논문 연결망의 특성 분석

배 덕호, 하 지운, 홍 지원, 김 상욱
한양대학교 전자컴퓨터통신공학과
e-mail: {dhbae, oneofus, lenas.laster, wook}@agape.hanyang.ac.kr

Analyzing a Scientific Literature Network

Duck-Ho Bae, Jiwoon Ha, Ji-Won Hong, Sang-Wook Kim
Dept. of Electronics Computer Engineering, Hanyang University

요 약

최근 다양한 논문 검색 사이트들이 등장함에 따라 논문 데이터에 대한 관심이 높아지고 있다. 본 논문에서는 논문 데이터 연결망을 다양한 실험을 통해 분석한다. 더 나아가, 기존의 웹 연결망, 블로그 연결망과의 특성 비교를 통해 연결망 간의 차이를 보이고, 이러한 차이가 나타나는 원인을 규명한다.

1. 서론

논문 연결망은 웹 연결망, 블로그 연결망 등과 더불어 대표적인 온라인 사회 연결망 중 하나이다. 최근, Google Scholar, Microsoft Libra, DBLP 등의 논문 검색 사이트들이 많이 등장함에 따라 논문 데이터의 확보가 용이해졌으며, 이에 따라 논문 연결망을 분석하여 유용한 정보들을 도출하고자 하는 시도들이 많아지고 있다. 이러한 논문 데이터는 웹이나 블로그 데이터와는 달리 노이즈와 abuse가 적은 데이터로서 더욱 각광 받고 있다.

논문 데이터는 논문을 노드로, 참조 정보를 에지로 하는 그래프로 모델링 가능하다[3][5]. 이 때, 논문 A가 논문 B를 참조했다면, 논문 A에서 논문 B로의 방향성 에지를 추가한다. 본 논문에서는 이러한 그래프를 논문 연결망이라 정의한다. 논문은 (1) 자신 이전에 발행된 논문만을 참조 가능하며, (2) 한 번 발행된 논문은 수정이 불가능하다. 따라서 모델링된 논문 연결망은 사이클이 존재하지 않는 DAG (directed acyclic graph) 형태를 띈다.

위에서 언급한 논문의 특성으로 인해 논문 연결망은 기존의 웹 연결망이나 블로그 연결망과는 많은 차이를 보이며, 이로 인해 웹 연결망이나 블로그 연결망의 분석 기법들이 논문 연결망에서는 효율적으로 동작하지 않을 것임을 충분히 예상할 수 있다.

기존 연구들에 의하면 논문의 특성이 실제로 기존의 연결망의 분석 기법들에 문제를 야기하고 있음을 알 수 있다. 첫째, 논문 랭킹에 있어, 최신 논문은 예전 논문에 비해 참조 받을 기회가 적으므로 예전 논문들이 상대적으로 높은 랭킹을 갖는 기득권 현상이 발생한다[3]. 웹과 블로그 연결망에서도 유사한 경향이 나타나지만, 참조의 수정이 불가능한 논문 연결망에서는 해당 경향이 더욱 심하게 나타난다. 둘째, 논문 데이터베이스에는 오래된 논문들이 누락되어 있거나 최신 논문들을 참조하는 논문들에 대한 정보는 누락된 경우가 많다[5]. 따라서 링크 기반 유사도 계산에 있어, 관련성이 높은 예전 (최신) 논문과 예전 (최신) 논문 간의 유사도를 정확하게 계산하지 못하는 문제점이 발생한다[5]. 이러한 문제점 역시 기존의 웹과 블로그 연결망에서도 나타나지만, DAG 형태를 갖는 논문 연결망에서 더욱 심하게 나타난다.

그 동안 웹 연결망과 블로그 연결망의 특성을 분석한 연구는 많이 이루어져 왔다[1][2]. 그러나 기존의 사회 연결망과는 다른 구조적 특성을 보일 것으로 예상된 논문 연결망의 특성 분석은 그 중요도에 비해 아직 미흡한 상태이다. 본 논문에서는 다양한 실험을 통해 논문 연결망의 특성을 분석한다. 더 나아가, 기존의 웹 연결망, 블로그 연결망과의 비교를 통해 연결망 간의 특성의 차이를 보이고, 이러한 차이가 나타나는 원인을 규명한다.

2. 논문 데이터

본 논문에서는 논문 연결망의 분석을 위해 2009년 3월에 다운로드한 DBLP 데이터[4]를 사용하였다. DBLP 데이터에는 참조 정보가 포함되어 있지 않으므로, Libra 데이터를 통해 논문들 간의 참조 정보를 수집하였다. 수집된 데이터의 논문은 총 1,071,973편이고, 논문 당 평균 참조 수는 7.69개이다.

표 1. 사회 연결망들의 노드, 평균 에지 수 비교

	웹 연결망 [1]	블로그 연결망[2]	논문 연결망
노드 수	203,549,046	3,693,788	1,071,973
평균 에지 수	7.17	7.13	7.69

3. Degree 분포 분석

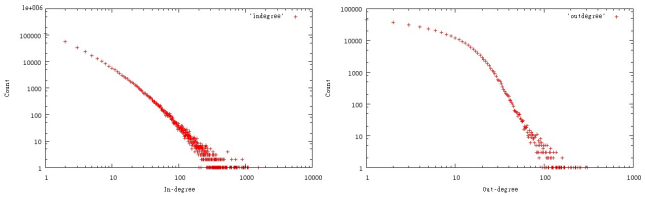
본 장에서는 논문 연결망의 노드의 degree 분포를 분석하였다. 척도 x와 척도 y가 식 1을 만족할 때, 척도 x와 y는 power-law 분포를 따른다고 정의한다[2]. 이 때, x는

in-degree (out-degree), y는 해당 in-degree (out-degree)를 갖는 노드의 수, a와 e는 상수이다.

$$y = \alpha x^{e} \quad \text{(식 1)}$$

연결망의 degree 분포가 power-law 분포를 따르는 현상은 큰 규모의 연결망에서 공통적으로 나타나는 특성으로, 이러한 연결망을 척도 없는 연결망 (scale-free network)라고 한다[1][2].

그림 1은 논문 연결망의 degree 분포를 나타낸다. 논문 연결망의 in-degree와 out-degree 분포 모두 power-law 분포를 따르는 것으로 나타났다. 그러나 out-degree 분포에서 degree가 5 ~ 50인 노드들의 분포는 power-law 분포를 벗어나는 것으로 나타났다. 이는 한 편의 논문은 한정된 지면과 관련 연구들로 인해 대부분 5 ~ 50편 정도의 다른 논문을 참조하며, 100편 이상의 다른 논문을 참조하는 논문의 수는 매우 적기 때문이다.



(a) In-degree. (b) Out-degree.

그림 1. 논문 연결망의 degree 분포.

4. Bow-Tie 구조 분석

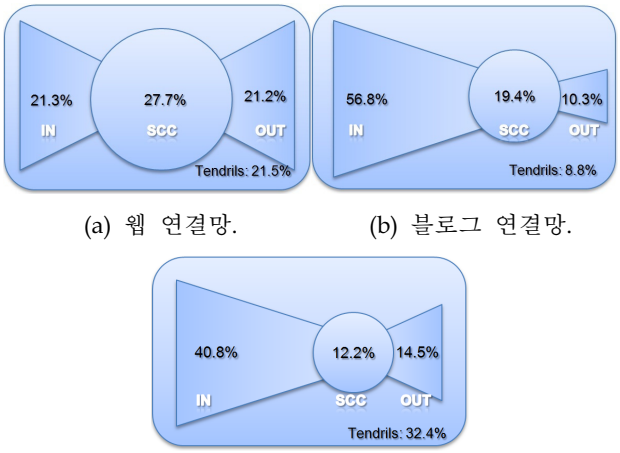
본 장에서는 논문 연결망을 거시적 관점에서 분석하기 위해 Bow-Tie 구조를 구성하고, 이를 웹 연결망과 블로그 연결망의 Bow-Tie 구조와 비교하였다.

Bow-Tie 구조는 크게 SCC, IN, OUT, Tendrils의 4가지 컴포넌트로 구성된다[1]. SCC (strongly connected component)의 모든 노드들은 SCC의 다른 모든 노드들로 도달 가능한 방향성 경로 (directed path)를 가진다. IN의 모든 노드들은 SCC로 도달 가능한 경로를 가지고 있지만, SCC에서 도달 가능한 경로는 가지고 있지 않다. OUT의 모든 노드들은 IN과는 반대로 SCC에서 도달 가능한 경로는 가지고 있지만, SCC로 도달 가능한 경로는 가지고 있지 않다. Tendrils의 모든 노드들은 SCC로 도달 가능한 경로와 SCC에서 도달 가능한 경로 모두 가지고 있지 않다.

Bow-Tie를 구성하기 위해서는 전체 연결망을 아우르는 Giant Component를 우선적으로 찾아야 한다. 논문 연결망에서는 전체 노드의 97%를 아우르는 Giant Component가 나타났다. 이는 각 분야 간에는 서로 교류가 있으며, 학제가 연구가 활발히 진행되고 있다는 것을 의미한다.

그림 2는 구성된 Bow-Tie 구조를 나타낸다. 논문 연결망에서도 Bow-Tie 구조가 나타났다. 이전에 발행된 논문만을 참조할 수 있는 논문의 특성상 논문 연결망에는 SCC가 존재하지 않아야 한다. 그러나 흥미롭게도 실제로는 이후에 발행되는 자신 혹은 공동 저자의 논문을 미리 참조하는 경우가 종종 발생하였으며, 이로 인해 SCC가 존재하였다.

논문 연결망은 IN이 OUT에 비해 3배 정도 컸으며, 전체적인 구조가 블로그 연결망과 유사하였다. OUT에 속하는 논문들의 대부분은 IN과 SCC에 속하는 논문들로부터 참조를 많이 받은 수준 높은 논문이다. IN에 속하는 논문들의 대부분은 참조를 거의 받지 못하고 참조를 하기만 하는 논문들이다. SCC에 속하는 논문들은 참조도 하고 받기도 한 논문들이다. 이렇듯, 수준높은 논문들의 수는 전체 논문 수에 비해 상대적으로 적으므로 OUT의 비율이 낮은 것을 알 수 있다. 논문 연결망의 경우, Tendrils의 비율이 웹 연결망과 블로그 연결망에 비해 높았다. 이는 논문 연결망의 경우, DAG 형태에 가까우며 이로 인해 SCC에 속하는 논문들이 적기 때문이다.



(a) 웹 연결망. (b) 블로그 연결망.

(c) 논문 연결망.

그림 2. Bow-Tie 구조 비교.

5. 결론

본 논문의 공헌은 다음과 같다. 첫째, 논문 연결망에서도 척도 없는 연결망에서 나타나는 공통적인 특성들이 나타남을 degree 분포 분석을 통해 보였다. 둘째, Bow-Tie 구조 분석을 통해 웹 연결망, 블로그 연결망과는 다른 논문 연결망만의 고유한 특성을 분석하고, 이러한 특성이 나타나는 원인을 규명하였다.

감사의 글

"본 연구는 지식경제부 및 정보통신산업진흥원의 IT융합 고급인력과정 지원사업의 연구결과로 수행되었음" (NIPA-2011-C6150-1101-0001)

참고문헌

[1] A. Broder et al., "Graph Structure in the Web," In *Computer Networks*, Vol. 33, No. 1, 2000.

[2] J. Ha et al., "Analyzing a Korean Blogosphere: A Social Network Analysis Perspective," In *ACM SAC*, 2011.

[3] W. Hwang, S. Chae, and S. Kim, "Yet Another Paper Ranking Algorithm Advocating Recent Publications," In *WWW*, 2010.

[4] M. Ley, "DBLP: Some Lessons Learned," In *VLDB*, 2009.

[5] S. Yoon, S. Kim, and S. Park, "A Link-based Similarity Measure for Scientific Literature," In *WWW*, 2010.