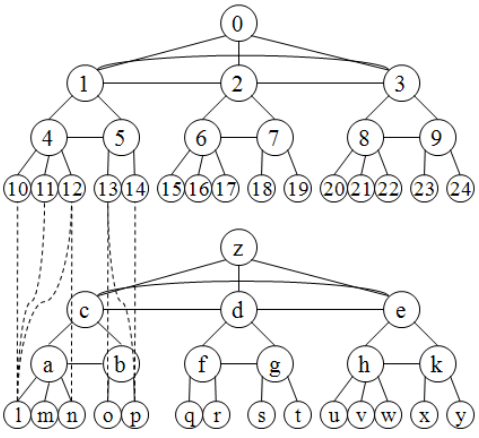

그림 1. SimTree의 구조[9].

3. LinkClus의 블로그 공간 적용 방안

3.1 블로그 공간 모델링

본 논문에서는 블로그 공간을 이분 그래프(bipartite graph)로 표현한다. 서로 다른 타입의 블로거들과 포스트들을 서로 다른 노드들의 집합으로 표현하고 블로거와 포스트 사이의 액션을 링크로 표현한다. 그림 2는 블로그 공간을 이분 그래프로 표현한 예이다.

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.

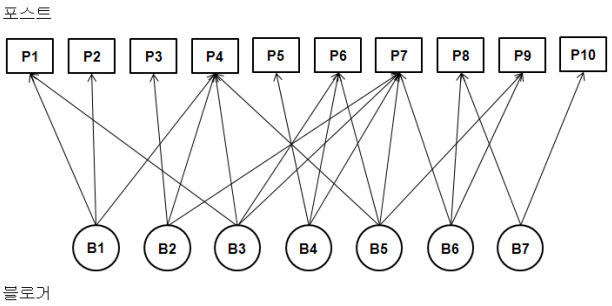

그림 2. 이분 그래프로 표현한 블로그 공간.

3.2 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.

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 3은 하나의 블로거를 여러 개의 폴더로 세분화한 예이다.

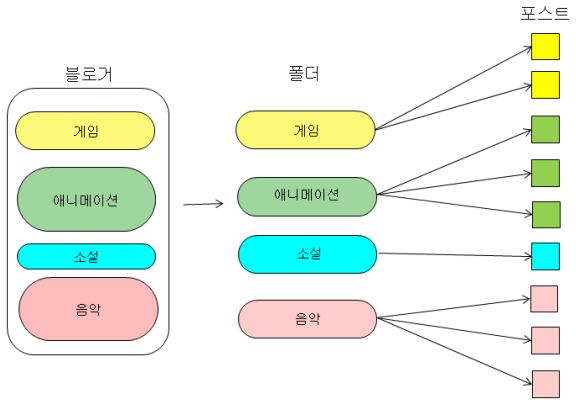

그림 3. 블로거를 폴더로 세분화한 예.

3.3 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 포스

트에게 액션을 취했다면 두 블로거의 유사도가 매우 높게 계산될 수 있다. 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.

4. 실험

블로그 공간에 적용한 LinkClus는 액션을 기반으로 블로거들과 포스트들을 클러스터링하였기 때문에 동일한 클러스터에 있는 포스트들의 내용이 동일한 주제라고 확신할 수 없다. 그러나 유사한 블로거들에게 공통으로 액션을 받은 포스트들이기 때문에 내용적으로 유사해 주제가 동일할 것이라고 기대한다. 따라서 본 논문에서는 LinkClus로 클러스터링된 포스트들의 주제가 동일한지 실험을 통하여 알아보고자 한다.

실험을 위해 국내 블로그 서비스 중 하나인 네이버 블로그에서 2006년 4월부터 수개월간 수집하여 익명으로 처리한 데이터를 사용하였다.

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. 따라서 본 논문에서는 같은 클러스터에 있는 포스트들의 주제가 얼마나 동일한지를 측정하기 위해서 태그를 이용한 클러스터의 정확도를 정의한다. 태그는 블로거가 포스트를 작성할 시 포스트의 주제어라고 판단하여 작성한 데이터를 의미한다. 즉, 만약 두 포스트들의 태그가 일치한다면 두 포스트들의 주제가 동일하다는 것을 의미한다.

$$\text{클러스터의 정확도} = \frac{\text{태그가 일치하는 포스트쌍의수}}{\text{태그가있는 포스트쌍의수}}$$

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. 따라서 실제 클러스터링의 정확도는 측정된 결과보다 더 높을 것으로 판단된다.

5. 결론

본 논문에서는 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.

<표 1> 블로그 공간의 포스트들을 클러스터링한 결과의 정확도

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

감사의 글

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.

참고문헌

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

[2] 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.

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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