제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

학술 데이터베이스에서 논문 랭킹을 위한 방안들의 평가

채수민*, 황원석*, 김상욱*
*한양대학교 전자컴퓨터통신공학과
e-mail: aesem@hanyang.ac.kr

Evaluating Ranking Methods in a Scientific Literature Database

Soo-Min Chae*, Won-Seok Hwang*, Sang-Wook Kim*
*Dept. of Electronics and Computer Engineering, Hanyang University

요 약

본 논문에서는 논문 랭킹 방안의 바탕이 되는 개념을 이해하고, 그 개념을 바탕으로 기존 논문 랭킹 방안들에 대한 특징을 파악한다. 또한 각 방안의 정확도를 비교하여, 논문 랭킹의 정확도를 높이는 요인이 무엇인지 판단한다.

1. 서론

다양한 분야의 논문을 모아 거대한 학술 데이터베이스를 구축하고, 이를 검색하기 위한 서비스가 발전하고 있다. 특히, 학술 데이터베이스에 저장된 논문의 수가 방대하기 때문에 다수의 논문들이 결과로 검색될 수 있다. 따라서 검색 결과인 다수의 논문들을 연구자가 필요로 할 것이라 생각되는 순서로 정렬해주는 논문 랭킹 방안이 필요하다.

기존 논문 랭킹 방안으로는 문헌 [1]에서 제안한 방안, PaperRank [2], Browsing-Based Model [3], PopRank [4], Co-Ranking [5], CiteRank [6], 문헌 [7]에서 제안한 방안이 있다. 그러나 이 방안들의 정확도를 정량화하여 비교한 연구는 아직까지 없었다.

본 논문에서는 실험을 통해 각 방안의 정확도를 측정하고, 이를 비교하여 더 높은 정확도를 보이는 방안들을 확인한다.

2. Random Walk with Restart

Random Walk with Restart (RWR)는 랭킹 방안 연구에서 널리 알려진 개념이다 [8]. RWR은 그래프 구조에서 랜덤 파티클(random particle)이 노드 사이를 이동하는 과정을 마르코프 연쇄(Markov Chain)로 설명한 개념이다. 랜덤 파티클은 하나의 노드에서 링크를 따라 연결된 다른 노드로 랜덤 워크(Random Walk)하거나, 링크와 무관하게 임의 노드에서 리스타트(Restart)하며 이동을 반복한다. RWR을 수식으로 나타내면 다음과 같이 표현된다.

$$\boldsymbol{r_{i+1}} = (1-\alpha)(C^T + \boldsymbol{w} \times \boldsymbol{d^T}) \times \boldsymbol{r_i} + \alpha \boldsymbol{w}$$

C는 랜덤 워크를 나타내는 확률행렬(stochastic matrix)로, 그래프의 인접행렬(adjacency matrix)을 열의 합이 1이 되도록 정규화한 행렬이다. $\boldsymbol{w}$는 리스타트를 나타내는 벡터로 각 인자 값은 각 노드에서 리스타트 할 확률에 비례한다. 일반적으로 리스타트 할 확률은 모든 노드에 동일하게 부여한다. $\boldsymbol{d}$는 댕글링(dangling) 노드를 표현하는 벡터이다. 그리고 α는 랜덤워크와 리스타트의 비율을 정하며, 일반적으로 0.15로 설정한다. $\boldsymbol{r_i}$는 i번째 단계에서의 논문들의 점수를 갖는 벡터이며 정상상태(stationary state)에 다다른 $\boldsymbol{r_i}$는 최종적으로 랜덤 파티클이 각 노드에 머무를 확률이다. 본 논문에서 소개하고자 할 논문 랭킹 방안들은 RWR을 바탕으로 한다.

3. 논문 랭킹 방안 소개

문헌 [1]에서 제안한 방안에서는 논문과 인용 관계를 통해 논문의 랭킹을 구한다. 논문을 노드로, 인용 관계를 링크로 하여 그래프를 모델링하고 RWR [8]을 이용하여 랭킹을 계산한다. 이 방안에서는 α를 일반적인 경우와 다르게 0.5로 하는 것을 제안한다. 이는 연구자가 논문을 찾을 때, 논문의 인용 관계를 많이 사용하지 않는다는 가정에 의해 결정된 값이다.

PaperRank [2] 또한 문헌 [1]에서 제안한 방안과 거의 유사하게 논문의 랭킹을 구하는 방안이다. 단, α를 수정 없이 0.15로 사용하고, 다른 그래프를 사용한다는 차이점이 있다. PaperRank는 인용 관계를 양방향성 링크(undirected link)로 하여 그래프로 표현하고, RWR을 이용하여 랭킹을 계산한다.

Browsing-Based Model [3]은 논문과의 인용 관계 뿐 아니라 동일 저자 여부까지 고려하여 논문의 랭킹을 계산하는 방안이다. 논문을 노드로 하고 논문 간의 인용 관계와 동일 저자 작성 여부를 링크로 하여 그래프로 표현한다. 또한 권위 있는 저자들이 작성한 논문에 높은 리스타트 확률을 부여하고, RWR을 통해 랭킹을 계산한다.

PopRank [4]는 논문과 저자, 학회 사이의 관계를 모두 고려하여 논문과 저자, 학회의 랭킹을 동시에 계산하는 방안이다. 논문과 저자, 학회를 노드로 하고 논문과 저자, 논문과 학회, 저자와 학회 사이의 인용, 작성, 게재 관계를 링크로 하여 3개의 이분 그래프(bipartite graph)로 각각 표현한다. 또한 이 방안에서 제안하는 일정한 규칙에 의해 서로 다른 타입의 노드를 이동하는 랜덤 파티클을 가정하고, RWR을 통해 각 노드의 랭킹을 계산한다. 이때, 각 노드의 인기도에 비례하게 리스타트 확률을 부여하는데, 각 노드의 인기도는 미리 정해진 기준을 따른다.

Co-Ranking [5]은 논문의 인용 관계뿐 아니라 저자와의 관계까지 고려하여 논문과 저자를 동시에 랭킹하는 방안

이다. 논문과 저자를 각 타입의 노드로 하고 논문 간의 인용 관계, 저자 간의 사회적 관계를 링크로 하여 그래프로 각각 표현한다. 또한 논문과 저자 사이의 작성 관계를 링크로 하여 이분 그래프로 표현한다. 이때, 사회적 관계는 공동 저자와 공동 게재 여부로 맺어진다. 이 방안에서도 PopRank와 유사하게 이 방안에서 제안한 일정한 규칙에 의해 동일한 타입의 노드와 서로 다른 타입의 노드를 이동하는 랜덤 파티클을 가정하고, RWR을 통해 각 노드의 랭킹을 계산한다. 단, 이 방안에서는 리스타트를 고려하지 않는다.

CiteRank [6]는 논문의 인용 관계뿐만 아니라 논문의 최신성까지 고려하여 논문의 랭킹을 구하는 방안이다. 논문을 노드로 하고 인용 관계를 링크로 하여 그래프로 표현한다. 이 방안에서는 논문의 최신성을 고려하기 위해 랜덤 파티클이 각 노드에서 출발할 확률을 출판년도에 따라 다르게 부여한다. 또한, 이 방안에서는 RWR의 매 상태(state)의 누적을 통해 랭킹을 구한다. Co-Ranking과 마찬가지로 리스타트를 고려하지 않는다.

문헌 [7]에서 제안한 방안은 논문의 인용 관계 뿐 아니라 논문의 최신성과 논문이 게재된 학회의 권위까지 고려하여 논문의 랭킹을 구하는 방안이다. 논문을 노드로 하고 인용 관계를 링크로 하여 그래프로 표현한다. 이 방안은 논문의 최신성과 학회의 권위를 랭킹에 부여하기 위해 이들을 통해 리스타트 확률을 수정하고, RWR을 통해 논문의 랭킹을 계산한다.

4. 실험

실험 데이터는 2009년 3월에 다운받은 DBLP 데이터를 사용하였으며, 인용 정보는 Libra를 통해 얻었다. 데이터는 총 1,071,973개의 논문과 각 논문의 저자와 학회 정보 그리고 논문 당 평균 7.67개의 인용 정보로 구성되었다.

실험은 데이터 마이닝과 연관된 키워드 7개를 질의어(질의어: "clustering", "sequential pattern mining", "graph pattern mining", "link mining", "spatial databases", "web mining", "multirelational data mining")로 선택하여 수행하였다. 각 질의어에 대해 각각의 랭킹 방안을 수행하여 상위 10개, 20개, 30개의 결과를 얻었다. 이 결과들이 유명 데이터 마이닝 책 [9]의 레퍼런스 목록에 포함된 정도를 정밀도(precision)를 통해 나타내고, 7개의 정밀도의 평균을 정확도로 간주하였다. 각 방안의 매개변수 값은 해당 방안을 제안한 논문에서 사용한 값을 그대로 사용하였다.

<표 1>은 위에서 소개한 총 7개의 논문 랭킹 방안에 대한 비교 결과이다. 상위 10개에서는 문헌 [7]에서 제안한 방안이 가장 높은 정확도를 보였고 Co-Ranking이 가장 낮은 정확도를 보였다. 또한 상위 20개에서는 PopRank와 문헌 [7]에서 제안한 방안의 정확도가 가장 높았고, Co-Ranking이 가장 낮은 정확도를 보였다. 상위 30개에서의 결과는 상위 10개의 결과와 유사한 양상을 보였다. 대체적으로 논문과 저자의 관계만을 고려한 Co-Ranking과 Browsing-Based Model 보다, 학회와의 관계까지 고려한 문헌 [7]에서 제안한 방안과 PopRank가 더 높은 정확도를 보임을 확인하였다.

<표 1> 논문 랭킹 방안 별 상위 10개, 20개, 30개의 정밀도의 평균

논문 랭킹 방안 \ 평가대상	상위 10개	상위 20개	상위 30개
문헌 [1]에서 제안한 방안	0.243	0.193	0.171
PaperRank [2]	0.243	0.179	0.162
BBM [3]	0.200	0.136	0.148
PopRank [4]	0.229	0.207	0.176
Co-Ranking [5]	0.157	0.121	0.114
CiteRank [6]	0.243	0.171	0.138
문헌 [7]에서 제안한 방안	0.257	0.207	0.186

5. 결론

본 논문은 기존에 제안된 주요 논문 랭킹 방안들에 대해 소개하고, 각 방안들의 정확도를 실제 데이터를 이용한 실험을 통해 정량화하여 비교하였다. 비교 결과, 저자 보다는 학회를 고려한 논문 랭킹 방안들의 정확도가 더 높음을 알 수 있었다. 이러한 특징은 논문 랭킹과 관련된 추후 연구에 도움이 될 것으로 판단된다.

감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업(NIPA-2010-(C1090-1011-0009))의 부분적인 지원과 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] P. Chen et al., "Finding scientific gems with Google's PageRank algorithm," *Journal of Informetrics*, Vol. 1, No. 1, pp. 8-15, 2007.

[2] M. Gori and A. Pucci, "Research Paper Recommender Systems: A Random-Walk Based Approach," In *Int'l. Conf. on Web Intelligence*, pp. 778-781, 2006.

[3] S. Yan and D. Lee, "Toward alternative measures for ranking venues: a case of database research community," In *JCDL*, pp. 235-244, 2007.

[4] Z. Nie et al., "Object-level ranking: bringing order to web objects," In *WWW*, pp. 567-574, 2005.

[5] D. Zhou et al., "Co-Ranking Authors and Documents in a Heterogeneous Network," In *ICDM*, pp. 739-744, 2007.

[6] D. Walker et al., "Ranking scientific publications using a simple model of network traffic," *Journal of Statistical Mechanics*, 2007.

[7] W.-S. Hwang, S.-M. Chae and S.-W. Kim, "Yet Another Paper Ranking Algorithm Advocating Recent Publications," In *WWW*, 2010. (accepted)

[8] H. Tong, C. Faloutsos and J.-Y. Pan, "Fast Random Walk with Restart and Its Applications," In *ICDM*, pp. 613-622, 2006.

[9] J. Han and M. Kambe,. *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2nd Edition, 2006.