International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

A Novel Co-authorship Prediction Model using Semantic Clustering and Supervised Prediction

Sivakumar P
Computer Science and Engineering
Govt. Engg. College, Thrissur
Kerala, India

Vipin Kumar K.S
Computer Science and Engineering
Govt. Engg. College, Thrissur
Kerala, India

***Abstract*— Co-authorship prediction has been studied in researches as a part of social network analysis. Co-author prediction is the problem of predicting missing or future links (collaboratons) between authors. Previous studies have dealt with this problem and have proposed various approaches. Out of these, there are mainly two approaches: similarity based and learning-based. The former approach uses similarity metrics between authors such as common neighbor, random walks, etc and rank them while the latter treats co-author prediction as binary classification and uses learning models with similarity metrics as features. In this work, we propose a novel co-authorship prediction model based on semantic clustering and supervised learning. We test our proposed model with some other keyword-based predictors and the results show that our predictor performs averagely better than the comparison predictors .**

Keywords*— Co-authorship prediction model, supervised learning ,semantic clustering.

I. INTRODUCTION

Social network mining has been a widely researched topic for the last few decades.A social network can be considered as a graph where links denote relationships(collaborations) and nodes represent actors(people).Social network analysis include mining useful insight like the underlying structure, types of interest groups, finding sentiments of people, or deep mining of patterns for future use. Two main concepts used are : community detection and link prediction.

Community detection is the problem of finding underlying community patterns or structure in a social network i.e. finding tightly knit groups of people.It can be overlapping or disjoint communities.The former has members participating in multiple communities, while the latter has a member confined to single community. Some of the methods proposed over years are Clique percolation, divisive and hierarchical clustering, modularity optimization, etc.
Some of its applications include detecting groups based on common interests, occupation,etc and diffusing pattern of information by analyzing a particular community structure , etc. Besides, community detection can also be used in other types of networks such as biological networks, citation networks, etc [1].
On the other hand, link prediction is the problem of predicting missing or future links in a network.Link prediction mainly has two approaches: similarity based and learning based approaches.A similarity-based approach computes the similarities on non-connected pairs of nodes like common neighbor, jaccard index,etc. Every potential node pair would be assigned a score, where higher score means higher probability of having or forming link, and vice versa.
A learning-based approach treats the link prediction problem as a binary classification task.Certain classifiers and probabilistic model can be used for solving this problem. Each non-connected pair of nodes corresponds to an instance with features describing nodes and the class label. If there is a potential link connecting a pair of nodes, this pair is labeled as positive(1), otherwise it is negative(0). The features can be either the similarity metrics or features derived from the network, such as the textual information of attributes and domain knowledge, or both combined.Some of its applications include recommendation of friends,collaborators, inference of complete networks based on partial data, predicting interaction of biological elements like proteins ,etc [2].

A co-author network is a social network where actors are authors and relationships are collaborations(co-authorships) eg.DBLP co-author network.There are two types of network considered : Homogeneous and heterogeneous model.In a homogeneous model , the network consists of only one author nodes. The features that can be extracted in this type of network are mainly structural. In heterogeneous model , different elements such as papers, keywords , etc. are considered as nodes along with author nodes.The set of paths connecting different layers are called as meta−paths.Meta-path based features are used for predicting links.For example , a relation between two authors are formed through their paper published in same journal or their citing of each others' paper ,etc.So, the number of such relation is a feature for ranking or training classifiers[5]. Mining Co-author networks can reveal significant information like how authors collaborate, what make them form communities and what are the features or criteria that predict future collaborations.Similarity features proposed in literature are mainly structural: node-based and path-based.Certain works also focuses on domain-specific and meta-path based features.But due to their complexity, majority of works focus on structure-based features.

This work uses a semantic clustering model [12] to group keywords and derive a novel similarity measure based on the authors existing publication titles for prediction of co-authorships.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

The semantic clustering model is a combination of a word embedding model (Word2Vec) [13] and clustering model (KMeans). Also, a supervised learning model is used to create a predictor from the proposed measure.

II. RELATED WORKS

Previous studies have dealt extensively with different approaches to co-author prediction.

Pavlov et.al [3] proposed a model of predicting collaborations in a homogeneous co-authorship network. They used topological metrics such as common neighbors , Jaccard Index,Katz index ,etc as features to train a classifier for learning a model.They used a variety of classifiers namely , SVM ,Decision tree , AdaBoost,etc.

Maruyama et.al [4] used a homogeneous model and used domain-specific features and topological features to predict co-authorship.The domain-specific features are no. of common journals,conference papers , whether the two author-pairs have undergone common graduate program ,etc.The topological features are Common neighbors , Katz index ,etc.They also used a horizontal filter that eliminates the author-pairs who have certain domain-specific attributes zero. A number of classifiers were used such as Regression ,Naive bayes , Logit boost, BFtree,etc.

Sun et.al [5] proposed a heterogeneous model of bibliographic network and used meta-paths between authors as a feature.They combined meta-paths like author-paper-author,author-venue-author,etc. and relation measures like relation count,random walk,etc. to obtain meta-features.The used logistic regression to learn a model for predicting co-authorships using meta-features as similarity features.

Zhang [6] studied meta-features and their combination as mechanisms of co-authorship evolution.They first created a prediction model with meta-features using logistic regression and learned weights to combine these features for improved result.

Yu et.al [7] proposed a path and node combined approach to rank links in social networks.Results show that path and node combined approach gives better results of prediction when compared with node-based or path-based approaches alone.

Ding et.al [8] proposes a novel community detection algorithm based on local information.They also propose to new community relevance indices to calculate relevance between each pair of communities.Finally, they propose a ruler inference based model to predict links.Their results show that their proposed method has better result and lower time complexity compared with other state-of-the-art approaches.

Bhardwaj et.al [9] studied the link prediction problem in DBLP co-authorship network considering the role of communities.Their result shows that links are not random and that structure-based predictors such as CN,JI,AA,PA outperform random predictors.Also, the role of communities play an important role in prediction result.

Soundaraj et.al [10] proposed community enhanced metrics based node similarities.Their results show that adding community information can improve the score of precision.

Deylami et.al [11] proposed different set of community based metrics that can be calculated in a more reasonable time compared to the metrics of [10].

III. PROPOSED PREDICTION MODEL

The overall framework of the proposed model is given in fig.1.

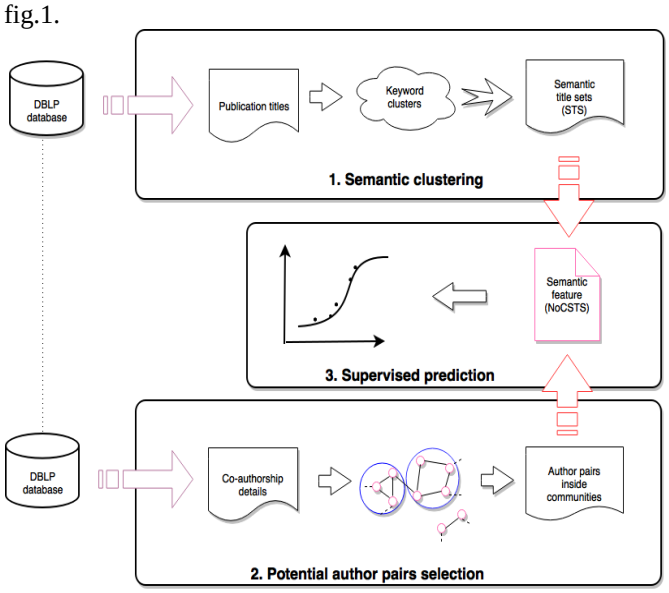

Fig. 1. Proposed model framework.

The proposed model comprises of main 3 modules.

- Semantic clustering.
- Potential author pairs selection.
- Supervised prediction.

A. *Semantic Clustering*

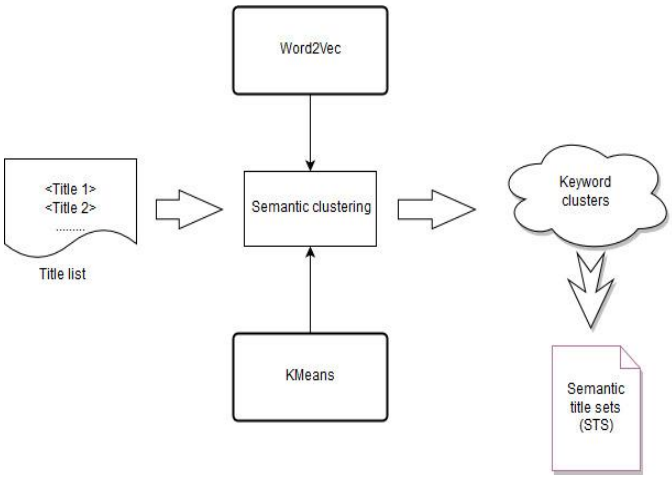

Fig. 2. Semantic clustering.

This module creates unique semantic title sets (STS) from the input publication titles using the combination of a word embedding architecture(Word2Vec) and a clustering algorithm (Kmeans).This module is further subdivided into 3 steps:

- Preprocessing of publication titles.
- Keyword clustering based on Word2Vec and Kmeans algorithm.
- Creation of Semantic Title Sets (STS).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

1) *Preprocessing*
The publication titles are stopped, stemmed and represented as a sequence of keywords.

2) *Keyword clustering*
The stemmed titles are fed to Word2Vec model.The Word2Vec model takes certain keywords from the input and maps them to a low-dimensional continuous vector space. The keywords that co-occur in the input corpus are positioned close in the vector space.The output word vectors are clustered using KMeans algorithm such that those words that are semantically similar are put in the same cluster.

3) *Semantic Title Sets(STS) creation*
After getting the keyword clusters,each title is represented as a sequence of cluster membership of its keywords (boolean vector). Each title is represented as a combination of keyword clusters.

If $title_i = k_{i1}k_{i2}...k_{it}$, $t = |K(title_i)|$ where $K(title_i)$ is set of all keywords of $title_i$ and if there are C keyword clusters, then $title_i = M(c_1)M(c_2)....M(c_l)$, for $l = |C|$ and

- $M(c_m) = 1$, if $k_{ij} \in c_m$ for any $k_{ij} \in k(title_i)$ and $c_m \in C$
- $M(c_m) = 0$, otherwise.

For creating STS, a parameter n is used that controls how much similar the titles in a STS are.The steps are:
Take each unique title representations (boolean vector).
For each unique title vector, group all the other title vector with that share atleast n clusters with it. This gives STS_n.
After creating STS, remove any duplicate STS if there are any.

B. Potential Author Pairs Selection

In this module, the co-authorship of authors are used to get potential author pairs who can collaborate. There are 3 steps.

- Co-author network creation.
- Community detection.
- Selecting author pairs inside communities (potential).

1) *Co-author network creation*
The co-authorships for the corresponding publication titles used for module 1 are extracted and a corresponding co-author network is formed.A co-author network defines a collaboration between two authors.

2) *Community detection*
To get potential author pairs, it is enough to consider author pairs in some dense communities.This work uses the LPA[14] for community detection.After getting communities, a threshold community size s is used to get significant communities.

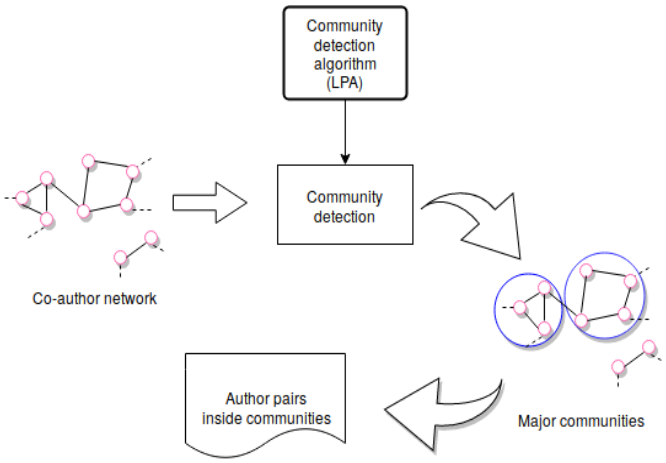

Fig. 3. Potential author pairs selection.

3) *Selection of potential author pairs*
Once the major communities are extracted, the author pairs inside them are chosen as potential.Equal number of both existing and nonexisting author collaborations are chosen.

C. Supervised Prediction

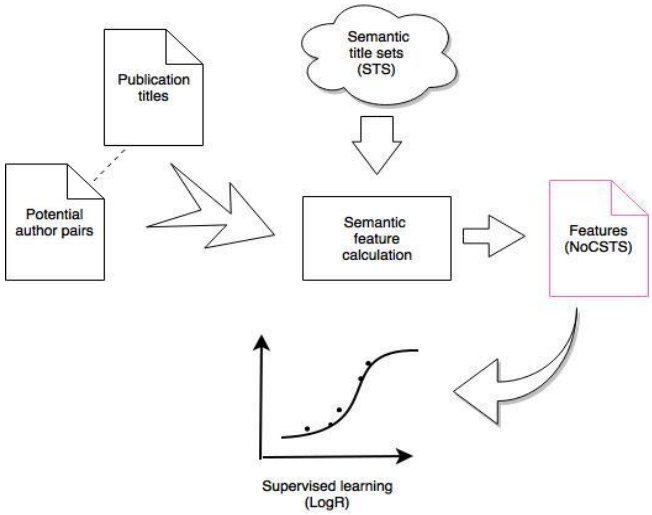

Fig. 4. Supervised prediction.

This module calculates the proposed semantic similarity measure (NoCSTS) and creates its corresponding link predictor using supervised learning.There are 2 steps.

- Semantic feature (NoCSTS) calculation.
- Supervised prediction model creation.

1) *NoCSTS calculation*
NoCSTS is defined as Number of common STS.For each author pair (A_i, A_j), it calculates the number of times their unique combination of publication titles occur in each STS.

$$\text{NoCSTS}(A_i, A_j) = \sum_{\substack{p \in P(A_i), \\ p' \in P(A_j)}} \sum_{S \in STS_n} S(p, p'), \text{ where}$$

- $S(p,p') = 1$ if $p \in S$ and $p' \in S$,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

- *S(p,p') = 0, otherwise.*

$P(A_i)$ is the set of all publication titles of A_i .

$P(A_j)$ is the set of all publication titles of A_j .

STS_n is the total semantic titlesets , each sharing atleast *n* same clusters.

2) NoCSTS calculation

After getting NoCSTS feature for each potential author pairs, logistic regression is used to create NoCSTS predictor.For that, the potential author pairs are divided into test and train pairs.A supervised model(predictor) is created for the train pairs and using that predictor, the test pairs are predicted.

NoCSTS predictor can be modeled as [5] :

$$\Pr(A_i, A_j) = \frac{e^X}{1+e^X}$$, where

X is a linear combination of weight and NoCSTS feature.

IV. COMPARISON PREDICTORS

This work uses 5 other keyword-based measures to compare with the proposed measure.

For that, first each measure is fed to a binary classifier to learn a model.This model is called a link predictor as it outputs probability of links.

These 5 comparison measures are created using a metapath and graph-based measures[5,6].

The metapath is APKPA based on a heterogeneous network.APKPA is the path connecting two authors through their publication keywords.Or, in other words, APKPA is a connection between two authors who share same publication keywords.

The measures are Relation count (RC), Normalized relation count (NRC), Random walk(RW) and Symmetric random walk(SRW).

A. APKPA-RC

This counts all the APKPA relations.

APKPA-RC (A_i, A_j) is the total number of APKPA paths between A_i and A_j.

B. APKPA-NRC

It normalizes RC from both sides.

$$APKPA\text{-}NRC(A_i, A_j) = \frac{APKPA-RC(A_i, A_j) + APKPA-RC(A_j, A_i)}{APKPA-RC(A_i, .\ ) + APKPA-RC(., A_j)}$$

C. APKPA-RW

It counts how a random walker from A_i to A_j.

$$\text{APKPA-RW}(A_i, A_j) = \frac{APKPA-RC(A_i, A_j)}{APKPA-RC(A_i, .)}$$

D. APKPA-SRW

It counts RW from both sides.

$$APKPA\text{-}SRW(A_i, A_j) = APKPA\text{-}RW(A_i, A_j) + APKPA\text{-}RW(A_j, A_i).$$

E. C-APKPA

It combines the above four predictors using Logistic regression (based on [6]).

$$C-APKPA(A_i, A_j) \sim \{APKPA-RC(A_i, A_j) + APKPA-NRC(A_i, A_j) + APKPA-RW(A_i, A_j) + APKPA-SRW(A_i, A_j)\}$$

$(A_i, .)$ or $(., A_i)$ represents all APKPA paths between A_i and all other authors.

$(A_j, .)$ or $(., A_j)$ represents all APKPA paths between A_j and all other authors.

The first four predictors are created using each keyword-based metrics and the fifth predictor is a combination of these four. Also, NoCSTS as a feature is used to create a link predictor.Then NoCSTS is compared with these 5 predictors using below evaluation metrics.

V. EXPERMENTS AND RESULTS

The proposed work is implemented in a dual-core system with 4 GB RAM.The Ubuntu OS is used.

A. Dataset

For the purpose of evaluating our proposed model, we use DBLP bibliographic database [15].A part of the DBLP dataset dump file is used for this research work : publication titles and co-authorships corresponding to 'data mining' and 'data eng.' journals that corresponds to 3681 publications, 6901 authors and 13789 co-authorships.

B. Tools

We use python and RStudio for implementing and evaluating our work.Main python packages used are *:nltk* for preprocessing, *gensim-word2vec* and *scikit-learn - Kmeans* for semantic clustering, and *igraph* for graph implementation and plotting.We use *glm()* in RStudio to implement the logistic regression model and use *ROCR* package for finding AUC.

C. Evaluation metrics

The performance of the proposed model is analysed by using five evaluation measures. They are accuracy,precision,recall,F-score and AUC .Dataset is divided into train data and test data .The system is trained with 90% of dataset and tested with remaining 10%.Since the model uses two random approaches LPA and KMeans, *10-fold* validation cant be used as each time the train and test author pairs will be different.The proposed model is run for ten times , each time 90% of train and 10% of test author pairs are selected randomly.Each time the comparison predictors namely the keyword-based metrics (APKPAs) are implemented and compared with the proposed

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

model.Average of the above mentioned five evaluation metrics is taken and recorded in a result table.

A binary classifier predicts all data instances of a test dataset as either positive(link exist) or negative(link not exist) . This classification (or prediction) produces four outcomes as given below[16].

- True positive (TP): correct positive prediction.
- False positive (FP): incorrect positive prediction.
- True negative (TN): correct negative prediction.
- False negative (FN): incorrect negative prediction.

We use the following evaluation measures to evaluate our proposed predictor.

1) Accuracy

Classification accuracy is the number of correct predictions made as a ratio of all predictions made[16].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

2) Precision

It is the ratio of true positive links to the total true predicted links[16].

$$Precision = \frac{TP}{TP + FP}$$

3) Recall

It is the ratio of true positive links to the total true links [16].

$$Recall = \frac{TP}{TP + FN}$$

4) F-score

It is the harmonic mean of precision and recall [16].

$$F\text{-}score = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

5) AUC

AUC(or AUROC) is Area under Receiver Operating Characteristics(ROC) Curve.ROC illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied [17].ROC is plotted between True Positive Rate (TPR) and False Positive Rate (FPR) for different thresholds of prediction.TPR is Recall and FPR is ratio of FP links to total actually false links (FP+TN).For each threshold from (0,1) in the probability score, find the TPR and FPR score and plot it in the TPR-FPR axis.

The area under this plot obtained (ROC curve) is AUC (value ranging from 0 to 1.) .

D. Result

The proposed and comparison predictors are trained and tested. The train and test author pairs were divided in the below given ratios.

- 90 % train and 10 % test author pairs.
- 80 % train and 20 % test author pairs.
- 70 % train and 30 % test author pairs
- 60 % train and 40 % test author pairs.
- 50 % train and 50 % test author pairs.

Each of the above specified sets were run for 10 times.

The average results for NoCSTS (proposed) and other 5 existing predictors : APKPA-RC, APKPA-NRC, APKPA-RW, APKPA-SRW, C-APKPA are shown below.

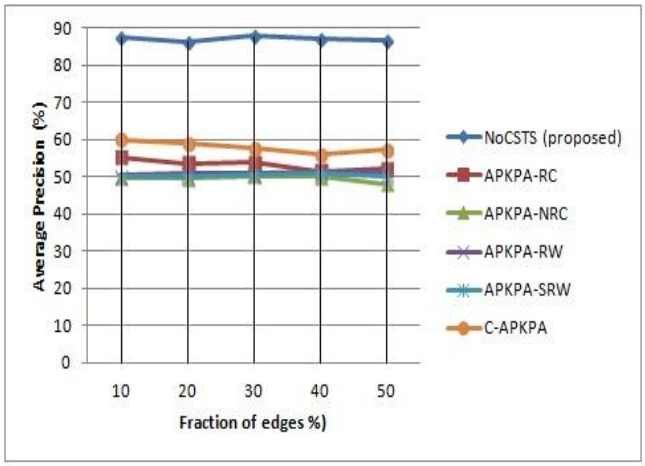

Fig. 5. Average Precision line graph for different fraction of edges. NoCSTS performs well compared to other predictors.

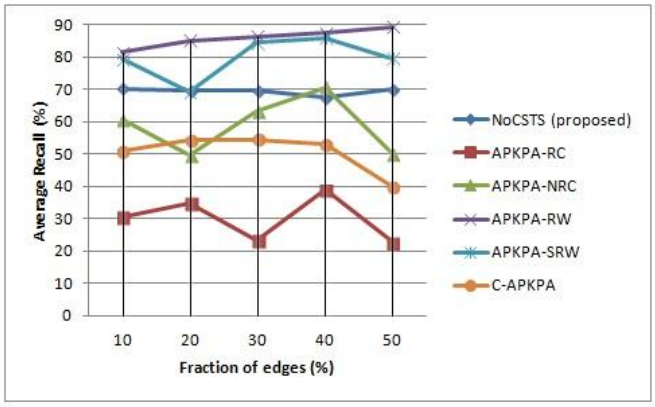

Fig. 6. Average Recall line graph for different fraction of edges. APKPA-RW has highest value, then APKPA-SRW and third is NoCSTS.

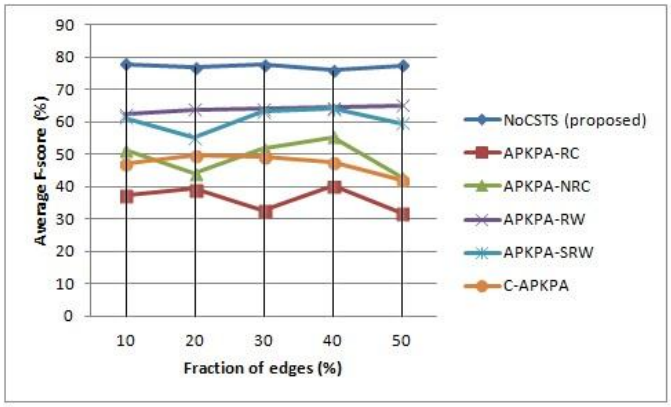

Fig. 7. Average F-score line graph for different fraction of edges. NoCSTS performs well compared to other predictors.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

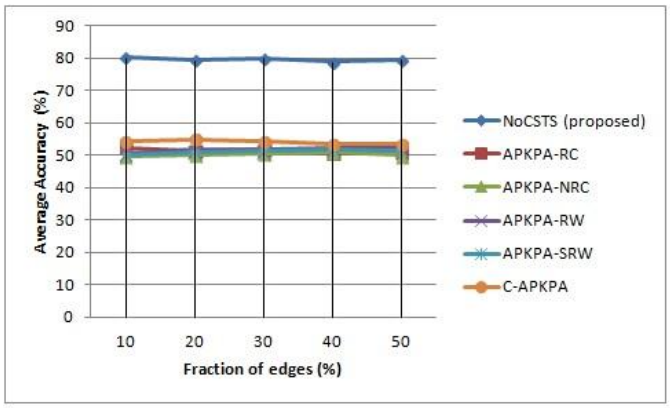

Fig. 8. Average Accuracy line graph for different fraction of edges. NoCSTS performs well compared to other predictors.

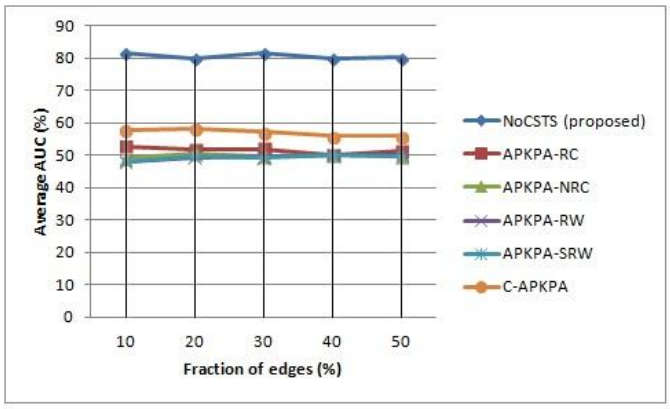

Fig. 9. Average AUC line graph for different fraction of edges. NoCSTS performs well compared to other predictors.

As the results show, the proposed predictor performs well with other predictors except for Recall.

VI. CONCLUSION

Predicting co-authors is an important mining task in social network analysis.It can help determine how a network grows, what are the topics that can converge or diverge and how authors form communities.

In this paper, a co-author prediction model based on semantic clustering and supervised prediction is proposed. By combining semantics of publication titles with supervised learning, significant co-authorships can be predicted. Our model shows overall improved results with other keyword-based metrics after evaluation.

As a future work, additional features such as h-index, etc can be added. The use of other metrics with the proposed measure can be used to enhance prediction. Also, combined classifiers can be implemented with regularization.

REFERENCES

[1] https://en.wikipedia.org/wiki/Community_structure.
[2] Wang, Peng, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. "Link prediction in social networks: the state-of-the-art." Science China Information Sciences 58.1 (2015): 1-38.
[3] Pavlov, Milen, and Ryutaro Ichise. "Finding experts by link prediction in co-authorship networks." Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics-Volume 290. CEUR-WS. org, 2007.
[4] Maruyama, W., and L. Digiampietri. "Co-authorship prediction in academic social network." ,BraSNAM-5 Brazilian Workshop on Social Network Analysis and Mining(2016).
[5] Sun, Yizhou, Rick Barber, Manish Gupta, Charu C. Aggarwal, and Jiawei Han. "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, 2011.
[6] Zhang, Jinzhu. "Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction." Information Processing & Management (2016).
[7] Yu, Chuanming, Xiaoli Zhao, Lu An, and Xia Lin. "Similarity-based link prediction in social networks: A path and node combined approach." Journal of Information Science (2016): 0165551516664039.
[8] Ding, Jingyi, Licheng Jiao, Jianshe Wu, and Fang Liu. "Prediction of missing links based on community relevance and ruler inference." Knowledge-Based Systems 98 (2016): 200-215.
[9] Bhardwaj, Onkar, and Xiaohui Lu. "Experiments with Link Prediction on DBLP coauthorship network."
[10] Soundarajan, Sucheta, and John Hopcroft. "Using community information to improve the precision of link prediction methods." In *Proceedings of the 21st International Conference on World Wide Web*, pp. 607-608. ACM, 2012.
[11] Deylami, Hasti Akbari, and Masoud Asadpour. "Link prediction in social networks using hierarchical community detection." In *Information and Knowledge Technology (IKT), 2015 7th Conference on*, pp. 1-5. IEEE, 2015.
[12] Kong, Xiangjie, Huizhen Jiang, Zhuo Yang, Zhenzhen Xu, Feng Xia, and Amr Tolba. "Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation." PloS one 11, no. 2 (2016): e0148492.
[13] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
[14] Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks." *Physical review E* 76, no. 3 (2007): 036106.
[15] http://dblp.uni-trier.de.
[16] https://en.wikipedia.org/wiki/Precision_and_recall
[17] https://en.wikipedia.org/wiki/Receiver_operating_characteristic