링크 분석 및 학습을 통한 공동연구성과 기반 공저자 관계 예측

전현주[1] · 김윤후[1] · 정재은[1*] · 김건오[2]

[1]중앙대학교 · [2]트윈워드

Predicting Co-Authorship based on Link analytics and learning

HyeonJu Jeon[1] · YunHu Kim[1] · Jason J. Jung[1*] · Kono Kim[2]

[1]Chung Ang University · [2]Twinword

E-mail : {hyeonju, yunhu0110, j3jung}@cau.ac.kr / Kono@twinword.com

요 약

본 연구는 공동연구성과를 고려하여 링크 분석 및 학습을 통해 기대효과가 높은 논문의 공저자 협업 관계를 예측하는 방법론을 제시한다. 기존의 공저자 관계는 높은 정확도로 예측됨에도 불구하고 예측된 관계가 얼마나 좋은 관계인지 고려하지 않는 한계점을 보이고 있다. 따라서 본 연구에서는 위의 문제를 해결하기 위해 기대성과에 도움이 되는 공저자 관계 예측 방법을 다음과 같이 3가지 단계로 제안한다. (1) 서지정보 이종 그래프(Heterogeneous graph)를 구축하여 공동연구성과를 측정한다. (2) 공동연구성과를 기반으로 링크를 분석 및 학습한다. (3) 기대성과가 높을 것으로 전망되는 링크를 예측한다. 공동연구성과를 고려한 본 연구는 예측된 공저자 관계에 신뢰도를 높일 수 있을 것으로 기대한다.

ABSTRACT

This study proposes a methodology for predicting co-authorship of contributors to a highly anticipated paper through link analysis and learning, taking into account the result of collaborative research. Previous studies predict the co-authorship with high accuracy, but this shows limitations in that the quality of the predicted relationship is not considered. Therefore, to solve the above problem, we propose three steps to predict the co-authorship that will help with the expected performance: (1) Construct a heterogeneous graph to measure results of collaborative research. (2) Analyze and learn links based on results of collaborative research. (3) Predict links that are anticipated to have high expectation. It is expected to be useful for increasing confidence in the predicted co-authorship.

키워드

co-authorship, collaborative research, graph analytics, link prediction

Ⅰ. 서 론

각 연구 분야에는 많은 논문의 저자들이 있고, 그들은 학회, 논문 등을 통해 다른 저자들과 연결된다. 이에 따라 저자 간 커뮤니티는 매우 크고, 매년 급속히 성장하고 있다. 선행 연구에 의하면 잘 연결된 협력 네트워크를 가진 연구자 또는 연구단체가 생산성이 높다[1]. 따라서 연구자들이 공저자 협업 네트워크에서 가치 있는 새로운 저자들과 협력하는 것이 필수적이고 중요하다[2]. 그러나 실제로 비슷한 주제의 연구자를 찾아 함께 협업하는 것은 어렵다. 따라서 서지 정보를 분석하여 협업 가능성이 있는 공저자 관계를 예측 및 추천하는 연구가 제안되고 있다[3,4,5,6]. 공저자의 관계에 대한 예측이 상당히 높은 정확도를 보이고 있음에도, 예측된 공저자 관계가 얼마나 좋은 관계인지 파악하는 것은 어렵다.

저자 a와 함께 협업할 가능성이 있는 저자로 저자 b와 저자 c가 예측되었다고 가정해 보자. 저자 a는 어떤 저자와 협업했을 때 높은 성과를 낼 수 있을까? 기존의 연구로는 이를 합리적으로 판단하기에 어려움이 있다. 따라서 잠재된 링크를 저

* corresponding author

자 a와 저자 b,c의 기대되는 공동연구성과 기반으로 예측할 수 있다면 저자 a는 보다 합리적인 의사결정을 할 가능성이 높을 것이다.

이 문제를 해결하고자 본 논문에서는 공동연구성과를 학습하여 기대되는 성과가 높은 공저자 협업관계를 예측하는 방법론을 제안한다.

II. 관련 연구

공저자 협업관계를 예측을 위해 링크 예측 기법을 사용하는 연구가 진행 중에 있다. 링크예측(link prediction)은 링크 형성 확률을 평가하여 잠재된 링크를 예측하는 기법[3]인데, [4]은 링크 형성 확률을 평가하기 위해 두 노드 사이의 의미적 유사성을 고려하였고, [5]는 위상(topology)적 유사성을 고려하였다. 실제 서지 네트워크에 있는 여러 유형의 객체와 링크를 이종그래프에 표현하였다. 그 후 새로운 방법론인 path predict라는 메타경로(meta path)기반 관계예측 모델을 제안하였다. [6]는 공저자 협업관계 예측의 정확도를 높이기 위해 빈도통계에서 파생된 확률 모델의 영역을 적용하였다. 또 확률 모델의 결과예측을 추가 기능으로 사용하여 예측의 정확도를 높였다.

하지만 일반적인 링크 예측 기법에서는 링크가 형성될 확률만 고려한다는 한계점이 있다. 따라서 예측된 링크 간 서로 다른 중요도에 대해서는 평가할 수 없다.

본 논문에서는 이 문제를 해결하기 위해 공동연구성과를 측정하여, 이를 기반으로 링크 예측 결과를 학습해 링크 간 중요도를 표현하는 방법을 제안하고자 한다.

III. 공동연구성과 학습기반 링크 예측

이 장에서는 공저자 협업관계를 예측하기 위해 본 논문에서 사용하는 데이터 표현을 정의하고, 기대되는 성과를 고려한 협업관계 예측을 위해 두 저자 간 공동연구성과의 측정 방법이 제시된다.

3.1 공저자 협업관계 네트워크

네트워크 구조와 노드의 특징을 결합하여 고려할 때[3], 서지 정보에 관한 원본 데이터와 가장 유사하게 표현되도록 노드와 링크가 한 가지 이상 있는 이종 그래프의 형태로 데이터를 표현한다[5]. 이종 그래프는 서로 다른 노드들 사이의 관계에 대한 정보를 각각 가지고 있어 링크 이동의 의미를 파악하는데 유용하다. 전체 그래프를 G라고 했을 때,

$$G^{hetero} = < A, P, V > \quad (1)$$

로 표현할 수 있다. A와 P는 각각 저자와 논문의 노드를 나타내고, V는 아래 [식 2]와 같이 그래프의 두 종류 링크를 나타낸다.

$$V = \{ W, C \} \quad (2)$$

링크 W는 저자와 논문 간의 저술관계를 나타내고, 링크 C는 논문과 논문 간의 인용관계를 나타낸다. [그림 1]에서 표현된 a, p는 각각 저자와 논문의 노드들을 나타낸다. a_i는 i번째 저자노드를 의미하고, i는 $1 \le i \le I$ 의 범위를 갖는다. (I는 전체 저자노드의 수이다.) p_n는 n번째 논문노드를 의미하고, n는 $1 \le n \le N$ 의 범위를 갖는다. (N는 전체 논문노드의 수이다.)

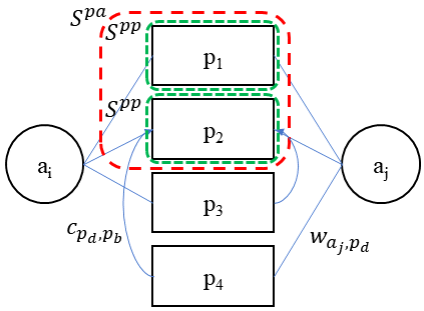

그림 1. 공저자 협업관계 네트워크

3.2 공동연구성과 측정

두 저자 간 공동연구성과는 공저논문수와 각 공저 논문들의 인용수를 구함으로써 측정할 수 있다. 우선 네트워크 내의 저자와 논문, 논문과 논문 간의 관계를 분석하여 전체 논문집합P의 부분집합 P_{ai}와 P_{p_i}를 정의한다.

$$P_{a_i} = \{ p_n \mid \exists\, w_{a_i p_j} \in W \} \quad (3)$$

P_{ai}는 전체 논문집합P에서 저자a_i가 저술한 논문들의 집합을 의미한다. 따라서 저자와 논문간의 저술관계를 나타내는 집합이므로 링크 w를 기반으로 정의한다.

$$P_{p_i} = \{ p_n \mid \exists\, c_{p_i p_j} \in C \} \quad (4)$$

P_{p_i}는 전체 논문집합P에서 논문p_i가 인용한 논문들의 집합을 의미한다. 따라서 논문과 논문간의 인용관계에 기반을 둔 집합이므로 링크 c를 기반으로 정의한다.

위에서 정의한 집합들의 관계를 바탕으로 공저

자 협업관계를 도출하기 위한 공저논문수와 논문 간 인용수를 측정한다.

$$S^{pa}(a_i, a_j) = \left| P_{a_i} \cap P_{a_j} \right| \quad (5)$$

$$S^{pp}(p_n) = \left| P_{p_i} \cap P_{p_j} \right| \quad (6)$$

S^{pa}는 두 저자 간의 공저논문수를 나타내고, S^{pp}는 두 논문 간의 인용수를 나타낸다.

공동연구성과는 [그림 1]과 같이 두 저자(a_i, a_j) 간 공저논문수와 각 공저논문(p_a, p_b) 인용수의 총합을 곱하여 다음 [식 7]과 같이 측정한다.

$$r(a_i, a_j) = \begin{cases} S^{pa} \times \sum_{p_n \in P_{a_i} \cap P_{a_j}} S^{pp}(p_n) & \text{if } S^{pa} \neq 0 \\ 0 & \text{if } S^{pa} = 0 \end{cases} \quad (7)$$

두 저자 간 공동연구성과를 반영한 관계는 $r(a_i, a_j)$로 나타낸다. 측정된 공동연구성과를 저자들의 행렬로 표현하면 다음[표 1]과 같다.

표 1. 두 저자 간 공동연구성과 행렬

	a1	a2	a3	…	aj
a1	r(a1, a1)	r(a1, a2)	r(a1, a3)		r(a1, aj)
a2	r(a2, a1)	r(a2, a2)	r(a2, a3)		r(a2, aj)
a3	r(a3, a1)	r(a3, a2)	r(a3, a3)		r(a3, aj)
…				…	
ai	r(ai, a1)	r(ai, a2)	r(ai, a3)		r(ai, aj)

3.3 공저자 협업관계 예측 모델

이번 장 에서는 공저자 협업관계 그래프에서 공동연구성과의 패턴을 학습하여 새로운 공저자 협업링크를 예측하는 방법을 제안한다. 예측 모델이 공동연구성과의 학습을 이용함으로써 예측된 저자 간의 기대성과를 함께 예측할 수 있다. 따라서 시스템은 기존보다 더 구체적인 예측을 보여줄 것이다.

공저자 협업 네트워크에서 잠재된 링크를 찾아내기 위해 기존의 공동연구성과의 패턴을 학습하는 합성곱 신경망(CNN)을 활용한다.

$$M_{(t)} \rightarrow \widetilde{M}_{(t+n)} \rightarrow \widehat{M}_{(t+n)} \quad (8)$$

t년도에서 n년 후인 $t+n$년도의 링크 예측 결과를 학습한다. 여기서 $\widehat{M}_{(t+n)}$는 최종 예측 값이다. 그러나 그래프 학습을 시작하기 전, 성능을 높이기 위해 추정된 값을 초기 값으로 공동연구성과를 학습한다. $\widetilde{M}_{(t+n)}$는 무작위 초기화를 방지하여 빨리 수렴할 수 있도록 만드는 추정된 값이다. 이를 구하기 위한 방법으로는 전통적인 링크 예측(link prediction)기법을 활용한다.

공동연구성과를 반영한 가중치를 수정(update)하여 실제 값과 예측 값의 차이가 가장 작은 신경망 모델로 학습시킨다.

$$L = r(a_i, a_j) \times \sum_{\substack{\forall a_i, a_j \in A, \\ a_i \neq a_j}} \| M_{(t+n)i,j} - \widehat{M}_{(t+n)i,j} \| \quad (9)$$

Ⅳ. 실험 계획 및 결론

서지 정보의 네트워크를 구성하기 위해 DBLP데이터를 사용한다. 컴퓨터 분야에서 발표된 연구논문으로 논문과 저자의 관계를 표현한다.

2001년부터 2015년까지의 공저자 네트워크를 3년 단위로 나누어 training data로 활용한다. 공저자 간의 공동연구성과를 바탕으로 공저자 관계예측 모델을 학습한다.

2016년부터 2018년까지의 데이터를 test data로 활용한다. 학습된 모델의 정확도를 측정한다.

예측한 데이터와 실제 데이터의 결과를 비교한다. 학습을 반복했을 때, 정확도의 변화를 관찰한다.

표 2. 훈련데이터와 시험데이터 구성의 세부사항

	Time peroid
Training Peroid	2001.01.01. to 2003.12.31.
	2002.01.01. to 2004.12.31.
	2003.01.01. to 2005.12.31.
	...
	2015.01.01. to 2017.12.31.
Test Period	2016.01.01. to 2018.12.31

각 연구분야에는 논문을 저술한 수많은 저자들이 있다. 이들이 속해있는 커다란 네트워크에는 비슷한 연구주제를 가진 저자들이 많이 있고, 이들이 함께 연구한다면 높은 성과를 낼 가능성이 크다는 것이 이전의 연구에서 밝혀진 바 있다. 따라서 기대성과를 고려한 공저자 협업관계를 예측하는 것은 많은 저자들에게 성과 높은 연구를 할 기회를 가져다 줄 수 있는 새로운 접근법이다. 이 논문에서는 그러한 공저자 협업관계 예측을 위해 공동연구성과를 측정하고 학습

하는 방법론을 제안하였다.

공동연구성과를 고려해 예측한 공저자 협업관계는 기대성과를 예측하는 것이 가능하다. 이를 통해 저자들은 수많은 예측된 관계 중 더 높은 성과를 가져올 수 있는 저자와 실제로 관계를 만들어 나갈 수 있고, 해당 분야의 연구성과가 더 깊어지는 것을 가능케 한다.

본 논문에서 제안한 방법론으로 앞으로의 작업에서는 실험결과를 제안할 것이다. 또한 저널과 학회, 논문의 주제 등 다양한 정보들을 고려하여, 예측결과의 정확도를 높이기 위한 방법에 대해서 제안하고자 한다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음 (20170001000031001).

References

[1] S. Lee and B. Bozeman. “The impact of research collaboration on scientific productivity.” Social Studies of Science, Vol. 35, No. 5, pp. 673－702, 2005.

[2] Hung-Hsuan Chen , Liang Gou, Xiaolong Zhang, Clyde Lee Giles, “CollabSeer: a search engine for collaboration discovery”, Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, Ottawa, Ontario, Canada, pp. 231-240, 2011.

[3] Linyuan Lü, Tao Zhou, “Link prediction in complex networks: A survey”, Physica A: Statistical Mechanics and its Applications, Volume 390, Issue 6, pp. 1150-1170, 2011.

[4] Lars Backstrom, Jure Leskovec, "Supervised random walks: predicting and recommending links in social networks", Proceedings of the fourth ACM international conference on Web search and data mining, Hong Kong, China, pp. 635-644, 2011.

[5] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal and J. Han, “Co-author Relationship Prediction in Heterogeneous Bibliographic Networks”, 2011 International Conference on Advances in Social Networks Analysis and Mining, Kaohsiung, pp. 121-128, 2011.

[6] C. Wang, V. Satuluri and S. Parthasarathy, “Local Probabilistic Models for Link Prediction”, Seventh IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, pp. 322-331. 2007.