Uncovering the Mechanisms of Co-authorship Network Evolution by Multirelations-based Link Prediction

Jinzhu Zhang, Chengzhi Zhang, Bikun Chen

{zhangjinzhu, zhangcz,chenbikun}@njust.edu.cn
Nanjing University of Science and Technology, Dept of Information Management, Xiaolinwei Str 200, Nanjing (China)

Introduction and literature review

Co-authorship network, a proxy of research collaboration, reveals the collaboration patterns and the determining factors through social network analysis perspective, with nodes representing authors and links representing co-authorships (Ortega, 2014; Yan & Ding, 2009). If we know what mechanisms push the evolution of co-authorship network, we could predict which authors may collaborate in future.

Most of the studies correlate co-authorship evolution mechanisms to similarity indicators which quantitatively compared by link prediction in homogeneous network (Lu & Zhou, 2010). In order to integrate multirelations between authors, path-based similarity indicators are proposed for co-authorship prediction in DBLP heterogeneous network (Sun et al., 2011; Sun & Han, 2013). However, what is the role of each mechanism plays and how to combine multiple mechanisms to suit the co-authorship network evolution need to be clarified, moreover, the method need to be verified in different domains.

Therefore, we integrate similarity indicators based on multirelations in heterogeneous network and quantitatively evaluate them by link prediction justly, to uncover and infer the mechanisms of co-authorship network evolution. Firstly, similarities between authors are represented by a matrix where the rows are multirelations and the columns are multirelations' measures. Secondly, the evaluation of similarities is processed based on link prediction, to reveal the importance of each mechanism which is the weight for combining multiple mechanisms. Finally, experiments are presented in the domain of Library and Information Science (LIS), which reveals the best appropriate mechanism, the significance of each mechanism and the combination strategy of different mechanisms.

Data and method

Data

We collect the data from the SCIE (Science Citation Index Expanded) databases in Thomson Reuters' Web of Science, using journal publications on subject category of LIS across 2000 to 2009.

We choose the authors that the frequency greater than or equal to five as the experiment data, which includes 669 authors, 3,948 articles, 6,476 keywords, 14 subject categories, 29 journals and 79,717 references.

We eliminate the subject categories because of too small numbers and references because of computing complexity. The co-author network has 1052 edges that indicate co-authorship, where we randomly choose 946 (90%) edges as training set and the remaining 106 edges as the testing set.

Multirelations-based link prediction

(1) Representation of co-authorships via multi-relations: Co-authorships via multirelations are systematically represented and extracted in a heterogeneous bibliographic network shown in Figure 1. Part of multirelations between authors could be represented in Table 1.

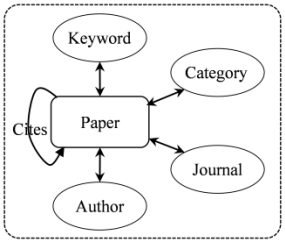

Figure 1. The nodes and relations in heterogeneous bibliographic network.

Table 1. Multirelations between authors.

Relations	*Description*
A-P-A-P-A	Common neighbours
A-P-A-P-A-P-A	Common neighbours' neighbours
A-P-J-P-A	Publish paper at the same journal
A-P-K-P-A	Authors have the same keyword
A-P-K-P-K-P-A	Authors' keywords co-word in same paper
A-P→P-A	Author x cite author y
A-P←P-A	Author x is cited by author y
A-P→P←P-A	Authors x and y cite the same paper
A-P←P→P-A	Authors x and y co-cited by same paper
A-P→P→P-A	Author x cite the paper that cite author y
A-P←P←P-A	The reverse relation of the above

(2) Measures of each relation: The four measures are the follows: path count (PC) is the number of

shortest path between two authors, normalized path count (NPC) is to discount PC by their overall connectivity, random walk (RW) and symmetric random walk (SRW) (Sun & Han, 2013).
(3) Evaluation of similarities based on link prediction: The relations and their measures combine the similarities, so there are 44 similarity indicators combined by 11 relations with four measures. We evaluate all the similarity indicators based on link prediction with precision and area under the curve (AUC).

Results

The three comparison perspectives are: (1) from the horizontal axis, compare which relation is best appropriate to the mechanism. (2) From the longitudinal axis, compare which measure is best to describe the mechanism. (3) Comparison between combined-relations-based and single-relation-based mechanisms.

The evolution mechanisms based on single-relation-based similarities

In Table 2 and Table 3, the entries emphasized in bold and italic corresponding to the highest accuracies from the horizontal axis.
In precision, the APAPA with NPC is the best appropriate and important mechanism in LIS where NPC plays the best in four measures, yet the APJPA with RW plays the worst. In AUC, the APAPA with SRW is the best mostly with little differences. There is lots of information loss in the projection from heterogeneous network to homogeneous network compared with CNs.

Table 2. The precision/AUC of single-relation-based similarities.

Relations	*PC(%)*	*NPC(%)*	*RW(%)*	*SRW(%)*
APAPA	38.4/87.5	***42.5***/87.5	31.7/87.7	41.4/***87.9***
APAPAPA	24.0/***86.2***	***32.9***/86	21.1/***86.2***	29.4/85.8
APJPA	3.2/76.8	***3.9***/77.2	0.9/76.7	2.6/***77.4***
APKPA	7.6/81.4	***20.4***/82.1	9.4/81.8	16.3/***82.3***
APKPKPA	2.2/70.8	***4.9/72.5***	2.5/70.9	4.3/72
CNs	23.4/84.1			

Comparison between combined-relations-based and single-relation-based mechanisms

The paper designs five combination strategies for comparison: (1) CR1: Combination of all relations without weights. (2) CR2: Combine all relations except APJPA. (3) CR3: Combination of all relations with weights denote by precision in Table 2. (4) CR4: the combination formed via just authors which is APAPA+APAPAPA. (5) CR5: the combination formed via just keywords, which is APKPA+APKPKPA. The precision and AUC are listed in Table 3.
In precision, the CR3 with NPC is the most appropriate and important mechanism in LIS where NPC plays the best in four measures, yet the CR5 with PC plays the worst. The AUC is consistent with the precision result mostly and others with little differences. The CR2 and CR3 with each measure are all outperformed the single-relation-based mechanisms. The CR4 performs much better than CR5 proves that in co-authorship formation the author is more important than research interest.

Table 3. The precision/AUC of different combinations of relations.

Relations	*PC(%)*	*NPC(%)*	*RW(%)*	*SRW(%)*
CR1	28.6/86.4	**40.8/88.6**	26.3/88.4	36/88.3
CR2	38.6/84.8	**43.7/87.4**	32.4/86.4	43.6/86.8
CR3	45.1/89.1	**49.2**/89.3	39.8/89.0	47.2/**89.5**
CR4	24.2/86	**38.6/86.4**	27.1/86.2	35.3/86.1
CR5	2.2/80.6	**16.7**/82.8	6.6/**83.1**	12/82.7

Conclusion and discussion

This paper uncovers the mechanisms of co-authorship network evolution by multirelations-based link prediction in LIS. In the next, we will consider other factors that influence research collaborations, all relations especially related to references to enhance the accuracy and validation in two or more different areas with different article types (e.g., journal and conference).

Acknowledgments

Our work is supported by the Ministry of Education of China Project of Humanities and Social Sciences (Grant No. 14YJC870025), the Fundamental Research Funds for the Central Universities (Grant No. 30915013101) and the National Natural Science Foundation of China (Grant No. 71173211).

References

Lu, L. & Zhou, T. (2010). Link prediction in complex networks: A survey. *Arxiv preprint arXiv:1010.0725*.

Ortega, J. L. (2014). Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search. *Journal of Informetrics, 8*(3), 728-737.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C. & Han, J. (2011). Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. *Proc. ASONAM*.

Sun, Y. & Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter, 14*(2), 20-28.

Yan, E. & Ding, Y. (2009). Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118.