International Journal of Pure and Applied Mathematics
Volume 120 No. 5 2018, 781-791
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

EXPLOITING ACADEMIC FACTORS FOR IMPROVING COLLABORATION RECOMMENDATION SYSTEM

Ghaidaa A. Al-Sultany
PhD at Department of Information Network
University of Babylon University, Babel, Iraq
ghaidaa.almulla@itnet.uobabylon.edu.iq

Abstract

For the sake of recommendation quality, in this research, we present a new method for recommending collaborators to scholars based on aggregating set of significant factors related with the authors' research history. Four academic factors have been considered to predict the significant links between the target author and other elective collaborators. The outputs of the factors are passed to the recommendation process to enhance the decision-making of the collaborators' recommender. PageRank algorithm ranks a number of k-author that have significant interaction with target author and then seek to nominate a new valuable partner to collaborate with target researcher. It has taken into account the aggregated collaborations' factors to discover the significance of links in academic network. the experiments analysis was conducted on *DBLP and Wikicfp* data set to evaluate the outcome of the recommendation system. The results have shown encouraging performance in comparison with the other research model *MVCWalker*. It has proved that aggregating collaborations' factors can increase the academic collaboration recommendations performance in terms of the precision, recall rate.

Keywords

Academic Collaboration, PageRank Algorithm, Collaborations' factors, link prediction

Introduction

The scientific collaboration becomes increasingly, complex, and goes to involve large-scale collaborations and multidisciplinary teams. Consequently, find the valuable collaborators it is often a complex task and time-consuming for researchers with large volume of big scholarly data (Smart and Bayer, 1986). Besides, academic collaboration begins supported by many countries to induced international collaboration (Gazni, et al.2012). In addition to staying in touch with close collaborate researcher prefer to work with valuable collaborators not yet known them (xia et al.2015). However, a collaboration between researchers provides better knowledge (Cheol Shin et al.2013), and has a positive impact on researcher productive (Bordons et al. 2015) particularly between different disciplines (Cheol Shin et al.2013). To be able to cope with the increasing scholarly materials and trying to moving away from the quantity towards the quality of collaborate. In this paper, proposed system to hybrid similarity measure and hits algorithm for collaborators suggestion. Using DBLP data set (build co-author network reflects the author – author, author-paper, and author-conference relationship. Follow the phrase "friend of my friend" to recommended new collaborators by exploiting coauthor network (Sielis, et al. 2015).

In this study, web mining based PageRank algorithm (Aggarwal, C. 2016) (Ricci, et al. 2015) has been applied to address the most ranked relevant k-collaborators authors to a given author with respect to number of collaborating factors (the order of co-authors, the time of collaboration, the number of collaborations and publishing similarities). PageRank algorithm is adapted to the network that supplies a chain for ranking using author's link Prediction process (Carullo, et al. 2014) (Batra, et al. 2013). The relationships between the target author and the other authors' collaborators are represented as a graph with nodes refer to the publishing authors and the edges reflect the importance between the nodes. Hence, computing the relation between the target author and the other authors has depended on their importance to each other from which each recommended node has a rank score.

Related work

Due to the enormous volume of scholarly data, the task of finding collaboration relationships among different authors is a challenging burden (Achakulvisut et al. 2016). Several recommender systems have been proposed to help the researcher to interact more easily and share information in the academic community. The research in (Hoang, et al. 2017) addressed research similarity and interaction strength mong authors via developing academic event recommendation method. The experiments of the work were tested on the the DBLP Computer Science Bibliography and Wiki Calls for Papers (WikiCFP). The researchers in (Xia et al. 2014) proposed a model named MVCWalker, in which most valuable collaborators are recommended to scholars. They explored some of the academic factors to discover the relationships between

authors. In (Luong et.al.2015), the research paper focused on finding new co-authors from existing relationship extracted from bibliographic DBLP data set. the research discussed a development of random walk model to find potential co-authors for a target author. Exploring the process of link production is quite significant task in network analysis. Consequently, the authors in Khrouf and Troncy (2013) investigated the researchers' behaviours and attendance records to recommend new collaboration between them. They utilised statistical analysis of scientific collaboration to discover the relationship among the researcher, in particular those whom have more interactions on the social networks. In 2011, (Sun et al.2011) proposed four measures along the lines of topological features in homogeneous networks. In addition, the authors used the logistic regression model as the co-authorship probability prediction model between two authors and studied the relationship prediction in the heterogeneous bibliographic network.

In our work, expanding to the research in (Xia et al. 2014), recommending k- authors collaborators to target author based on the similarity of authors' publications using the PageRank Algorithm.

PageRank based Collaborators Ranking

Relatedness and importance are two measure of link analysis. Relatedness measures the connection between two nodes in the same graph. PageRank Algorithm is one of the sophisticated web mining algorithms for evaluating the importance of documents. Importance is a measure for ordering nodes based on their effect, important or popularity in given graph. As Page Rank considers one of the probability distribution, it can measure likelihood that a person will arrive at any particular page by random click on links (Aggarwal, C. 2016) (Ricci, et al. 2015). it's used by Google Search to rank websites in their search engine results and can be calculated for any-size collection of documents. It's used for ranking webpages that indexed by a search engine (Grover, et al 2012). PageRank and its hyperlinks described by a directed graph $G = \langle V, E \rangle$ where the nodes V correspond to the pages and the edges E is the set of hyperlinks that link the pages to each other, and the directed edge $(p, q) \in E$ indicates the existence of a link from p to q. In PageRank algorithm, the initial value to each page will be depend on the number of page, it will be equal to 1/n where n refer to the total number of pages in web graph. Then, for each iteration the PageRank of all pages depend on the PageRank of the pages that refer to it proportional to the number of outgoing links for that pages (Batra, et al. 2013) (Ricci, et al. 2015) .

Methodology

Principally, Academic collaboration among academic researchers has been considered for most successful scientific research achievements as most the productive scholars prefer to collaborate with other researcher to produce valued work. However, the process of finding the valuable collaborators from a big scholarly data centre is often time consumable. In this work, we present a new method for recommending collaborators to scholars based on aggregating set of significant factors related with the authors' research history. Five academic factors have been considered to predict the significant links between the target author and other elective collaborators. The outputs of the factors are passed to the recommendation process to enhance the decision-making of the collaborators' recommender. PageRank Algorithm based recommendation system are implemented in this work in which academic collaboration factors are aggregated to produce set of valuable potential collaborator to the target author.

Data Prepressing and Representation

In this work, the data was provided by ***DBLP (database systems and logic programming)*** (Ley, M. 2009) on the webpage "https://dblp.uni-trier.de" in the form of XML files. It contains around 4,096,214 publications and more than 2 million authors with about 5.370 conferences and 1.568 journals (Ley.2009). In addition, the ***Wikicfp (Wiki call for paper)*** dataset available on the website "http://www.wikicfp.com/cfp/" has utilized to help the researchers to find organized information for call of papers. It is represented as XML files also. In this research, the dataset has been divided into two partitions (Training and Test Set) based on the year of the publications in which the published data before the year 2010 has constituted the training set while the data after 2010 was saved for testing set.

Collaboration Factors based Link Importance

Basically, the network hyperlink between two nodes indicates the cooperation relationship between them. As it is known that choosing certain nodes with high values is preferred for measuring the relationships strengths, in our work, the PageRank recommendation algorithm has assigned the edges among the authors nodes with factors based weight to quantify the cooperation between one user and potential collaborators.

In (Xia et al. 2014) three factors (as will be detailed below) have been taken into account to calculate the relationship between given author and other past collaborators, however the research has not taken into account the research relatedness for a given author with other collaborators, which might represent one of the most significant factor to predict new collaborators to the target author. Inspired from the work of the authors Xia et al, the similarity among a given author and other new collaborators has been aggregated with the factors (*Co-*

author order, Time of collaboration, Time of collaboration) to conclude closer relationships between authors as detailed below.

- ***Co-author order***

The author order can be measured with respect to his contribution to the work of a given paper. Since the first and second researchers' names in one paper usually reflect the highest contribution to the that paper, the co-author order may reveal the cooperation strength among the researchers as stated in equation (1) that measures the link importance using the co-author order.

$$DCL(p_i; p_j) = \begin{cases} \frac{1}{i} + \frac{1}{j} & j < 3 \\ \frac{1}{i} + \frac{2}{j} & j > 3, i \leq 3 \\ \frac{2}{i} + \frac{2}{j} & i > 3 \end{cases} \quad \ldots (1)$$

Where **DCL** reflect the significance of link between two nodes (a_m, a_n) in co-author network.

- ***Time of collaboration***

It's obvious that the links among academics change over time Where the academic researcher tends to share with the researchers whom co-authored a paper in recent time rather than those whom co-authored of decade ago, as specified in equation (2).

$$Time\ (a_m, a_n) = \frac{t_i - t_0 - 1}{t_c - t_0 - 1} \quad \ldots (2)$$

where t_0 is the time of first published co-authored research, t_i is the time of publishing research paper *i*, in which two researchers co-authored and t_C is the current time.

- ***Times of collaboration***

This factor reflects the times of collaboration in academic social networks as illustrated in equation (3).

$$LIM(a_m, a_n) = \sum_{t=t_1}^{t_2} DCL * Time(a_m, a_n) \quad \ldots (3)$$

Research Relatedness based Link Importance

In order to calculate the similarities between researchers, we must know the similarities between their publications. Here we must consider the content of their research and the summary

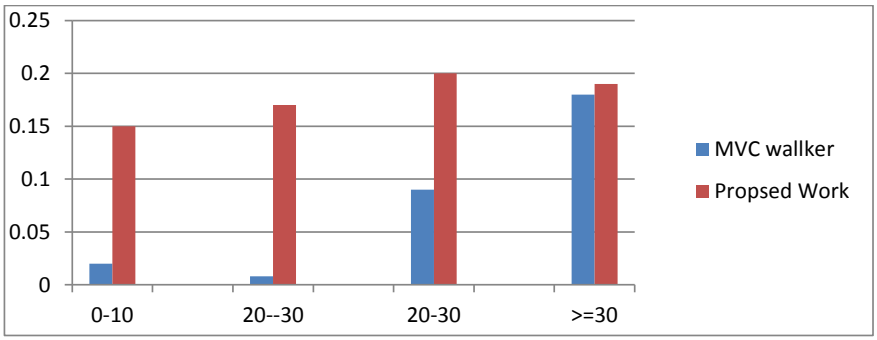
of the research found in the abstract and the title of the research. Here we will use the title in addition to using the DOI to derive the research abstract to calculate the similarity. Today most of publishers use DOIs to address publications. (DOI) link "digital object identifier" is record available in DBLP data set in <ee> tag, in which <ee> represent the global URL for electronic version of authors paper using Beautiful soup tool to fetch abstract from web.

$$R(a_m,a_n) = \sum_{p_i \in p_{a_m}} \sum_{p_j \in p_{a_n}} \frac{sim(p_i,p_j)}{|p_{a_m}| \times |p_{a_n}|} \quad \ldots (5)$$

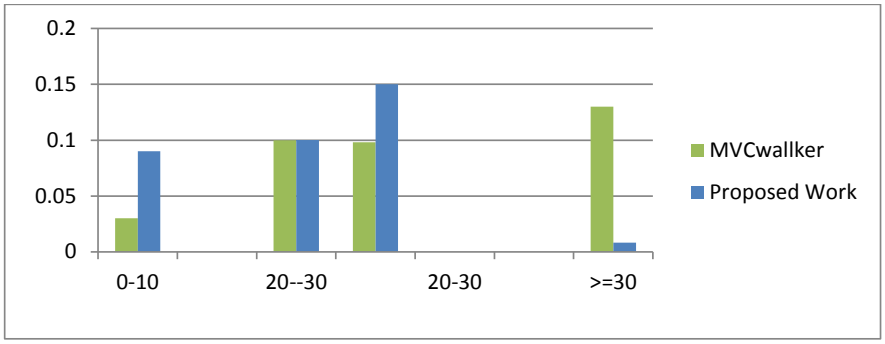
Where $R(a_m; a_n)$ is the research relation between authors a_m and a_n, and $|p_{a_m}|, |p_{a_n}|$ Indicate the number of all papers they wrote by authors a_m and a_n. The function $sim(p_i, p_j)$ returns the content similarity between the abstracts of publications ***pi*** and ***pj***.

Result and Evaluation

the effectiveness recommended collaborators to given author in our work has been examined in terms of the precision and recall measures different parameters. The experiment has been done on 100 target authors whom have a variety number of collaborators between (0 - 30). The number of collaborators was divided into four parts ((0 -10), (10 – 20), (20 - 30) and more than 30). The proposed research was evaluated and compared with MVC Walker algorithm developed in [] and the outcome has shown outperformance with respect to the precision and recall. The highest precision was 0.2 when the target author has (20 -30) collaborator as shown in Figure (1-a). while recall equal to 0.15 when the number collaborators between (0 – 10) as shown in Figure (1-b).



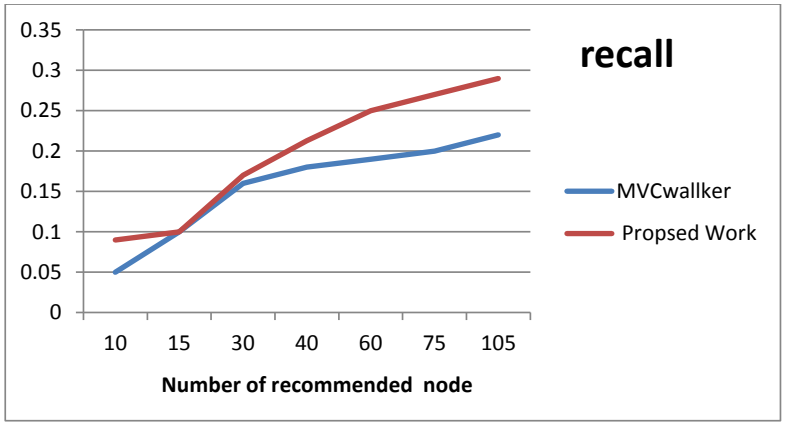
Figure(1-a): ***The Precision according to the number of collaborators partitions***



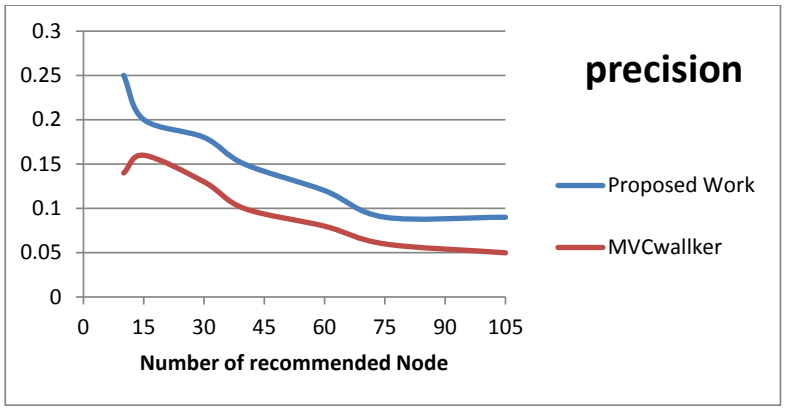
Figure(1-b): ***The Recall according to the number of collaborators partitions***

Number of recommended Nodes

A number of recommended Nodes have great influence on Precision and recall. Basically, recommending more users raising the recall measure up and decreasing the precision measure, whereas, few recommended users have the opposite effect. It has shown that the higher precision and recall was obtained when the number of recommended nodes equaled to 10 as illustrated in Figure (2-a). the value of precision decreases dramatically when the number of recommended nodes get increased. On the other hand, recall has increased while the number of recommended nodes is increased also, as shown in Figure (2-b).



Figure(2-a): ***The Recall with respect to number of recommended nodes***



Figure(2-b): ***The Precision with respect to number of recommended nodes***

Conclusion

In this paper, using collaborative filtering for analyzing the author behavior in social academic networks has been proposed. It explores set of academics' collaboration factors to compute similarities between the target author and a set of collaborators. PageRank algorithm ranks a number of k-author that have significant interaction with target author and then seek to nominate a new valuable partner to collaborate with target researcher. It depended on four academic factors that have been taken into account to produce set of valuable potential collaborator to the target author.

The research evaluation has shown that increasing the number of factors effect positively on the recommendation system as the results outperform the performance of the past model *MVCWalker*. At last but not least, Other features can be considered in this direction as well as, more experiments tests can be conducted on the data to improve the work.

Reference

- Achakulvisut, T., D. E. Acuna, T. Ruangrong, and K. Kording. 2016. Science concierge: A fast content-based recommendation system for scientific publications. *PLoS ONE* 11:e0158423. doi:10.1371/journal.pone.0158423
- Aggarwal, C. C. (2016). Recommender systems (pp. 1-28). Springer International Publishing. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), 604-632.

- Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. Journal of Informetrics, 9(1), 135-144.
- Carullo, G., Castiglione, A., & De Santis, A. (2014, September). Friendship recommendations in online social networks. In Intelligent Networking and Collaborative Systems (INCoS), 2014 International Conference on (pp. 42-48). IEEE.
- Cheol Shin, J., Jeung Lee, S., & Kim, Y. (2013). Research collaboration across higher education systems: maturity, language use, and regional differences. Studies in Higher Education, 38(3), 425-440.
- Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. Journal of the Association for Information Science and Technology, 63(2), 323-335.
- Grover, N., & Wason, R. (2012). Comparative analysis of PageRank and hits algorithms. International Journal of Engineering Research & Technology (IJERT), 1(8), 1-15.
- Hoang, D. T., Tran, V. C., Nguyen, V. D., Nguyen, N. T., & Hwang, D. (2017). Improving Academic Event Recommendation Using Research Similarity and Interaction Strength Between Authors. Cybernetics and Systems, 48(3), 210-230.
- Ley, M. (2009). DBLP: some lessons learned. Proceedings of the VLDB Endowment, 2(2), 1493-1500.
- Khrouf, H., and R. Troncy. 2013. Hybrid event recommendation using linked data and user diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems*, ACM, 185–92.
- Luong, N. T., Nguyen, T. T., Hwang, D., Lee, C. H., & Jung, J. J. (2015). Similarity-based Complex Publication Network Analytics for Recommending Potential Collaborations. J. UCS, 21(6), 871-889.
- Sielis, G. A., Tzanavari, A., & Papadopoulos, G. A. (2015). Recommender systems review of types, techniques, and applications. In Encyclopedia of Information Science and Technology, Third Edition (pp. 7260-7270). IGI Global.
- Xia, F., Chen, Z., Wang, W., Li, J., & Yang, L. T. (2014). Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. IEEE Transactions on Emerging Topics in Computing, 2(3), 364-375.
- Smart, J., & Bayer, A. (1986). Author collaboration and impact: A note on citation rates of single and multiple authored articles. Scientometrics, 10(5-6), 297-305.
- Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., & Han, J. (2011, July). Co-author relationship prediction in heterogeneous bibliographic networks. In Advances in Social

Networks Analysis and Mining (ASONAM), 2011 International Conference on (pp. 121-128). IEEE.5.

- Batra M., Sharma S., Prof A., Of D., Applications C., and Rachna M., "Comparative Study of Page Rank Algorithm With Different Ranking Algorithms Adopted By Search Engine For Website Ranking," *Int. J. Comput. Technol. Appl.*, vol. 4, no. 1, pp. 8–18, 2013.
- Aggarwal C. C., (2016), Recommender Systems, Springe publisher, IBM T.J. Watson Research Center Yorktown Heights, NY, USA.
- Francesco R., Rokach L., Shapira B. Kantor P. B., (2011), Recommender Systems Handbook, Springer Science and Business Media, LLC, Springer New York Dordrecht Heidelberg London.