Tracing and Predicting Collaboration for Junior Scholars

Chun-Hua Tsai
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260, USA
cht77@pitt.edu

Yu-Ru Lin
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260, USA
yurulin@pitt.edu

ABSTRACT

Academic publication is a key indicator for measuring scholars' scientific productivity and has a crucial impact on their future career. Previous work has identified the positive association between the number of collaborators and academic productivity, which motivates the problem of tracing and predicting potential collaborators for junior scholars. Nevertheless, the insufficient publication record makes current approaches less effective for junior scholars. In this paper, we present an exploratory study of predicting junior scholars' future co-authorship in three different network density. By combining features based on affiliation, geographic and content information, the proposed model significantly outperforms the baseline methods by 12% in terms of sensitivity. Furthermore, the experiment result shows the association between network density and feature selection strategy. Our study sheds light on the re-evaluation of existing approaches to connect scholars in the emerging worldwide Web of Scholars.

Categories and Subject Descriptors

I.5.2 [**Computing Methodologies**]: Design Methodology—*classifier design and evaluation, feature evaluation and selection*

General Terms

Measurement; Experimentation

Keywords

collaboration; cold-start; link prediction

1. INTRODUCTION

Academic publications are critical to assess scholars' scientific productivity. However, in our analysis, there is only 8.8% of scholars keep publishing after 6 years. To help the newcomer of academia, the studies of [8, 22] indicate the importance of a junior scholar to identify and maintain appropriate collaboration relations in their early professional careers. Meanwhile, researchers around the world are currently producing more and more scholarly data over web digital database. The abundant and large-scale of bibliographic information have been proven useful in connecting scholars [21] and predicting future productivity [22] through the Web. However, for junior scholars, it is challenging to predict or suggest the effective collaboration with a limited academic record (publications). Further, the rapid growth of global scholarly data imposes a challenging of filtering the meaningful information for the academic newcomer. These constraints make the current network-based collaborative prediction approaches less useful. This inspires us to trace the productivity of junior scholars from a different perspective.

In this work-in-progress paper, we present an exploratory study of the high-utility features for junior scholars to predict the future co-authorship in three evolving networks. The experiment result indicates the prediction model performance is negatively correlated with the network density, i.e. with the growth of network density, the utility of prediction features are decreasing. Besides, the network-based features perform well in predicting the future co-authorship, but with lower model sensitivity (recall score). In other words, simply adopting the network-based features for junior scholar co-authorship prediction is insufficient. We then improve the prediction model by integrating the content, affiliation and geographic features. Our experiment result shows the best-tuned model improves the baseline model by 12% of sensitivity measurement in low-density network.

The major contribution of this paper is that we provide a solid co-authorship predictor for junior scholars in multiple network density settings. In a rapid growth web digital library, researchers might sample the bibliographic data in different conditions, e.g. discipline, time and geography. It is unrealistic to adopt the same prediction features in all tasks. The criteria generate different network density and require the corresponding feature selection strategy. This finding is critical for scholarly recommendation system in the emerging worldwide Web of Scholars.

In the remainder of this paper, we firstly review the related work of junior scholars and link prediction in section 2. In section 3, we describe our dataset, feature selection and experiment setting. The experimental result will be discussed in section 4. Finally, we summarize our findings and future research directions.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
Include the http://dx.doi.org/10.1145/2872518.2890516.

2. RELATED WORK

Collaboration prediction between two scholars is a critical function of scholarly collaboration recommendation system. This is often studied as a link prediction problem. The goal is to identify the probability of future co-authorship [9]. There are different models to solve this problem, e.g. the unsupervised approach by citation network [6], the supervised learning approach by the global and local network [8] and social network-based features [17, 9, 20]. The predictive feature is varied in different link prediction problems, e.g. 1) adopts the publish sequence, topic and language model to find out the domain experts [10, 5, 4]; 2) considered the organizational overlap from the SNS (Social Network Service) user profile to predict the social relation [11, 12].

In prediction tasks, to extract suitable features in different link prediction problems is essential. The challenges for junior scholar of co-author link prediction are the data sparseness [8], which is also known as the cold start link prediction problem. The studies from [14] have adopted a network-based bootstrap probabilistic graph to predict the possible future network links. Their finding indicates that the precision is worse than the non-cold start link prediction problems. Hence, the typical link prediction features are not enough for cold start link prediction problem. It is necessary to examine effective features for different research problems.

3. PREDICTING COLLABORATION

Our research goal is to develop useful predictors that can be used to suggest suitable collaborators to facilitate junior scholars' early research stages. These problems have been discussed in several studies. However, the current link prediction method is inadequate for junior scholars due to lack of sufficient publication records. In this section, we propose the prediction features from proximity network, string distance, content similarity, and geographic distance feature to solve these constraints.

3.1 Problem Statement

We focus on the link prediction problem for junior scholar collaboration. We define a social network as G = (V, E), in which each edge $e = (u, v) \in E$ is an interaction between authors u and v at a particular time t. Here, the interaction is defined as collaboration in terms of coauthoring an academic publication. We can construct a collaboration network based on publication coauthoring information. Our goal is to predict the future co-authorship at $t' > t$. In other words, the goal is to find a future collaborative link that will be formed at the future time t' based on data observed at t. We can treat this as a binary classification problem that to distinguish the positive and negative links at the future time t'.

3.2 Data Description

We retrieved 247,147 publications from ACM Digital Library. These papers were published from 1990 to 2011 and included title, author, abstract, citation and affiliation information. Figure 1(a) shows the trend of papers, authors and junior authors with a yearly growth between 1990 and 2011. The distribution is left-skewed that most of the papers published after 2000, this trend also meets the positive growth of scholarly data over web digital database.

We characterize how junior scholars continuously publish academic works by computing the "retention rate" [22] of a set of a junior scholar. We first define Junior Scholar J_y as the set of authors who published their first paper in year y. The retention rate is computed based on whether or not the scholar has, at least, one paper within each of the next three or more years. Let $J_y^{(d)}$ be the set of authors in J_y who published continually, at least, one paper within each year between years $y+1$ and $y+d$. We define the retention rate $R_y^{(d)}$ as:

$$R_y^{(d)} = \frac{|J_y^{(d)}|}{|J_y|}.$$

Figure 1(b) shows the pattern of juniors who continued to publish papers in ACM conference in the next 3, 4, 5, and 6 years. Retention rate decreases naturally because the retained scholars in the next $k+1$ years are a subset of retained scholars in the next k years (who keep publishing for an additional year). With the passage of time, there are fewer and fewer scholars remaining the academic publication record. The junior scholars who start publish at 2004, 17.73% of them continue to publish in the next 3 years and only 8.8% continue to publish in the next 6 years. This gap increases from around 6.46% in 2000 to around 8.9% in 2004. The retention rate is lower than the finding in [22]. This is caused by ACM dataset only cover the single publisher, the probability of scholars keep publishing in one single publisher is lower.

3.3 Feature Extraction

For link prediction, it is critical to extracting the features that represent some properties between two paired nodes in a network. In this experiment, we consider 4 classic network-based features [16] as baseline:

Common Neighbors (CN): The CN [19] indicates the intersection set of neighbors of a given author. Here we define the set of neighbors as all co-authors observed at t.

Jaccard Coefficient (JC): The JC [3] measures similarity between finite neighbor sets. Here we defined neighbors sets as co-authors sets at t. For any two given authors, it is the intersection of their co-authors sets divided by the union of their co-authors sets.

Adamic/Adar (AA): The AA [1] is a typical local network similarity measurement that considers a weighting parameter between network nodes.

Katz_Weighted (KatzW): the KatzW [13] measure the direct sums over a collection of paths between two network nodes, but exponentially damping by length to count short paths more heavily. We defined the path as the collaboration network path of any two authors.

The above features have been found effective in prior work for general link prediction task [9, 16, 8, 11]. However, these features could be ineffective in the junior scholar collaboration link prediction task because those features consider only the network proximity properties. For junior scholars, the collaboration network is limited in their early research age. Hence, the network proximity approaches can only get a portion range of predictive power. Besides, the data sparsity issue is another challenge for junior scholar link prediction. To overcome these challenges, we propose additional features:

Affiliation Overlap (AO): The AO is to measure the similarity of two authors' career trajectories. We have equation

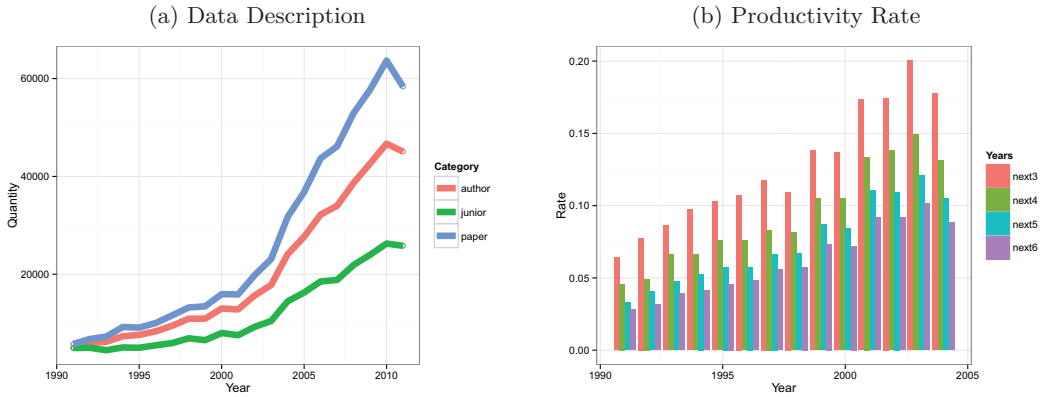

Figure 1: (a) Publication trend: the numbers of publications, authors and junior authors from 1990 to 2011 of ACM dataset; (b) The retention rate for next 3, 4, 5, and 6 years for scholars starting at year between 1990 and 2004 of ACM dataset. The retention rate is computed based on whether or not the scholar has, at least, one paper within each of the next three or more years.

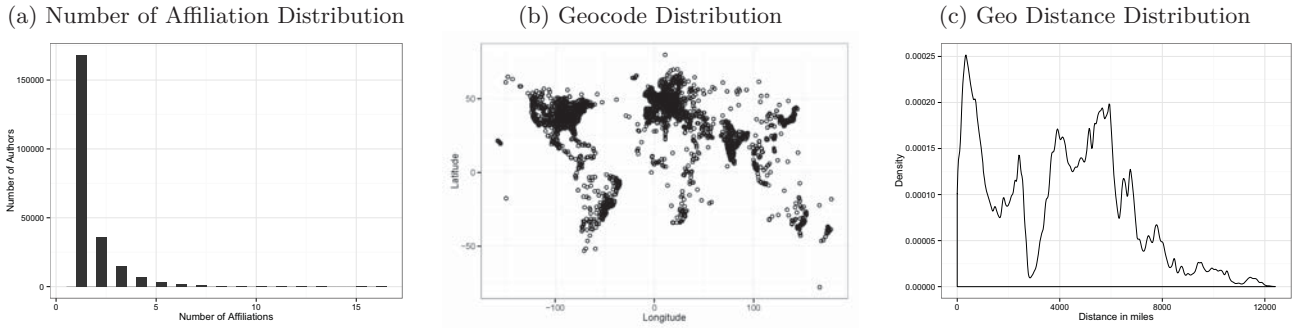

Figure 2: (a) A number of affiliation distribution of ACM dataset. Most of the authors belong to one single affiliation. (b) ACM dataset geocodes distribution: x axis present the longitude from -180 to +180, y axis represents the latitude from -90 to +90. The western world published more papers than other regions. (c) ACM dataset density distribution of geographic distance by author-pairs, the distance is in miles. The second peak around 4000 miles corresponds to the long distance transnational co-work over the oceans.

$Sim_{AO}(x, y) = \|x \cap y\|/\|x \cup y\|$, where x, y is the affiliation list of each author. This feature is inspired by the research of [11] that argued the organizational overlap in the industry was a good feature to predict the similarity of two users' in social networking services (SNS). Hence, we consider the affiliation information between user's publication. Because the affiliation might change or one author might have multiple organizations, we extract a list [publication, affiliation] for each user and compute the similarity of these two lists. However, according to Figure 2(a), most authors are with only one affiliation. The overlapping of any two authors' career trajectories is expected to be sparse. Hence, we provide string distance to overcome this limitation.

String Distance (SD): the SD [15] is a string metric for measuring the difference between two string sequences. We have $Sim_{SD}(a, b) = lev_{a,b}(\|a\|, \|b\|)$, where a, b are the two authors' latest affiliation name. The $lev_{a,b}$ is defined by [15].

The rationale behind this feature comes from the inconsistent affiliation data formats on ACM website. For example, one author from "Computer Sciences at the University of Pittsburgh" can be represented as "Department of Computer Sciences, University of Pittsburgh", "Department of Computer Sciences, University of Pittsburgh, Pittsburgh, PA, USA" or only "University of Pittsburgh". This increases the difficulty of comparing the discipline similarity between authors. Hence, we adopt the string distance to compute the affiliation similarity. This approach considers the sequence between two strings to prevent possible format inconsistency. For example, two computer sciences scientists from two different universities can have shorter string distance, but those who with completely different discipline, organization and location will have long string distance.

Geographic distance and rank distance (GD): The GD is to measure the actual geographic distance between two authors. We used the Haversine formula to compute the geographic distance between two points on earth based on longitude and latitude data. We define the geographic data as each author's latest affiliation location. The idea behind this feature is the spatially clustered academic co-work. According to [7], the citation structure between spatially clustered and globally dispersed teamwork were different in Bioresearch papers. This suggested us that distance could influence the co-work properties in research works. In other words, the local connection around campus could be an important collaboration source in their early research stage, e.g. colleague and scholars in nearby universities.

In order to obtain geographic data, we used the Google Maps API to retrieve the latitude and longitude informa-

tion for each author. For example, if we have an author from "School of Information Sciences, University of Pittsburgh", we send this affiliation name as a query to Google Maps API, so as to retrieve a latitude and longitude set, e.g. (40.444353,-79.960835), which represent the geographic location of University of Pittsburgh. In ACM dataset, we have 612,786 [author,affiliation] paired geocodes. In Figure 2(b), the geocode distribution is represented as a world map, the most publications are published in the western world. Figure 2(c) shows the density distribution over the geographic distance between authors. The second peak point around 4000 miles corresponds to the long distance transnational co-work over the oceans. For example, scientists from America and Europe will have a long distance due to the gap of the Atlantic, so they are co-working at a long geographic distance. Hence, to prevent the influence of the non-uniform geographical distribution of population, we also consider the rank-based format to represent the geographic routing in social network [17].

Content Similarity (CS): the CS [18] is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between the strings. We define the formula as: $Sim_{CS}(x, y) = (x \cdot y)/\|x\|\|y\|$, where x, y belong to the tf-idf (term frequency-inverse document frequency) vector of each author's publication text, i.e. title and abstract. We used tf-idf to create the vector with a word frequency upper bound 0.5 and lower bound 0.01 to eliminate the common and rarely used words. We consider unigram and bigram term sequence to cover more term composition in academic publications. The rationale behind this feature is simple and straightforward: to consider the authors' writing text to compute the similarity between them. Moreover, the related work from [2] utilized content information to relieve the data sparseness. For junior scholars, their early stage publication can provide meaningful information in co-author prediction task.

3.4 Prediction via Binary Classification

There are many classification algorithms for the supervised classifier. In this experiment, the proposed features are combined by seven different classification methods and compare the performance. The classifiers include kNN, Logistic regression, Naive Bayes, Decision Tree, ADA, SVM, and SVM_Tuned. All the classifiers are implemented in R with the application packages.

4. EXPERIMENT RESULT

4.1 Experiment Setting

Let J_t be the set of authors who published their first paper in year t, $J_{t'}$ is the set of authors who published, at least, one paper in a year $t' = [t+1, t+3]$ since their junior starting year. To predict a collaborative link for a junior scholar, we divide J_t into two non-overlapping partitions. The first partition is selected as the training dataset and the later one as the test dataset. We use $J_{t'}$ as ground truth to generate the positive and negative link. The positive link is the authors who established, at least, one collaboration in J'_t and the negative link is those who has no collaboration record in J'_t.

We divide the whole dataset into three 4-year periods: the year 2000 to 2003, 2004 to 2007 and 2008 to 2011. We will predict the co-authorship in t' years, based on the feature information in time t. We also limited 4 hops ego network because this covered possible co-author links. However, the positive and negative links are unbalanced (the negative links are much more than the positive link). Hence, we randomly choose 1:1 positive and negative link to represent the performance of our proposed model. All the performance measure will be reported by average values of 10-fold cross-validation.

4.2 Classifiers

The Table 1 and 2 we show the performance of different classification methods on ACM dataset. We randomly sample the 1:1 size of positive and negative cases distinguishes the best classifier due to the unbalanced positive and negative links. In this 1:1 experiment setting and binary classification, we expect the performance should exceed random guessing (50%). Table 1 presents the baseline model with Common Neighbors (CN), Jaccard Coefficient (JC), Adamic/Adar (AA) and Katz_Weighted (KatzW) features. All the classification methods, except for kNN, have at least 67% accuracy, 70% precision, and 54% recall. On AUC metrics, ADA performed the best with an AUC of 96.25%. The metrics indicate this is a good classifier to distinguish the positive and negative cases. However, the value of recall is only around 56.95-74.11%. In other words, the baseline model is not sensitive enough to cover all the possible co-authorship.

Table 2 provides the experimental result of the proposed model with baseline and newly added features (All_Features), including Affiliation Overlap (AO), String Distance of affiliation name (SD), Geographic distance and ranking (GD) and Content Similarity of title and abstract (CS). The result shows ADA is still the best classifier in this experiment. In ADA classifier, the performance is 5-10% higher in accuracy & precision and 4-7% higher in recall than baseline model. This is a minor improvement of the performance due to the junior scholars' data limitation. However, the newly proposed features effectively increase the recall value. Our best model improved baseline features over 10% by Naive Bayes classifier of 2008. We also observed a similar pattern in the year 2000 and 2004. This finding indicates the proposed model was highly sensitive to the prediction task.

Baseline Model: CN + JC + AA + KatzW					
Year 2000: 172 positive and equal size negative case					
Classification Model	Accuracy	Precision	Recall	F-value	AUC
kNN	53.28%	75.72%	46.64%	44.50%	58.50%
Logistic Regression	75.28%	85.12%	70.34%	70.89%	93.14%
Naive Bayes	76.18%	85.74%	70.44%	72.52%	92.62%
Decision Tree	76.72%	81.37%	74.81%	**74.81%**	94.83%
ADA	78.11%	**87.51%**	74.11%	74.58%	**96.25%**
SVM	**78.98%**	84.09%	**74.17%**	73.73%	94.69%
SVM_Tuned	76.94%	85.98%	70.91%	73.18%	93.09%
Year 2004: 328 positive and equal size negative case					
Classification Model	Accuracy	Precision	Recall	F-value	AUC
kNN	53.32%	66.51%	40.93%	39.85%	59.39%
Logistic Regression	67.92%	75.30%	55.96%	56.22%	85.36%
Naive Bayes	**72.37%**	78.140	64.93%	66.33%	87.42%
Decision Tree	72.07%	79.31%	**66.10%**	**64.30%**	90.96%
ADA	71.22%	**84.58%**	56.56%	59.13%	**91.40%**
SVM	69.82%	80.99%	55.90%	57.63%	89.13%
SVM_Tuned	70.39%	80.99%	57.51%	59.00%	89.13%
Year 2008: 419 positive and equal size negative case					
Classification Model	Accuracy	Precision	Recall	F-value	AUC
kNN	52.76%	55.38%	46.83%	42.27%	59.60%
Logistic Regression	69.22%	76.05%	64.46%	62.88%	82.76%
Naive Bayes	68.95%	76.62%	58.92%	59.88%	81.61%
Decision Tree	69.68%	70.89%	**74.96%**	**69.72%**	82.64%
ADA	**70.34%**	**79.81%**	56.95%	59.64%	**86.25%**
SVM	68.05%	79.78%	54.56%	55.65%	86.08%
SVM_Tuned	68.63%	80.77%	56.25%	57.11%	85.27%

Table 1: Baseline model of 2000, 2004 and 2008 with seven classifiers.

Proposed model: baseline + all proposed features.					
2000: 172 positive and equal size negative case					
Classification Model	Accuracy	Precision	Recall	F-value	AUC
kNN	53.85%	76.94%	46.67%	43.62%	60.19%
Logistic Regression	71.85%	79.02%	71.43%	67.97%	93.86%
Naive Bayes	74.97%	80.09%	**80.91%**	75.74%	95.41%
Decision Tree	**80.85%**	87.00%	80.43%	**80.77%**	96.44%
ADA	77.63%	**87.44%**	74.40%	73.96%	**98.33%**
SVM	72.77%	82.09%	72.40%	69.05%	95.90%
SVM_Tuned	74.11%	82.05%	73.87%	70.58%	98.17%
2004: 328 positive and equal size negative case					
Classification Model	Accuracy	Precision	Recall	F-value	AUC
kNN	53.63%	64.43%	41.52%	40.50%	61.48%
Logistic Regression	70.78%	75.64%	70.63%	66.15%	91.96%
Naive Bayes	70.67%	76.83%	71.24%	67.03%	90.69%
Decision Tree	**80.66%**	87.17%	**73.93%**	**75.84%**	93.74%
ADA	76.15%	**88.72%**	64.51%	67.89%	**96.21%**
SVM	72.46%	80.48%	72.49%	68.55%	94.89%
SVM_Tuned	73.27%	81.45%	72.82%	69.15%	96.37
2008: 419 positive and equal size negative case					
Classification Model	Accuracy	Precision	Recall	F-value	AUC
kNN	53.29%	60.55%	38.86%	38.05%	59.66%
Logistic Regression	69.54%	76.45%	69.41%	65.28%	89.21%
Naive Bayes	69.19%	75.50%	69.25%	65.06%	88.05%
Decision Tree	**76.04%**	78.15%	**77.20%**	**74.31%**	90.47%
ADA	74.73%	**88.39%**	61.99%	65.86%	**94.01%**
SVM	71.11%	79.18%	71.09%	67.17%	92.39%
SVM_Tuned	71.69%	78.27%	71.42%	67.52%	93.58%

Table 2: Proposed model of 2000, 2004 and 2008 with seven classifiers.

4.3 Performance Evaluation

The All_Features model is with the highest the area under the curve (AUC) in all three year periods. The AUC metric is defined as the probability that a classifier will rank a randomly chosen positive case higher than a randomly chosen negative case. This is also a performance indicator to evaluate a better classification model. This curve presents the All_Features model provides the best results. The sensitivity of positive: negative ratio in Figure 3 also shows the All_Features model is with higher sensitivity than the baseline model of 2000, 2004 and 2008. The plot indicates the sensitivity metric changes between different positive and negative ratio settings. We also present a statistical test to reveal the significance of the experiment result (Table 3). The result indicates the proposed All_Features model is significantly better (p-value<0.001) than the baseline model of 2000, 2004 and 2008. The All_Features model is performed better in F-Value and AUC metrics than baseline and other single feature models.

We find that our proposed All_Features model performed better than a single feature model and the baseline model in 2008. The string distance (SD), affiliation overlaps (AO), Geo distance (GD) and content similarity (CS) features gain recall improvement by 11%, 6%, 12% & 8%. In other words, the SD, AO, GD and CS of author's affiliation information are significantly increasing the model sensitivity. However, although the All_Features model is significantly better than the baseline model in all three year periods. In the year of 2000 and 2004, only CS feature has the significant result in 2004. The rest features are not significant in the correlated statistical test. This result corresponds to the academic circle growth from 2000 to 2011. There are more authors and publications during the later years. This generates the diversity of academia that we need more different features to predict the possible co-authorship. In a lower network density setting (the year 2008), the new proposed features are performed better than higher network density (the year 2000 and 2004). This result also corresponds to the higher recall value in the early years (2000 and 2004) than the recent year (2008) due to the network-based feature has higher predictive power in high network density.

Based on the experimental result, we believe the network-based features are effective to cover the partial prediction power. In a high-density network (fewer author and publications), the network-based feature has better performance due to the authors are easier to connect each other with their neighbors. However, we need to append some extra information to extend the model sensitivity when the network complexity increased. I.e. the SD, AO, and GD features can improve the model sensitivity in a larger author and publication basis. These features are considered affiliation and geographic information to train the prediction model with higher sensitivity, but lose effectiveness when network complexity increased. Besides, the content based features (CS) are also significantly better than baseline model. For any two authors who do not connect by their collaborator, it is hard to predict the collaboration relation, even they share the same research interests. Hence, the content-based feature is an effective feature even the network is dense. This is also a feature that not constrained by data sparseness issue. For junior scholars, this would be a generally accessible information in their early research age. We consider it as a good predictor for junior scholars to prevent the data sparseness issue.

Year 2000						
	Accuracy	Precision	Recall	F-Socre	AUC	P-value
Baseline	76%	85%	69%	70%	92%	-
All_F	76%	86%	72%	71%	97%	0.0004**
SD	75%	82%	75%	72%	96%	0.0547
AO	76%	88%	66%	69%	94%	0.3502
GD	74%	82%	73%	71%	95%	0.1034
CS	76%	87%	69%	70%	96%	0.0618
Year 2004						
Baseline	71%	84%	57%	60%	91%	-
All_F	76%	90%	64%	68%	96%	0.0002**
SD	74%	86%	64%	66%	94%	0.1325
AO	72%	88%	55%	61%	93%	0.4317
GD	74%	86%	62%	66%	93%	0.2593
CS	76%	91%	63%	68%	96%	0.0003**
Year 2008						
Baseline	68%	80%	52%	55%	84%	-
All_F	74%	88%	62%	66%	94%	0.0000**
SD	71%	81%	63%	63%	88%	0.0109*
AO	71%	86%	58%	63%	88%	0.0097**
GD	71%	80%	64%	64%	89%	0.0055**
CS	73%	88%	60%	64%	93%	0.0000**

*: p-value<0.05; ** p-value<0.01; All_F=All_Features

Table 3: The correlated statistical test result of baseline and proposed model of the year 2000, 2004 and 2008.

5. CONCLUSION

In this paper, we proceed an exploratory study of the effective co-authorship predictor of junior scholars. In addition to the typical network-based prediction, we further consider the features from affiliation, geographic and content information. The experiment results show the effective prediction performance in different network of density. We suggest in a low-density network environment, the junior scholars require non-network-based features to extend the collaboration prediction performance. In the experiment, the best-tuned model can increase the model sensitivity by 12% and best model AUC over 10%. This finding supports, at the early stage of junior scholars, the proposed non-network-based features would be more useful and easily accessible to fulfill the prediction task.

In summary, this paper provides a solid co-authorship predictor evaluation for junior scholars in multiple network density settings. We suggest the feature selection strategy in different sampling criteria is varied. This finding is critical for a scholarly recommendation system which fetch data

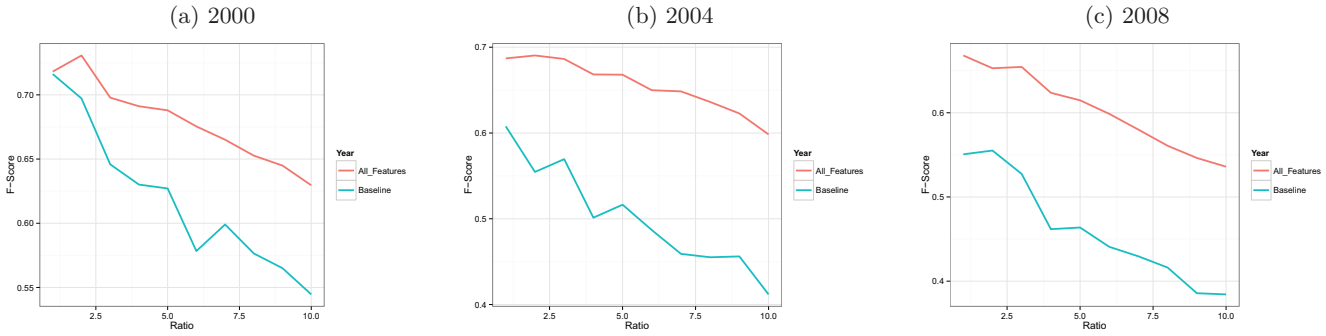

Figure 3: Prediction performance over the ratios of negative vs. positive links (from 1:1 to 10:1). The plots indicate that the performance decreases with the increase of negative links, and our proposed features perform significantly better even with the large portion of negative links. Especially, in higher network density, the proposed model outperformed the baseline model.

from web digital libraries. Our study sheds some light on the re-evaluation of existing approaches to associate scholars in the emerging worldwide Web of Scholars.

In future work, we plan to extend the current work to a social support system for conference new comers (e.g. Junior scholar). The system will be personalized by the social needs of junior and senior scholars in research community.

6. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] K. Balog, L. Azzopardi, and M. De Rijke. Formal models for expert finding in enterprise corpora. In *ACM SIGIR*, pages 43–50. ACM, 2006.

[3] G. Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[4] G. Cormode, S. Muthukrishnan, and J. Yan. People like us: Mining scholarly data for comparable researchers. In *Proc. WWW Companion*, WWW Companion '14, pages 1227–1232. Intl. WWW Conferences Steering Committee, 2014.

[5] H. Deng, I. King, and M. R. Lyu. Formal models for expert finding on DBLP bibliography data. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 163–172. IEEE, 2008.

[6] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proc. of ICML*, pages 233–240, New York, NY, USA, 2007. ACM.

[7] M. Gittelman. Does geography matter for science-based firms? epistemic communities and the geography of research and patenting in biotechnology. *Organization Science*, 18(4):724–741, 2007.

[8] S. Han, D. He, P. Brusilovsky, and Z. Yue. Coauthor prediction for junior researchers. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 274–283. Springer, 2013.

[9] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[10] S. H. Hashemi, M. Neshati, and H. Beigy. Expertise retrieval in bibliographic network: a topic dominance learning approach. In *Proc. of CIKM*, pages 1117–1126. ACM, 2013.

[11] C.-J. Hsieh, M. Tiwari, D. Agarwal, X. L. Huang, and S. Shah. Organizational overlap on social networks and its applications. In *Proc. of WWW*, WWW '13, pages 571–582. Intl. WWW Conferences Steering Committee, 2013.

[12] M. Karimzadehgan, R. W. White, and M. Richardson. Enhancing expert finding using organizational hierarchies. In *Advances in Information Retrieval*, pages 177–188. Springer, 2009.

[13] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[14] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *ACM SIGKDD*, KDD '10, pages 393–402. ACM, 2010.

[15] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.

[16] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559. ACM, 2003.

[17] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. 102(33):11623–11628, 2005.

[18] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[19] M. E. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

[20] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *ASONAM*, pages 121–128, 2011.

[21] C.-H. Tsai and P. Brusilovsky. A personalized people recommender system using global search approach. *iConference 2016 Proceedings*, 2016.

[22] C.-H. Tsai and Y.-R. Lin. The evolution of scientific productivity of junior scholars. *iConference 2015 Proceedings*, 2015.