제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

논문 계보 탐색 방안

배덕호*, 황세미*, 김상욱*
*한양대학교 전자컴퓨터통신공학과
e-mail: smith@agape.hanyang.ac.kr

A Method for Constructing Paper Genealogy

Duck-Ho Bae*, Se-Mi Hwang*, Sang-Wook Kim*
*Dept. of Electronics Computer Engineering, Hanyang University

요 약

새로운 논문은 대부분 기존 논문들의 영향을 받아 발행된다.따라서 논문들 간의 발행 계보를 파악할 수 있다면, 해당 분야의 연구 발전 과정을 파악하는데 큰 도움이 될 수 있다.본 논문에서는 논문들 간의 계보를 탐색하기 위한 방안을 제안하고,실험을 통해 제안하는 방안의 우수성을 검증한다.

1. 서론

매년 다양한 컨퍼런스 및 저널에서 방대한 양의 새로운 논문들이 발행되고 있다. 이러한 논문들은 기존 논문들의 아이디어를 발전시킨 새로운 아이디어를 제안하거나, 기존 논문의 성능을 향상시키는 방안을 제안하는 등 이미 발행되었던 논문들의 영향을 받아 발행된다. 따라서 논문들 간의 영향력 관계를 알 수 있다면, 한 분야의 연구 발전 과정을 파악하는데 큰 도움이 될 수 있다.

기존의 연구들은 주로 논문들과 논문들 간의 참조 관계로 이루어진 연결망에서의 개별 논문의 영향력을 측정하는 방안에 집중되었다[1][2][3]. 그러나 논문들 간의 영향력 관계를 탐색하는 연구는 아직까지 이루어지지 않았다.

본 논문에서는 한 논문이 다른 논문의 영향을 받아 발행되었을 때, 영향을 준 논문을 *부모 논문*, 영향을 받은 논문을 *자식 논문*이라 정의한다. 또한, 논문들 간의 부모-자식 관계를 *계보*라 정의한다.

본 논문에서는 주어진 논문들의 집합에서 논문들 간의 계보를 탐색하는 방안을 제안한다. 이를 위해 부모 논문 찾기 문제를 정의하고, 이를 해결하는 방안을 제안한다. 또한, 실험을 통해 제안하는 방법의 우수성을 검증한다.

2. 계보 탐색 방안

본 장에서는 한 분야의 연구의 발전 과정을 파악하기 위해 주어진 논문들의 집합에서 계보를 탐색하는 방안을 제안한다.

계보를 탐색하기 위해서는 먼저 해당 논문의 부모 논문을 찾아야 한다. 본 논문에서는 다음의 조건을 만족할 때, 자식 논문 c는 부모 논문 p의 영향을 발행되었다고 정의한다. 첫째, 논문 c는 논문 p를 반드시 참조하여야 한다. 둘째, 논문 p의 논문 c에 대한 '*부모 논문일*' 점수가 임계값 이상이어야 한다.

'부모 논문일' 점수는 한 논문이 다른 논문의 발행의 영향을 준 정도를 계량화한 점수로서, 참조 관계에 두 논문 사이에 부여되는 점수이다[1].

논문 p의 논문 c에 대한 '부모 논문일' 점수는 (1) 두 논문 간의 유사도와 (2) 논문 c와 유사한 논문들의 집합 S_c와 논문 p와의 유사도의 합에 의해 결정된다.

논문 p의 '부모 논문일' 점수를 계산하는 데 논문 c와 유사한 논문들의 집합 S_c를 이용하는 이유는 다음과 같다. 첫째, 논문 간의 정확한 유사도를 계산하기 위해서이다. 웹에서 수집 가능한 논문 데이터의 대부분은 본문이 없이 제목과 요약만 존재하며, 이마저도 제목만 있는 경우가 대부분이다. 이로 인해, 두 논문 간의 정확한 유사도를 계산하기 힘들며, 따라서 더 많은 데이터를 확보하기 위해,논문 c와 유사한 논문들의 집합 S_c를 이용한다.

둘째, 집합 S_c에 속하는 논문들의 대부분이 논문 p를 참조하고, 논문 p와의 유사도가 높을 경우, 집합 S_c를 논문 p로부터 발전된 하나의 분야라고 가정할 수 있다. 따라서 본 논문에서는 집합 S_c를 이용하여 두 논문 간의 유사도의 정확도를 향상시킨다[2].

그러나 오래된 논문들은 최신 논문에 비해 참조를 받을 기회가 많으므로, 부모 논문이 될 가능성도 커지게 된다. 본 논문에서는 이러한 문제를 해결하기 위해 참고 문헌 [2]와 동일하게 논문 c와 논문 p의 발행 연도의 차이에 따라 '부모 논문일' 점수를 정규화 하는 방안을 사용한다.

논문 p의 논문 c에 대한 '부모 논문일' 점수를 계산하는 방안은 수식은 다음과 같다. '부모 논문일' 점수는 Sc와 논문 p와의 유사도가 크거나, 두 논문간의 발행 연도의 차

1) 특정 논문의 경우, 더 이상 연구가 발전되지 않아 자식논문이 존재하지 않을 수 있다. 따라서 '자식 논문일' 점수 대신 '부모 논문일' 점수를 계산한다.

2) 논문 p 이후에 발행된 논문들만을 이용하여 집합 S_c를 구성한다.

표 1. 클러스터링 분야 논문 12편

No.	Title	Conf/Journal	Year
1	Efficient and Effective Clustering Methods for Spatial Data Mining	VLDB	1994
2	A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise	KDD	1996
3	BIRCH: An Efficient Data Clustering Method for Very Large Databases	SIGMOD	1996
4	STING: A Statistical Information Grid Approach to Spatial Data Mining	VLDB	1997
5	Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications	SIGMOD	1998
6	CURE: An Efficient Clustering Algorithm for Large Databases	SIGMOD	1998
7	An Efficient Approach to Clustering in Large Multimedia Databases with Noise	KDD	1998
8	Scaling Clustering Algorithms to Large Databases	KDD	1998
9	Entropy based Subspace Clustering Algorithm for Categorical Attributes	KDD	1999
10	ROCK: A Robust Clustering Algorithm for Categorical Attributes	ICDE	1999
11	Fast Algorithms for Projected Clustering	SIGMOD	1999
12	OPTICS: Ordering Points to Identify the Clustering Structure	SIGMOD	1999

이가 적을 때 커진다.

$$PS(c,p) = cite(c,p)\cdot(e^{-age}) \sum_{s \in S_c} \{cite(s,p)\cdot sim(s,p)\}$$

where cite(c,p)=1, if c cites p, and cite(c,p) = 0 otherwise.

age = 논문 p와 논문 c의 발행 연도의 차

3. 성능 평가

실험에는 2009년 3월에 다운로드한 DBLP 데이터[3]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 논문에서는 데이터 마이닝 기법인 클러스터링 분야에 대한 계보 탐색을 수행하였다. 실험을 위해 [4]의 클러스터링 챕터의 참고 문헌에 기술된 논문들을 대상으로 실험을 수행하였다. 표 1은 대상 논문 12편을 발행 연도를 기준으로 정렬한 결과이다.

본 실험에서는 내용 유사도가 높은 5편의 논문을 포함시켰으며, '부모 논문일' 점수가 가장 큰 한편만을 계보로 연결하였다. 이 외에도 순차 패턴 마이닝 분야, 공간 데이터베이스 분야 등 다양한 분야 관해 계보 탐색을 수행하였으나, 본 논문에서는 지면 관계상 생략한다.

그림 1은 생성된 클러스터링 분야의 계보를 나타낸다. 논문 8번은 논문 1번과 동일하게 k-means 기반 클러스터링 방안을 제안한 논문이며, 논문 2, 5, 7, 12는 모두 다밀도 기반 클러스터링 방안을 제안한 논문들이다. 또한, 논문 10은 논문 6과 동일한 저자들이 논문 6의 개념을 발전시킨 논문이다. 이렇듯, 제안하는 계보 탐색 방안은 우수한 결과를 보임을 알 수 있다.

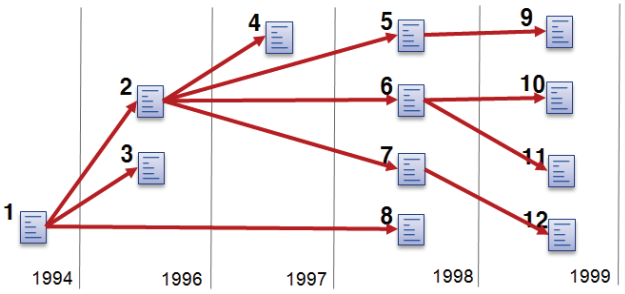

그림 1. 생성된 클러스터링 분야 계보.

4. 결론

본 논문의 공헌은 다음과 같다. 첫째, 논문들 간의 영향력 관계를 탐색하기 위해 부모 논문 찾기 문제를 정의하고, 이를 해결하기 위한 방안을 제안하였다. 둘째, 실험을 통해 제안하는 방안의 우수성을 검증하였다.

5. 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*, 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*, 2005.

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

3) http://www.informatik.uni-trier.de/~ley/db/

4) http://libra.msra.cn