# Important Citations Identification with Semi-supervised Classification Model

Xin An[1][0000-0001-7413-9396], Xin Sun[1] and Shuo Xu[2*][0000-0002-8602-1819]

[1] School of Economics & Management, Beijing Forestry University, Beijing 100083, P.R. China
anxin@bjfu.edu.cn (Xin An), sx0118@outlook.com (Xin Sun)
[2] College of Economics and Management, Beijing University of Technology, Beijing 100124, P.R. China
xushuo@bjut.edu.cn (Shuo Xu)
* Corresponding author

**Abstract.** Given that citations are not equally important, various techniques have been presented to identify important citations on the basis of supervised machine learning models. However, only a small volume of data has been annotated manually with the labels. To make full use of unlabeled data and promote the learning performance, the semi-supervised self-training technique is utilized to identify important citations in this work. After six groups of features are engineered, the semi-supervised versions of SVM and RF models improve significantly the performance of the conventional supervised versions when un-annotated samples under 75% and 95% confidence level are rejoined to the training set, respectively. The AUC-PR and AUC-ROC of SVM model are 0.8102 and 0.9622, and those of RF model reach 0.9248 and 0.9841, which outperform their counterparts. This demonstrates the effectiveness of our semi-supervised self-training strategy for important citation identification.

**Keywords:** Important Citation, Semi-supervised Learning, Self-training.

## 1 Introduction

Citations are reckoned as a proxy of scientific knowledge flow in the literature, thus they are usually utilized for multifarious academic evaluation purposes, such as ranking of researchers [1], journals [2], organizations [3], etc. But most studies treat all references as equally important to an interested citing publication. This is obviously not in line with actual situations. In recent years, researchers have argued that citations are not equally important and presented various techniques to identify important citations [4-11].

The supervised learning methods are commonly used for this task, which learn the feature space of the labeled data to form a classification model. However, most supervised learning methods require a large amount of labeled data to ensure the performance of the resulting machines [12]. Currently, only a small number of citations are labeled

manually due to the time-consuming annotation and heavy workload. That is to say, large amounts of unlabeled data have not been exploited. Last two decades have witnessed significant progress in the field of semi-supervised learning, and many successful cases from various fields are reported in the literature [12-15]. However, important citations identification with semi-supervised model remains largely under-studied.

To make full use of unlabeled data and promote the model performance, a semi-supervised self-training method is deployed in this work. After Section 2 briefly describes the related work, the framework of semi-supervised self-training for important citation identification is introduced in Section 3 along with six groups of features [11]. Section 4 shows the statistics of labeled and unlabeled data. In Section 5, the experiments of SVM and RF models armed with semi-supervised self-training strategy are conducted, and Section 6 concludes this work.

## 2 Related work

In the literature, various techniques have been presented to identify important citations. Valenzuela et al. [4] annotated 465 citations from ACL anthology and used two supervised learning models (SVM and RF) to conduct important citations classification. Since then, a plethora of studies have been implemented with different super-vised learning models on this annotated dataset [6-11], including SVM, RF, Naïve Bayes, K-Nearest Neighbors, Decision Tree, Deep Learning, etc. Among all these supervised models, SVM and RF were the most commonly used and outperformed the other counterparts. It can be seen that the supervised learning model is a main-stream technique in this task. However, it relies on large amount of labeled data to maintain the performance, which is in contrast with the reality that labeled data costly to obtain.

In practice, to overcome the limitation of little amount of labeled data and make full use of unlabeled data, the semi-supervised learning algorithm have received more attention. Many semi-supervised learning methods are raised, such as co-training [13], semi-supervised support vector machine (S3VM) [14], self-training [15], etc. These methods have been indicated the effectiveness in improving the predictive performance when leveraging large amounts of unlabeled data with a small amount of labeled data.

Among these approaches, the self-training method expands the training data with predictions on unlabeled data. It is easy to conduct and has great flexibility in threshold setting, which gives more choices on model selection. Therefore, to make full use of the unlabeled data, the semi-supervised self-training method is preferred to identify important citations in this paper.

## 3 Methodology

Figure 1 depicts the framework of important citations identification on the basis of semi-supervised self-training learning strategy. First of all, a supervised learning model (such as SVM and RF) is trained on the labeled data under 5-fold cross validation. After

learning the training set of each fold, the labels of the unlabeled data are predicted respectively. We selected samples with 95%, 90%, 85%, 80%, 75%, and 70% confidence level as the pseudo-labeled data to rejoin the training set. For each fold, the model is retrained on the new combined data and evaluated on the testing set. The involved parameters are optimized correspondingly. The areas under the curve of PR and ROC are used as indicators for evaluating the performance.

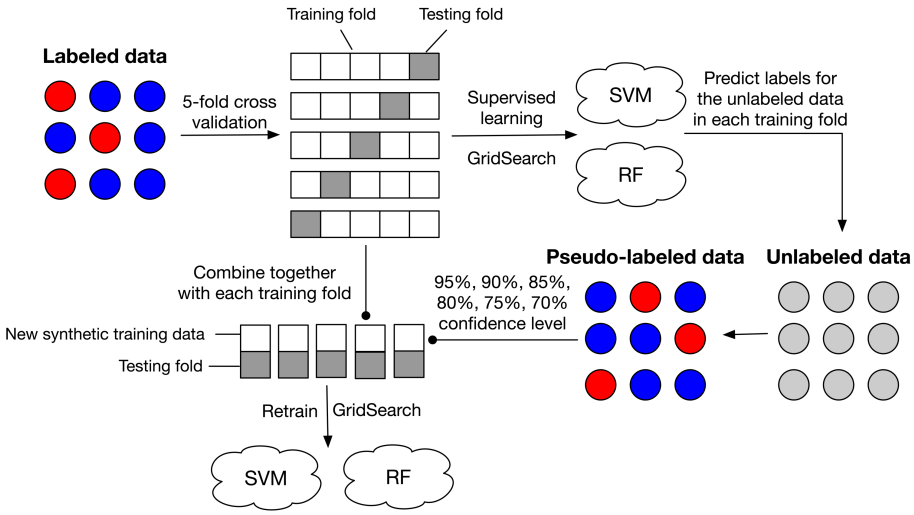



**Fig. 1.** Framework for identifying important citations with the semi-supervised learning model.

As for the feature engineering, the following six groups of features from our previous study [11] are utilized here: G1 (two generative features extracted from the CIM model), G2 (Structural based features, containing 7 features), G3 (Separate citation-based feature, containing 1 feature), G4 (Author overlap-based feature, containing 1 feature), G5 (Cue words-based feature, containing 2 feature), G6 (Similarity based feature, containing 1 feature). Please refer to [11] for more details.

## 4 Data and preprocessing

The annotated corpus in [4] is used in this work. This dataset was randomly chosen from the ACL anthology and were manually annotated by one expert with the label 0 (related work), 1 (comparison), 2 (using the work), and 3 (extending the work). For conducting the experiment of identifying important citations, we combine the related work and comparison classes into incidental class with the label 0, and using the work and extending the work classes into important class with the label 1. The inner-annotator agreement was verified between two experts to reduce the bias raised by human annotation and reached 93.9% in this coarse label set. Table 1 lists the summary of the labeled dataset. In the end, 456 pairs of labeled data were collected after preprocessing, of which 14.7% are important citations.

**Table 1.** Summary of labeled dataset.

| Label | Class | Number of Samples |
|---|---|---|
| 0 | Incidental | 389 (85.3%) |
| 1 | Important | 67 (14.7%) |

The preprocessing steps include: (1) Collecting PDF format of citing papers and converting to text format by Xpdf; (2) Parsing the text format data by ParsCit to extract title, author, abstract and references of each citing paper as well as the generic section headers; (3) Extracting citation contexts based on regular expressions; (4) Preprocessing all textual information including citation contexts and abstract using NLTK toolkit. During the preprocessing, 434 citing papers are collected, which yields 8,541 citing and cited pairs totally. Table 2 lists the statistics of citing paper and references. Apart from the labeled data described above, 8,085 unlabeled citations come into being. Similar to the labeled data, the feature engineering and preprocessing are also conducted on all unlabeled data.

**Table 2.** Statistics of citing paper and references.

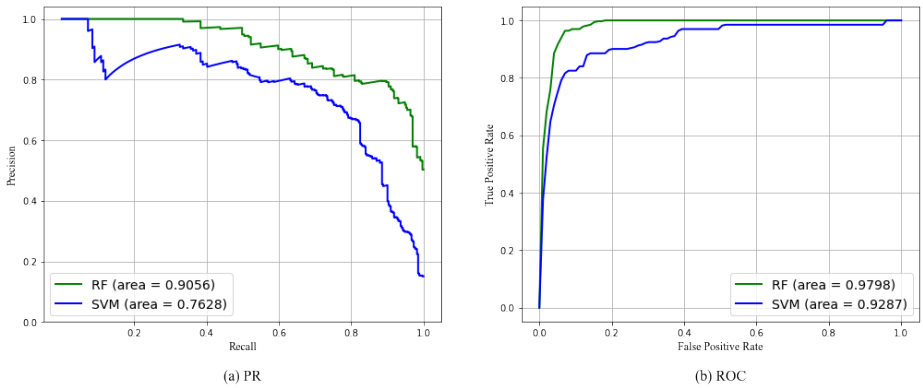
| Number of Citing papers | Number of unique references | Number of total citing and cited pairs | Number of unlabeled data |
|---|---|---|---|
| 434 | 4,590 | 8,541 | 8,085 |

## 5 Experimental results and discussion

As two state-of-the-art discriminative models, SVM and RF are utilized here as our classifiers. First of all, these two models were trained on the labeled data. To tune the parameters of these two classifiers, grid search with 5-fold cross-validation [16] is used in this study. Figure 2 shows the PR curves and ROC curves of SVM and RF. As one can see, the area under the ROC curve (AUC-ROC) of SVM and RF models are 0.9287 and 0.9798 respectively, and the areas under the PR curve (AUC-PR) are 0.7628 and 0.9056 respectively. The RF model outperforms the SVM model, which is in accordance with most of previous studies [4-11].

Then, a semi-supervised self-training on the unlabeled data is conducted. After learning the training set of each fold based on the above 5-fold data, the labels of the unlabeled data are predicted. We select samples with 95%, 90%, 85%, 80%, 75%, and 70% confidence level to rejoin the training set. Table 3 lists the number of new samples of each fold at different confidence level. After that, for each fold, the resulting model is retrained on new combined data and evaluated on the testing set. Similarly, grid search is also used to tune the involved parameters. Table 4 reports the results of mean AUC-ROC and AUC-PR of 5-fold under different confidence level. It can be seen that the AUC-PR and AUC-ROC for SVM model reach the maximum at the 75% confidence level, which are 0.8102 and 0.9622 respectively. The RF model has the highest AUC-

PR and AUC-ROC at 95% confidence level (0.9248 and 0.9841). Both are better than the results of the above supervised learning counterparts.




(a) PR (b) ROC

**Fig. 2.** The PR curves (a) and ROC curves (b) of SVM and RF models on labeled data with supervised learning strategy.

**Table 3.** Number of new samples under different confidence levels.

| Fold | Model | Confidence Level | | | | | |
|---|---|---|---|---|---|---|---|
| | | 95% | 90% | 85% | 80% | 75% | 70% |
| 1 | SVM | 4,444 | 5,977 | 6,714 | 7,067 | 7,334 | 7,533 |
| | RF | 1,002 | 2,406 | 3,670 | 4,709 | 5,368 | 5,909 |
| 2 | SVM | 3,538 | 5,863 | 6,567 | 6,999 | 7,279 | 7,502 |
| | RF | 944 | 2,462 | 3,674 | 4,663 | 5,387 | 6,054 |
| 3 | SVM | 3,993 | 5,913 | 6,649 | 7,025 | 7,306 | 7,517 |
| | RF | 925 | 2,462 | 3,620 | 4,624 | 5,369 | 6,086 |
| 4 | SVM | 4,362 | 5,940 | 6,688 | 7,040 | 7,319 | 7,521 |
| | RF | 944 | 2,462 | 3,674 | 4,663 | 5,387 | 6,054 |
| 5 | SVM | 3,411 | 5,853 | 6,555 | 6,994 | 7,271 | 7,499 |
| | RF | 944 | 2,462 | 3,674 | 4,663 | 5,387 | 6,054 |

**Table 4.** Performance of SVM and RF models with semi-supervised strategy under different confidence levels.

| Confidence level | SVM | | RF | |
|---|---|---|---|---|
| | AUC-PR | AUC-ROC | AUC-PR | AUC-ROC |
| 95% | 0.7380 | 0.9217 | 0.9248 | 0.9841 |
| 90% | 0.7290 | 0.9078 | 0.9015 | 0.9804 |
| 85% | 0.7525 | 0.9225 | 0.8811 | 0.9759 |
| 80% | 0.7545 | 0.9248 | 0.8463 | 0.9702 |
| 75% | 0.8102 | 0.9622 | 0.8331 | 0.9674 |
| 70% | 0.7522 | 0.9292 | 0.8374 | 0.9666 |

Further, to find out the contribution of each group of features, we perform an additional experiment to observe the changes of mean AUC-PR and mean AUC-ROC. Table 5 shows the scores of mean AUC-PR and AUC-ROC of the SVM model under 75%

confidence level and the RF model under 95% confidence level and their rankings (in parentheses) as well as the average rank using different groups of features under 5-fold cross-validation by controlling for structure features (G2). For each com-bination, the resulting parameters are optimized separately. As we can observe, the baseline model based on the structural features achieves a mean AUCPR of about 0.7600 and 0.7903, and AUCROC of about 0.8906 and 0.4743. The author-overlap based features (G4) ranks first, which increase respectively the AUC-PR to 0.9462 and 0.8145, AUC-ROC to 0.8145 and 0.4798. The CIM (Citation Influence Model) [17] model-based features (G1) rank the second, which demonstrates that the features generated from the generative model can improve the performance of important citations identification. This observation is in accordance with the previous work [11].

**Table 5.** The performance of semi-supervised SVM and RF models with different groups of features in terms of mean AUC-PR, AUC-ROC, and their ranks.

| Feature | SVM | | RF | | Average_rank |
|---|---|---|---|---|---|
| | PR | ROC | PR | ROC | |
| G2 | 0.7600(3) | 0.8906(6) | 0.7903(5) | 0.4743(5) | 4.75 |
| G2+G1 | 0.7558(4) | 0.8935(5) | 0.9035(1) | 0.4968(1) | 2.75 |
| G2+G3 | 0.7448(5) | 0.8971(4) | 0.8183(2) | 0.4885(3) | 3.5 |
| G2+G4 | 0.9462(1) | 0.9875(1) | 0.8145(3) | 0.4798(4) | 2.25 |
| G2+G5 | 0.7822(2) | 0.9065(3) | 0.7065(6) | 0.4604(6) | 4.25 |
| G2+G6 | 0.6947(6) | 0.9181(2) | 0.7997(4) | 0.4889(2) | 3.5 |

## 6 Conclusion

In this paper, we refer to the practices in [4] to divide citations into important and incidental classes and use semi-supervised self-training strategy to identify important citations by leveraging labeled data and unlabeled data to promote the performance and generalization ability. Through the semi-supervised self-training on the unlabeled data, the performance of the SVM model can be promoted from 0.9287 to 0.9622 and from 0.7628 to 0.8102 and that of the RF model from 0.9798 to 0.9841 and from 0.9056 to 0.9248 in terms of mean AUC-ROC and mean AUC-PR. This demonstrates the effectiveness of our semi-supervised self-training strategy for important citation identification. Additionally, the CIM model-based features, structural based features and author-overlap based features contribute greatly on important citations identification.

### Acknowledgements

This research received the financial support from the National Natural Science Foundation of China under grant number 72004012 and 72074014.

### References

1. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences 102(46), 16569-16572 (2005).

2. Garfield, E.: Citation indexes to science: a new dimension in documentation through association of ideas. Science, 122:108-111 (1955).
3. Lazaridis, T.: Ranking university departments using the mean h-index. Scientometrics 82(2), 211-216 (2010).
4. Valenzuela, M., Ha, V., Etzioni, O.: Identifying meaningful citations. In: Workshops at the twenty-ninth AAAI conference on artificial intelligence, pp. 21-26. AAAI , Austin (2015).
5. Zhu, X., Turney, P., Lemire, D., Vellino, A.: Measuring academic influence: not all citations are equal. Journal of the Association for Information Science and Technology 66(2), 408-427(2015).
6. Hassan, S.U., Akram, A. and Haddawy, P. Identifying important citations using contextual information from full text. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1-8. IEEE, New York (2017).
7. Hassan, S.U., Safder, I., Akram, A., Kamiran, F.: A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. Scientometrics 116(2), 973-996 (2018).
8. Hassan, S.U., Imran, M., Iqbal, S., Aljohani, N.R., Nawaz, R.: Deep context of citations using machine-learning models in scholarly full-text articles. Scientometrics 117(3), 1645-1662 (2018).
9. Qayyum, F., Afzal, M.T.: Identification of important citations by exploiting research articles' metadata and cue-terms from content. Scientometrics 118(1), 21-43 (2019).
10. Wang, M., Zhang, J., Jiao, S., Zhang, X., Zhu, N., Chen, G.: Important citation identification by exploiting the syntactic and contextual information of citations. Scientometrics 125(3), 1-21 (2020).
11. An, X., Sun, X., Xu, S., Hao, L., Li, J.: Important Citations Identification by Exploiting Generative Model into Discriminative Model. Journal of Information Science. (2021) doi:10.1177/0165551521991034.
12. Xu, S., An, X., Qiao, X., Zhu, L., Li, L.: Semi-supervised least-squares support vector regression machines. Journal of Information & Computational 8(6), 885-892 (2011).
13. Blum, A., Mitchell, T.: In: Proceeding of the eleventh annual conference on Computational learning theory, pp.92-100. ACM, Madison, Wisconsin (1998).
14. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. Journal of Machine Learning Research 9(2), (2008).
15. Tanha, J., van Someren, M., Afsarmanesh, H.: Semi-supervised self-training for decision tree classifiers. Journal of Machine Learning and Cybernetics 8(1), 355-370 (2017).
16. Xu, S., Ma, F., Tao, L.: Learn from the Information Contained in the False Splice Sites as well as in the True Splice Sites using SVM. Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering, 1360-1366 (2007).
17. Xu, S., Hao, L., An, X., Yang, G., Wang, F.: Emerging Research Topics Detection with Multiple Machine Learning Models. Journal of Informetrics, 13(4), 100983 (2019).