제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

핵심 논문 추출 방안

황세미, 배덕호, 황원석, 채수민, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: semiya@hanyang.ac.kr

A Method for Extracting Seminal Papers

Se-Mi Hwang, Duck-Ho Bae, Won-Seok Hwang,
Soo-Min Chae, Sang-Wook Kim
Dept. of Electronics Computer Engineering, Hanyang University

요 약

새로운 분야의 연구를 시작할 때, 해당 분야를 대표하는 핵심 논문을 파악할 수 있다면, 많은 도움이 된다. 본 논문에서는 특정 분야의 핵심 논문들을 추출하는 방안을 제안하고, 다양한 실험을 통해 제안하는 방안의 정확성과 효용성을 검증한다.

1. 서론

새로운 분야의 연구를 시작할 때, 해당 분야의 모든 논문을 조사하는 것은 매우 힘들다. 우수한 컨퍼런스나 저널에 실린 논문들만 조사하더라도 많은 시간과 노력이 필요하다. 따라서 해당 분야를 대표하는 몇 편의 핵심 논문을 알 수 있다면 해당 분야를 파악하는데 드는 시간과 노력을 크게 줄일 수 있다.

핵심 논문이란 해당 분야에서 영향력이 큰 논문으로,독창적인 아이디어를 제안하거나 해당 분야 연구의 기원이 되는 논문을 의미한다. 이러한 핵심 논문은 동일한 분야의 영향력이 큰 다른 논문들과 해당 논문과의 연관성이 높은 논문들로부터 참조를 많이 받는 경향이 있다.

논문의 영향력을 측정하는 기존 연구로는 [1,2,3] 등이 있어 왔다. 기존 연구들은 중요한 논문들에게서 참조를 많이 받는 논문일수록 큰 영향력 점수를 부여하였다. 그러나 참조 관계의 두 논문 사이의 연관성은 고려하지 않았다.

논문 p 이후에 발행된 유사한 논문들은 차이점을 설명하기 위해 논문 p를 참조할 의무를 지닌다. 만약, 논문 p를 참조하지 않았다면, 논문 p의 품질이 낮거나 해당 분야에서의 중요도가 높지 않다는 것을 간접적으로 의미한다. 따라서 연관성이 높은 논문들로부터 참조를 많이 받은 논문이 해당 분야에서 영향력이 큰 논문이라고 볼 수 있다.

또한, 오래전에 발행된 중요한 논문들뿐만 아니라 최근에 이슈가 되고 있는 논문들도 알 수 있다면, 해당 분야의 최신 연구 동향을 더 쉽게 파악할 수 있다. 따라서 본 논문에서는 최근 발행된 중요한 논문들도 추출할 수 있는 방안을 제안한다.

본 논문의 공헌은 다음과 같다. 첫째, 논문의 중요도와 연관성을 함께 고려한 논문의 영향력 측정 방안을 제안한다. 둘째, 오래된 논문들뿐 아니라 최신 논문까지 포함한 핵심 논문 추출 방안을 제안한다. 셋째, 다양한 실험을 통해 제안하는 방안의 정확성 및 효용성을 검증한다.

2. 관련 연구

논문의 영향력을 측정하는 기존 연구들은 RWR(random walk with restart) 개념을 기반으로 한다. 연구자는 다른 논문을 찾아볼 때에 해당 논문의 참고문헌을 보고 연관된 논문을 찾거나, 일정 확률로 참고문헌에 없는 다른 논문을 찾아보는 경향이 있다. RWR은 이러한 연구자의 논문 참조 패턴을 모델링한 것으로 널리 알려진 개념이다.

기존의 논문 영향력 측정 방안에는 참조, 저자, 논문의 관계에 RWR 개념[1]을 적용하여 각각의 영향력을 측정한 Browsing-based Model 방안[1], 참조와 출판 연도를 RWR에 적용하여 영향력을 측정함으로써 논문의 품질과 최신 정도를 반영하는 CiteRank 방안[2], 참조, 저자-논문, 학회-논문의 관계에 유사 RWR을 적용하여 영향력을 측정한 PopRank 방안[3] 등이 존재하였다.

기존 연구들과 본 논문의 제안하는 방안은 논문의 영향력을 측정한다는 목적은 동일하지만, 그 방안에는 차이가 존재한다. 첫째, 기존 연구들은 참조하는 논문의 중요도만을 고려하였을 뿐, 두 논문간의 연관성은 고려하지 않았다. 둘째, 기존 연구들은 RWR 개념에 기반을 두고 있으므로, 예전 논문들에게 상대적으로 높은 영향력을 부여하였다.

본 논문에서는 참조하는 논문의 중요도뿐만 아니라 연관성을 함께 고려하여 영향력을 부여하는 방안을 제안한다. 더 나아가, 예전 논문들의 점수 집중 현상을 완화할 수 있는 영향력 부여 방안을 제안한다.

3. 핵심 논문 추출 방안

본 논문에서는 동일 분야에 속하는 논문들의 집합[1]과 집합 내 논문들 간의 참조 정보가 주어졌을 때, 주어진 집합을 대표하는 핵심 논문 k개를 추출하고자 한다.

사용자는 분야의 파악 정도를 k를 통해 조절할 수 있다. 만약 k를 낮게 설정한다면, 해당 분야에서 영향력이 매우 큰 논문의 흐름만을 개괄적으로 볼 수 있는 반면, k를 높게 설정한다면, 이보다 더 세세한 흐름까지도 볼 수 있다.

제안하는 방안은 논문의 영향력을 측정하기 위해 RWR 개념을 이용한다. 이를 위해 논문을 노드로, 논문 사이의 참조 관계를 방향성 에지로 모델링한 후 RWR을 수행한

1) 동일한 분야에 속하는 논문들의 집합은 (1)클러스터링 기법을 통해 추출하거나, (2)사용자로부터 시드 논문을 입력 받아 해당 논문과 유사도가 높은 n편을 추출하는 방법을 사용할 수 있다.

다[2]. RWR을 통해 직접 연결된 노드의 영향력뿐만 아니라 그들의 이웃 노드들의 영향력까지 반영할 수 있다.

그러나 기존의 RWR은 다음과 같은 문제점이 존재한다. 첫째, RWR은 참조 관계에 있는 두 논문 사이의 연관성을 고려하지 않는다. 해당 논문과 유사한 논문들은 해당 논문을 참조할 의무가 있으므로 두 논문이 유사하지 않을 때보다 유사할 때의 참조가 더 중요하다. 따라서 자신과 유사도가 높은 논문들로부터 참조를 많이 받은 논문은 해당 분야에서 중요한 역할을 한 논문이라고 볼 수 있으며,이를 위해 논문들 간의 연관성을 고려하여야 한다.

둘째, RWR은 오래된 논문이 참조당한 수가 적음에도 불구하고 최신 논문에 비해 상대적으로 높은 영향력 점수를 갖게 되는 기득권 현상을 보인다. 이로 인해 최신 논문들은 핵심 논문으로 추출되지 못하는 결과를 가져온다. 한 분야의 연구의 기원부터 최신 동향에 이르기까지의 흐름을 파악하기 위해서는 오래된 논문의 기득권 현상을 완화시킬 필요가 있다.

본 논문에서는 RWR의 문제점을 해결하기 위해 WCP(weighted citation probability)를 제안한다. 제안하는 WCP는 이후에 발행된 논문들 중 유사한 논문이 해당 논문을 참조한 비율에 유사도를 반영한 것이다.

$$WCP(p) = \frac{\sum_{c \in C_p} sim(p,c)}{\sum_{t \in T_p} sim(p,t)} \quad (1)$$

WCP(p): 논문 p의 WCP 점수
sim(a,b): 두 논문간의 유사도[3]
C_p: 논문 p를 참조하는 논문들의 집합
T_p: 논문 p가 발행된 이후에 발행된 논문들의 집합

수식(1)에서 유사도는 논문들 간의 연관성을 반영하고, T_p는 오래된 논문들의 기득권 현상을 완화한다. WCP는 해당 논문 이후에 발행된 유사한 논문들이 많이 참조할수록 높은 점수를 가진다. 오래된 논문일수록 T_p는 증가하는 경향을 보이지만, 유사한 논문들로부터 많은 참조를 받을 경우, C_p도 커져, 전체적인 WCP(p)는 커진다.

그러나 WCP는 논문들 간의 참조 관계를 나타내지 못하므로 WCP와 RWR를 곱하여 최종 점수를 구한다. 이로 인해 논문의 연관성 및 중요도를 반영하면서 오래된 논문의 기득권 현상을 완화한 논문의 영향력을 구할 수 있다. 수식(2)은 최종 영향력 점수를 나타내며 계산된 점수가 높은 k편의 논문을 해당 분야의 핵심 논문으로 간주한다.

$$\text{영향력}(p) = RWR(p) \times WCP(p) \quad (2)$$

영향력(p): 논문 p의 최종 영향력 점수
RWR(p): 논문 p의 수렴 단계의 RWR 점수

4. 실험

실험에는 2009년 3월에 다운로드한 DBLP 데이터[4]를 사용한다. DBLP 데이터에는 참조 정보가 포함되어 있지 않으므로, Libra 데이터[4]를 통해 논문들 간의 참조 정보를 수집하였다. 수집된 데이터의 논문은 1,071,973개, 논문 당 평균 참조 수는 7.69개이다.

본 실험에서는 데이터 마이닝 기법인 클러스터링 분야, 순차 패턴 마이닝 분야, 공간 데이터베이스 분야, 총 3개 분야에 관해 핵심 논문 추출을 수행한다. 실험을 위해 각 분야에서 널리 알려진 논문들[5,6,7]을 시드 논문으로 해당 논문과 유사도가 높은 논문 300편을 추출하여 실험을 수행하였다.

실험 결과의 정량적인 분석을 위해 precision을 측정하였다. 이를 위해 [8]의 참고 문헌에 기술된 논문들을 정답으로 간주하였다. 표 1은 기존의 방안들과 제안한 방안으로 3개 분야에 대해 측정한 평균 precision을 나타낸다.

표 1. R-precision 평균 결과

	5	10	15	20
RWR	0.4	0.23	0.22	0.28
Browsing-based Model[1]	0.47	0.33	0.22	0.22
citeRank[2]	0.27	0.13	0.13	0.15
popRank[3]	0.27	0.33	0.31	0.23
제안 방안	0.47	0.43	0.38	0.33

실험 결과, 제안하는 방안의 precision이 기존 방안들보다 높은 것을 알 수 있다. 기존 방안들의 경우 RWR과 비슷하거나 오히려 낮은 precision을 보였지만, 제안하는 방안의 경우, WCP를 통해 RWR의 성능을 향상시킴으로써, 가장 높은 precision을 보였다.

5. 결론

본 논문의 공헌은 다음과 같다. 첫째, 논문의 중요도뿐만 아니라 연관성을 함께 고려한 영향력 측정 방안을 제안하였다. 둘째, 오래된 논문의 기득권 현상을 완화하는 영향력 측정 방안을 제안하였다. 셋째, 실험을 통해 제안하는 방안의 우수성을 검증하였다.

6. 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] S. Yan and D. Lee, "Toward Alternative Measures for Ranking Venues: a Case of Database Research Community," In *JCDL*, pp. 235-244, 2007.
[2] D. Walker et al., "Ranking Scientific Publications Using a Simple Model of Network Traffic," *Journal of Statistical Mechanics*, 2007.
[3] Z. Nie et al., "Object-Level Ranking: Bringing Order to Web Objects," In *WWW*, pp. 567-574, 2005.
[4] M. Ley, "DBLP: Some Lessons Learned," In *VLDB*, pp. 1493-1500, 2009.
[5] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," In *SIGMOD*, pp. 103-114, 1996.
[6] T. Raymond and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," In *VLDB*, pp. 144-155, 1994.
[7] R. Agrawal and R. Srikant, "Mining Sequential Patterns," In *ICDE*, pp. 3-14, 1995.
[8] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques,* Morgan Kaufmann, 2006.

2) RWR 수행 시, 에지 가중치는 두 논문 간의 내용 유사도로 부여한다
3) 본 논문에서는 내용 유사도를 사용한다
4) http://libra.msra.cn