2010

Unraveling the Relationship between Co-Authorship and Research Interest

Jiejia Lin
Kingdee International Software Group Co. Ltd, jiejialin@foxmail.com

Yunhong Xu
University of Science and Technology of China, xuyunhong@mail.ustc.edu.cn

Shujin Cao
Sun Yat-sen University, caosj@mail.sysu.edu.cn

UNRAVELING THE RELATIONSHIP BETWEEN CO-AUTHORSHIP AND RESEARCH INTEREST

Jiejia Lin, Kingdee International Software Group Co.Ltd, High-tech Industrial Park, Nanshan, Shenzhen, GuangDong, PRC. jiejialin@foxmail.com

Yunhong Xu, Management School, University of Science and Technology of China, Heifei, AnHui, PRC. xuyunhong@mail.ustc.edu.cn

Shujin Cao, School of Information Management, Sun Yat-sen University, Guangzhou, GuangDong, PRC. caosj@mail.sysu.edu.cn

Abstract

Co-authorship in scientific research is increasing in the past decades. There are lots of researches focusing on the pattern of co-authorship by using social network analysis. However, most of them merely concentrated on the properties of graphs or networks rather than take the contribution of authors to publications and the semantic information of publications into consideration. In this paper, we employ a contribution index to weight word vectors generated from publications so as to represent authors' research interest, and try to explore the relationship between research interest and co-authorship. Result of curve estimation indicates that research interest couldn't be employed to predict co-authorship. Therefore,graph-based researcher recommendation needs further examination.

Keywords: co-authorships, contribution index, research interest

1 INTRODUCTION

Co-authorship in scientific research has been increasing for the past decades. It is one of the most well documented forms of scientific collaboration, which can be easily observed whenever we retrieve publications in academic databases, such as ISI Web of Knowledge, IEEE, ACM and so forth(Newman, 2004). The prosperity of co-authorship results from many reasons. First of all, due to the complexity of scientific research, collaboration among different scientific areas is conducive to solve persistent problems that beyond the competence of a single researcher and institution. Secondly, the fast development of computer and network technologies make it easier for researchers to construct collaborative relationship. Collaboration could benefits various research communities by promoting the development of theories.However, this trend of increasing co-authorship has raised researchers' attention (Kretschmer, 2004). On one hand, the increasing number of authors per paper could give birth to academic parasitism(Ioannidis, 2008). On the other hand, inspirational activity in scientific research could be sabotaged by dishonesty of senior collaborators who use their experience and academic power to distort the membership and order of authors on publications(Kwok, 2005). Thus, investigating co-authorship of researchers is necessary, which could bring about two main benefits. On the individual level, testifying co-authorship could eliminate academic misconduct, and produce great incentive to creativities of young researchers. Besides, making sure of the reliability of co-authorship networks could provide firm support for network-based researcher recommendation. On the organizational level, the benefits could be more obvious. In many organizations, researchers' academic competency is usually measured by the quantity and quality of his/her publications. But only evaluating the competency of researchers by the number of publications is not enough. Testifying co-authorships in publications would make the evaluation more objective and fair. However, there are two challenges on testifying co-authorship,namely:

- Measuring contribution of coauthors: In practice, contribution of author may be measured by referring to the order of coauthors in author list. Document(Slone, 1996) has provide empirical evidence for this.
- Representing co-author's research interest: In this paper, we employ word vectors generated from abstracts of publications to represent research interest.

Common sense has it that researchers who share similar research interest may collaborate with each other much more frequent than those who don't. In this research, the hypothesis is that research interest is significantly correlated to coauthor frequency.

1.1 Related Work

1.1.1 Quantify Contribution of Author in Publications

Co-authorship in scientific publications has dramatically increased over the past decades, while the number of publications with one or two authors decreases gradually.A publication would have more than on author due to the complexity of research topic. However, it doesn't mean that every author in publication is qualified, nor it doesn't means that all the authors contribute to the research equivalently. To be a qualified coauthor, researchers should have done one or more of the following: provide the idea, design the protocol, play a leadership role in the acquisition of the data, execute the study, analyze the data, review the literature, and/or write and revises the manuscript (Berk, 1989). Though these criteria are proposed as principles, they are not written rule. Even though most of the researchers would abide by these principles. There still exists the possibility that co-authorship is abused because of the asymmetry of academic power among coauthors(Friedman, 1993). Moreover, some metrics, such as impact factor, H-index, citation times and so forth, fail to take the contribution of coauthors into account, therefore bring about negative effect when they are used to select candidates or to hire faculty by institutions (Sekercioglu, 2008). Because these measurements may make researchers change their research strategy to avoid taking risk, to slice the findings up as much as possible, to compress the results, to push themselves to co-author publications with others, which results in larger and larger co-author group(Lawrence, 2007). Therefore, it's imperative to take the contribution of authors into account when evaluating the academic competence of researchers rather than merely depending on the numbers of their publications. Obviously, using author rank as a weighting factor is a simple, intuitive, transparent, and objective approach. Richard M. Slone's research indicates that there's a strong correlation between authorship position and contribution($r = -.69$, $p < .001$) (Slone, 1996).

1.1.2 Social Network Analysis in Co-authorships Research

A social network is a social structure between actors(nodes), which is viewed as graph in term of a mathematic structure. These actors(nodes) could be individuals or organizations. A social network is usually represented as a pair (V, E) where V is a set of actors(nodes) or vertices and E is a set of edges or links (Said et al., 2008). It can be represented in several ways, including mathematics-based formal methods, statistical models, matrices with either binary or weighted value,etc. Usually, social network analysis focuses on mapping and measuring of relationships and flows between peoples, groups, organizations(Jamali and Abolhassani, 2006). As a method for studying social structure, social network analysis is widely applied in various domains, including knowledge management in enterprise(Fan et al., 2008), graph-based recommendation in customized search engines, personalized shopping agents, and E-commerce(Naderi et al., 2008, Mirza et al., 2003), author-coauthor relationships, co-author network pattern, co-citation analysis in scientific research(Newman, 2004), etc. Research on social network analysis places lots of emphasis on investigating properties of graph, such as the degree, closeness, betweenness of vertices. These properties are used to measure the centrality of vertices and the centralization of networks. Basing on this, the influence, productivity and performance of a certain actor(node) in a network is quantified. For example, the number of publications per author is often used as the productivity measurement to group authors in social network analysis. (Kretschmer, 2004). Similarly, co-authoring and co-citation of scientific publications are treated as tangible evidences of formal work relations to measure scientific performance (Mika et al., 2006). However, merely basing on the properties of network while completely disregarding the content of publications of authors could be biased (Viermetz and Skubacz, 2007). Extension of network analysis to include content is a conclusive and relevant step to further understanding co-author networks. With these semantic information integrated we can yield a clearer and more differentiated view of network data, especially when large texts of publications, such as keywords, abstracts are taken into consideration(Viermetz and Skubacz, 2007). In this research, we study the co-authorship among researchers from two aspects. On the one hand, we measure the contribution of coauthors to form a contribution index by referring to the order of authors' name. On the other hand, this contribution index is used as weighting factor to weight word vectors generated from abstracts of publication to signify coauthors' research interest.

1.2 Contributions

This paper has two primary research contributions:

- Finding out a way to quantify the contribution of authors for publications so as to evaluate a researcher's competency fairly.
- By investigating on the relationship between co-authorship and research similarity, we could provide supports to graph-based researcher recommendation.

1.3 Organization

Section 2 illustrates the metric used to quantifying contribution of authors in publications and the way to represent research interest of coauthors. Section 3 provides details of how to collect the data and to testify the relationship between co-authorships and research interest. Result of our experimental work is also presented and dicussed. Section 4 presents concluding remarks and future research.

2 CONTRIBUTION AND RESEARCH INTEREST

2.1 Quantify Contribution for Coauthors

Order of author name in publication is used to quantify the author's contribution. **DEFINITION 1: For authors in any publication, the contribution of authors to the publication is:**

$$S_k = 1/(k*H_n) \qquad (1)$$

where ***k*** is the author rank, ***n*** is the number of coauthors, and

$$H_n = \sum_{k=1}^{n} 1/k \qquad (2)$$

The k^{th} ranked author's contribution is considered as *1/k* as much as the first author. And the sum of

contributions for all coauthors in a publication is always 1 (Sekercioglu, 2008).

2.2 Representing Research Interest for Coauthors

Research interest of individuals is highly dynamic and unstructured. Therefore, it's difficult to portrait research interest for authors in a long period of time. Research interest could be either explicit or tacit. Explicit research interest can be represented in the form of word vectors in a short time, while tacit knowledge is articulated in individual's experience, which is much harder to be captured. Due to the intangibility of tacit knowledge and the difficulty on collecting data needed, we prefer word vector to represent research interest for researchers. Ways of representing research interest varies in different subjects. In IS, techniques such as data mining, text mining and clustering algorithm are used to discover shared interest of researchers, while in social network analysis, collaborative information is utilized. For example, Naderi, etc proposes a graph-based profile similarity calculation method for collaborative information retrieval. In their research, user profile is presented as a set of pairs(q, D_q) in which q is a query and D_q is a set of relevant documents (Naderi et al., 2008). Hungarian algorithm on this weighted bipartite graph is used to calculate the maximum weighted matching between user profiles. Another related research focuses on calculating the semantic similarity among researchers. Citation network and author citation network which center around the important figures in a certain research field are taken into consideration. In the research, a compound document method is employed. The compound document X_i is represented as:

$$\mathbf{X}_i=\lambda_{i1}d_1+\ldots+\lambda_{iJ}d_J=\begin{matrix}\lambda_{i1}w_{11} & \lambda_{i1}w_{12} & \cdots & \lambda_{i1}w_{1K} & d_1\\ \lambda_{i2}w_{21} & \lambda_{i2}w_{22} & \cdots & \lambda_{i2}w_{2K} & d_2\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \lambda_{iJ}w_{J1} & \lambda_{iJ}w_{J2} & \cdots & \lambda_{iJ}w_{JK} & d_J\\ a_1 & a_2 & \cdots & a_K & \end{matrix} \quad (3)$$

Where $\boldsymbol{\lambda_{ij}}$ is the weight of $\mathbf{d_j}$ (the j^{th} document) in $\mathbf{X_i}$, and "+" is not the sign for addition of vectors, but a sign of form. Compound document vectors are employed to define semantic similarity among network members (Jiang et al., 2008). **DEFINITION 2: For any author, the research interest is denoted by using the contribution of the author for a particular publication as a weighting factor to multiply the value of word vector, and then word vectors that generated from all the publications of the author are aggregated.**

$$\textbf{Research Interest} = \lambda_1\textbf{Vect}_1 + \lambda_2\textbf{Vect}_2 + \ldots + \lambda_n\textbf{Vect}_n \quad (4)$$

Where ***n*** is the number of publications an author has published, and λ_i equals to the contribution index S_k of authors. The $Vect_i$ is a world vector, which is generated by extracting the words from abstract of a publication. By employing the contribution index, we manage to differentiate the weight of the words in vectors for different authors in every single publication.

2.3 Measuring Research Similarity among Coauthors

We employ Pearson Coefficient of word vectors to measure research similarity for coauthors with a target author. Besides, research similarity could also be measured by the following definition.**DEFINITION 3**: **For any couple of coauthors in a set of publications, the similarity between them could be denoted by Jaccard coefficient, which could be calculated by the formula below.**

$$\text{Similarity Score}\,(a,b)=\frac{|A\cap B|}{|A\cup B|} \quad (5)$$

Where $|A\cap B|$ denotes the number of authors who have collaborated with both A and B, while $|A\cup B|$ denotes the number of authors who either collaborate with A or B (A and B are excluded). The more similar research interest two researchers share, the larger the similarity score may be.

3 EXPERIMENTAL PROCEDURE AND DATA PROCESSING

3.1 Experimental Procedure on Data Collecting and Processing

Step 1: We firstly collect publications for the target author, whose name and address are used to generate a

retrieval strategy for ISI web of Knowledge. The strategy is constructed as "*au=(chu pk) and ad=(city univ same hong kong) and py= 2008*".

Step 2: Because authors except for the target author may have published some other publications. Therefore, we retrieve all the publications for them as well. Due to the possibility that the names of authors may overlap, if an author is not mapped to the right institution, the corresponding publications are discarded.

Step 3: Basing on the pulications, we calculate the contribution index for every author in publications and form a coauthor matrix. Then by using formula 5, we calculate the similarity score for coauthors with the target author. We then calculate the collaborative frequency for authors with the target author. Collaborative frequency is represented as {*'Chu, PK'*:{$coauthor_1$:$frequency_1$, $coauthor_2$: $frequency_2$, …, $coauthor_n$: $frequency_n$} }

Step 4: In addition to the coauthor matrix and the collaborative frequency dictionary, word vectors are extracted from abstracts and stored in form of dictionaries, that is {*Pub_id* : {$word_1$: $frequency_1$, $word_2$: $frequency_2$, …, $word_n$: $frequency_n$}}. Some words may contribute little to representation of research interest, they are usually called stop words. In order to minimize their negative effect, we filter these words. Because morphological variants of words would have similar semantic interpretations, we stem the words with Porter's word stemming algorithm. Subsequently, the word frequency of word vectors is weighted by the contribution index of author. Finally we form a compound word vector for coauthor by formula 4. And the compound word vector is represented as {*Author:*{$word_1$:$frequency_1$*S_k, $word_2$: $frequency_2$*S_k, …, $word_n$: $frequency_n$*S_k}}. After this, TF-IDF algorithm is employed to further differentiate the importance for words. Subsequently, we use Pearson coefficient to calculate the similarity of research interest among coauthors with the target author.

Step 5: Finally Pearson coefficient is used as independent variable, while collaborative frequency and Jaccard coefficient as dependent variable respectively to conduct curve estimation to verify if research interest could predict co-author frequency and the scale of coauthor network.

3.2 Experimental Results

3.2.1 Descriptive Statistics of Data

We collected 376 different publications in year 2008, in which there are 68 publications that are directly related to *Chu, PK*. Only one of them was authored merely on *Chu, PK*'s own. However, not every author in 376 publications has collaborated with *Chu, PK*, he/she may just share the same coauthors with *Chu, PK*. Therefore, we eliminate these authors from our dataset. At last, there are 158 out of 879 authors who have collaborated directly with Mr. CPK in year 2008.

3.2.2 Changing Parameter for Adjusted TF-IDF Algorithm

We employed TF-IDF algorithm to conduct feature selection so as to get a more accurate representation of research interest for authors. Because the threshold of TF-IDF may influence the selection of words, we started the threshold with 0.6, and gradually increase it by 0.2 until 1.2. The TF-IDF formula is as follow:

$$\mathbf{WordWeight = TF*IDF}$$
$$\mathbf{= ((0.5 + 0.5*Term\_Frequency)/max(Term\_Frequency))*Log_{10}(N/Term\_Appearance)} \quad (6)$$

where Term_Frequency represents the frequency of word, **N** is the number of authors, Term_Appearance represents how many times a word appears in different authors. 0.5 is used to adjust the TF-IDF Weight.

3.2.3 Results of Curve Estimation

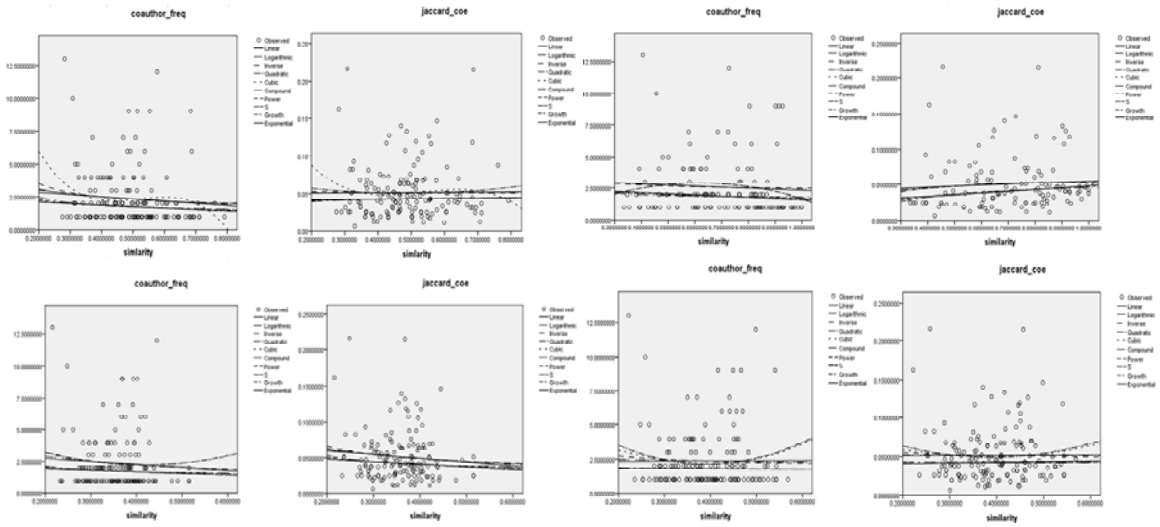

Figure 1. Curve estimation when threshold of TF-IDF ranges from 0.6 to 1.2

As can be seen from figure 1, no matter how we change the threshold of TF-IDF algorithm, the scatter plots imply that there isn't proper model to reflect the relationship between similarity of research interest and collaborative frequency or Jaccard coefficien. Beside scatter plot, R-Squared coefficient is usually used to judge how good a variable is at predicting another. According to our experiment, none of the curve models(10 models: Linear, Logarithmic, Inverse, Quadratic, Cubic, Compound, Power, S, Growth, Exponential.) has a good enough R-Squared coefficient and none of the significant coefficients is smaller than 0.05. Thus, research interest couldn't predict co-authorship well.

3.2.4 Discussion

The result shows that there's no significant relationship between research interest and co-authorship. There may be several reasons. First of all, we only explore the publications of coauthors in year 2008. More publications would be needed so as to fully portrait the research interest of authors. However, the more publications we collect in a longer period of time, the more possible that we will suffer from "Topic drift" (Jiang et al., 2008). Secondly, authors could have serveral close research interest, thus would collaborate with different groups of people. It's possible that some of the collaborators cooperate with the target author frequently while the others don't, though these collaborators may share similar research interest. Thus, using Jaccard Coefficient to measure similarity of research interest could cause insignificant results, research interest correlation may be found low while collaborative frequency is found high.Thirdly, only employing keywords to represent research interest may be inadequate. Some other factors should be taken into consideration, such as co-citation, research interest descriped by authors themselves and so on. Besides, other ways of feature selection for word vectors could be used to help get a more accurate representation of research interest. Thus help to improve the result.

4 CONCLUSION AND FUTURE RESEARCH

In this paper, we take the semantic information of publications into consideration, and employ contribution index to quantify the contribution of authors to publications. By using the semantic information, we try to unravel the relationship between research interest and co-authorship. Though, the insignificance of curve estimation may be caused by the listed reasons above, however, what if the result truly reflects the situation, in this sense, we may conclude that there exists possible undeserved co-authorship. Therefore, graph-based researcher recommendation deserves further examination.

In the future, we will dedicate in employing co-citation, authors' description of research interest and some other feature selection methods, such as Chi-Square, Information Gains, Principal Component Analysis, to better discover the relationship between co-authorship and similarity of research interest.

5 ACKNOWLEDGEMANT

We would like to acknowledge the kind help of Tao Yang from Indiana University for his precious modification advices and the help of data collection of JingJing Wu from IRIS(ShenZhen) software co. ltd.

References:

Berk, R. N. (1989) Irresponsible Coauthorship.American Journal of Roentgenology, American Journal of Roentgenology, 152, 719-720.

Fan, B., Liu, L., Li, M. and Wu, Y. (2008) Knowledge Recommendation Based on Social Network Theory.In 2008 Ieee Symposium on Advanced Management of Information for Globalized Enterprises, ProceedingsIeee, New York, pp. 322-324.

Friedman, P. J. (1993) Standards for Authorship and Publication in Academic Radiology - Aur Ad-Hoc Committee on Standards for the Responsible Conduct of Research. Investigative Radiology, 28, 879-881.

Ioannidis, J. P. A. (2008) Measuring Co-Authorship and Networking-Adjusted Scientific Impact. Plos One, 3, 8.

Jamali, M. and Abolhassani, H. (2006) Different aspects of social network analysis. In 2006 IEEE/WIC/ACM International Conference on Web IntelligenceIeee Computer Soc, Los Alamitos, pp. 66-72.

Jiang, K. Z., Wu, Y. Q., Lv, Z. and Gu, J. Z. (2008) Research on author's semantic similarity based on collaborative network.In Proceedings of the Fifth International Conference on Information Technology: New Generations(Ed, Latifi, S.) Ieee Computer Soc, Los Alamitos, pp. 1034-1039.

Kretschmer, H. (2004) Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web.Scientometrics, 60, 409-420.

Kwok, L. S. (2005) The White Bull effect: abusive coauthorship and publication parasitism.Journal of Medical Ethics, 31, 554-556.

Lawrence, P. A. (2007) The mismeasurement of science.Current Biology, 17, R583-R585.

Mika, P., Elfring, T. and Groenewegen, P. (2006) Application of semantic technology for social network analysis in the sciences.Scientometrics, 68, 3-27.

Mirza, B. J., Keller, B. J. and Ramakrishnan, N. (2003) Studying recommendation algorithms by graph analysis. Journal of Intelligent Information Systems, 20, 131-160.

Naderi, H., Rumpler, B. and Pinon, J. M. (2008) A Graph-based Profile Similarity Calculation Method for Collaborative Information Retrieval.In Proceedings of the 23rd Annual Acm Symposium on Applied ComputingAssoc Computing Machinery, New York, pp. 1127-1131.

Newman, M. E. J. (2004) Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America, 101, 5200-5205.

Said, Y. H., Wegman, E. J., Sharabati, W. K. and Rigsby, J. T. (2008) Social networks of author-coauthor relationships. Computational Statistics & Data Analysis, 52, 2177-2184.

Sekercioglu, C. H. (2008) Quantifying coauthor contributions.Science, 322, 371-371.

Slone, R. M. (1996) Coauthors' contributions to major papers published in the AJR: Frequency of undeserved coauthorship.American Journal of Roentgenology, 167, 571-579.

Viermetz, M. and Skubacz, M. (2007) Using topic discovery to segment large communication graphs for social network analysis.In Proceedings of the Ieee/Wic/Acm International Conference on Web Intelligence - Wi 2007(Ed, Haas, L. L. K. J. M. R. B. A. H. H.) Ieee Computer Soc, Los Alamitos, pp. 95-99.