



Important citations identification with semi-supervised classification model

Xin An¹ · Xin Sun² · Shuo Xu³

Received: 19 June 2021 / Accepted: 9 November 2021 / Published online: 20 January 2022
© Akadémiai Kiadó, Budapest, Hungary 2021

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 instances have been annotated manually with the labels. To make full use of unlabeled instances and promote the identification performance, the semi-supervised self-training technique is utilized here to identify important citations in this work. After six groups of features are engineered, the SVM and RF models are chosen as the base classifiers for self-training strategy. Then two experiments based on two different types of datasets are conducted. The experiment on the expert-labeled dataset from one single discipline shows that the semi-supervised versions of SVM and RF models significantly improve the performance of the conventional supervised versions when unannotated 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 and the benchmark methods in the literature. This demonstrates the effectiveness of our semi-supervised self-training strategy for important citation identification. Another experiment on the author-labeled dataset from multiple disciplines, semi-supervised learning models can perform better than their supervised learning counterparts in term of AUC-PR when the ratio of labeled instances is less than 20%. Compared to our first experiment, insufficient amount of instances from each discipline in our second experiment enables the performance of the models to be unsatisfactory.

Keywords Important citation · Semi-supervised learning · Self-training · Expert-labeled dataset · Author-labeled dataset

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 (Hirsch, 2005), journals (Garfield, 2006), organizations (Lazaridis, 2010), etc.

✉ Xin An
anxin@bjfu.edu.cn

Extended author information available on the last page of the article

But most studies treat all references as equally important to an interested citing publication. This is obviously not in line with actual situations. Therefore, important citations identification plays a vital role in scientific evaluation procedure. It has promising potentials in fair distribution of academic resources and fair evaluation of talents. In recent years, researchers have argued that citations are not equally important and presented various techniques to identify important citations (An et al., 2021a; Hassan et al., 2017, 2018a; 2018b; Qayyum & Afzal, 2019; Valenzuela et al., 2015; Wang et al., 2020; Zhu et al., 2015).

Given a scholarly article, its important citations are actually the references that greatly contribute to this article. It is easy to see that the citation importance is closely related with citation function (viz. the reason for citing a paper) (Teufel et al., 2006; Valenzuela et al., 2015). Though various classification schemes for citation function were constructed in the literature (Abu-Jbara et al., 2013; Dong & Schafer, 2011; Li et al., 2013; Radoulov, 2008; Teufel et al., 2006), these schemes were greatly simplified after 2015 to facilitate annotation and machine-learning model building with satisfactory performance (An et al., 2021a). For example, Zhu et al. (2015) distinguished influential references from incidental ones from the role which a reference plays in the core idea, method of a given citing paper. Valenzuela et al. (2015) classified citations into *related work*, *comparison*, *using the work* and *extending the work*, and then further folded these categories into two ones: important citations (*related work* and *comparison*) and incidental ones (*using the work* and *extending the work*). Many follow-up studies followed these simplified classification schemes (An et al., 2021a; Hassan et al., 2017, 2018a; 2018b, Qayyum & Afzal, 2019; Wang et al., 2020).

To identify important citations, the supervised learning methods are commonly used, which can learn a discriminant pattern from a labeled dataset to form a classification model. However, most supervised learning methods require a large amount of labeled instances to ensure the performance of the resulting models (Xu et al., 2011). To the best of our knowledge, only annotated datasets in Valenzuela et al. (2015) and Zhu et al. (2015) can be accessed publicly due to the time-consuming annotation and heavy workload. The number of pairs of citing-cited articles is 456 and 2,685 in these two datasets respectively (cf. Section *Datasets*). Consequently, the overwhelming majority of classification models for identifying important citations are built on the basis of hundreds of labeled instances in previous studies (Hassan et al., 2017, 2018a; 2018b; Qayyum & Afzal, 2019; Valenzuela et al., 2015; Wang et al., 2020).

As a matter of fact, large amount of relatively inexpensive unlabeled instances can be available, but have not been exploited for identifying important citations. One branch of machine learning techniques, semi-supervised learning, is able to leverage large amount of un-annotated instances along with small amount of annotated instances. Last two decades have witnessed significant progress in the field of semi-supervised learning. Correspondingly, many learning strategies and methods have been proposed in the literature, such as self-training (Yarowsky, 1995), co-training (Blum & Mitchell, 1998), transductive support vector machine (TSVM) (Bennett & Demiriz, 1999; Joachims, 1999), and graph-based method (Zhu et al., 2005). Among these approaches, the self-training strategy provides more choices on base classifiers and has great flexibility in threshold setting.

However, important citations identification with semi-supervised model remains largely under-studied. To make full use of unlabeled instances and promote the model performance, a semi-supervised self-training learning strategy is deployed here to identify important citations. The SVM and Random Forest (RF) model are taken here as base classifiers of self-training learning strategy. In this study, we devote to exploiting whether and to what extent the unlabeled instances can benefit a supervised model. Besides, from the

perspective of practical significance, we hope the proposed strategy for important citations identification in this research can contribute fair evaluation of scientific research and academic achievements.

The rest of the article is structured as follows. After Section [Related work](#) briefly describes the important citations identification and semi-supervised learning, the framework of semi-supervised self-training for important citations identification is introduced in Section [Methodology](#). Section [Datasets](#) shows the statistics of two different types of datasets from Valenzuela et al. (2015) and Zhu et al. (2015). In Section [Experimental results and discussion](#), two experiments of SVM and RF models armed with semi-supervised self-training strategy are conducted, and Section [Conclusions](#) concludes this work.

Related work

Important citations identification

In recent years, the classification of citations has shifted from manual classification (Garfield, 1965) into automatic identification, from multi-categories (Abu-Jbara et al., 2013; Dong & Schafer, 2011; Li et al., 2013; Radoulov, 2008; Teufel et al., 2006) into only two categories (important vs. incidental) (Iqbal et al., 2021). Various approaches have been developed in the literature to identify important citations automatically.

Zhu et al. (2015) collected about 100 scholarly articles from 40 researchers with their opinions on the most essential references to their works, which generated 3,143 labeled pairs of citing-cited papers. Then, they used the SVM model as their supervised learning algorithm to classify the citations into influential category and non-influential one. Valenzuela et al. (2015) annotated 465 citations from the Association for Computational Linguistics (ACL) anthology into important and incidental categories and two supervised learning models (SVM and RF) were used to classify important citations. Since then, a plethora of studies have been implemented with different supervised learning models on these annotated datasets.

Hassan et al. (2017) employed five classification techniques (SVM, RF, Naïve Bayes, K-Nearest Neighbors and Decision Tree) on the dataset in Valenzuela et al. (2015) with 14 features, including context-based features, cue words-based features and textual features. Hassan et al. (2017) found that the RF model performed the best in terms of the area under the curve of receiver operating characteristic (AUC-ROC) and precision-recall (AUC-PR), followed by the SVM model. Hassan et al. (2018b) further exploited the potential of a deep learning model, Long Short-Term Memory (LSTM) model, for this task on the same dataset. Hassan et al. (2018b) observed that the LSTM model outperformed the traditional counterparts, but its performance is limited by the unavailable large-scale annotated instances.

Compared to Valenzuela et al. (2015), Qayyum and Afzal (2019) improved the performance of the SVM and RF models in term of precision by relying on freely available meta-data on the dataset in Valenzuela et al. (2015) and their self-collected dataset with 488 labeled citation pairs. Wang et al. (2020) distinguished important and non-important citations by engineering the syntactic and contextual features on the dataset in Valenzuela et al. (2015) and their self-annotated dataset with 458 citation pairs. Zeng et al. (2020) detected citation worthiness by using a Bidirectional Long Short-Term Memory (Bi-LSTM) network with attention mechanism and contextual information. An et al. (2021a) combined

generative model and discriminative model for identifying important citations on the datasets in Valenzuela et al. (2015) and Zhu et al. (2015). An et al. (2021a) found that the RF model outperforms the SVM model, but the Convolutional Neural Network (CNN) model did not achieve the desired performance due to the small volume of annotated instances. Aljuaid et al. (2021) improved the performance by using sentiment analysis of in-text citations to identify important citations with the SVM, Kernel Logistic Regression (KLR) and RF models on the dataset in Valenzuela et al. (2015) and the dataset in Qayyum and Afzal (2019).

It can be seen that the supervised learning model is a main-stream technique in this task. Among all these supervised models, the SVM and RF models were the most commonly used and outperformed the other counterparts. However, the supervised learning technique heavily relies on large amount of labeled instances to maintain the performance, which is in contrast with the reality that labeled instances are costly to obtain. In fact, only two small-scale labeled datasets in Valenzuela et al. (2015) and Zhu et al. (2015) are publicly available. Until now, large amount of unlabeled instances have not still been exploited.

Semi-supervised learning

In practice, to overcome the limitation of little amount of labeled instances and make full use of unlabeled instances, the semi-supervised learning technique has been receiving more attention. It attempts to harness the unlabeled instances to exceed the performance of supervised learning models. Over the past two decades, many semi-supervised classification methods have been raised on the basis of different assumptions, such as smoothness, low-density and manifold (van Engelen & Hoos, 2020).

According to the distinct optimization procedures, the semi-supervised classification algorithms can be divided into two groups, namely inductive algorithms and transductive algorithms (van Engelen & Hoos, 2020). The inductive algorithm aims to form a classification model to predict the whole input space. Among these, generative mixture models and expectation–maximization (EM) are considered as the earliest semi-supervised learning method (Zhu, 2008), which needs the identifiability and model correctness to maintain the performance. The wrapper methods are the most commonly used, which train supervised base classifiers on labeled instances and utilize pseudo-labeled instances to augment the performance, including self-training (Yarowsky, 1995), co-training (Blum & Mitchell, 1998), etc. Theoretically, any supervised classifier can be used as a base learner in this group of methods, which is deemed as one of the most significant advantages.

Another group of semi-supervised methods is transductive algorithms, which can only predict the given set of unlabeled instances. The semi-supervised SVM (S^3VM) was proposed as the extension of SVM to the semi-supervised learning, of which transductive SVM (TSVM) (Joachims, 1999; Vapnik, 1998) aims to find the maximum margin on labeled and unlabeled instances. But the TSVM encounters the problem of NP-hard. In addition, the graph-based methods define a graph over labeled and unlabeled instances and reflect the pairwise similarity using edges (Zhu et al., 2005). The more similar the edges, the more likely two instances share a same label. The graph-based methods include Mincut (Blum & Chawla, 2001), Gaussian Random field and Harmonic Functions (Zhu et al., 2003), etc. But the graph construction relies on domain knowledge and has high time complexity.

In general, the self-training method expands the training set with predictions on unlabeled instances. It is easy to operationalize and has great flexibility in threshold

setting. This gives more choices on base classifier selection and has been utilized in many domains, such as word sense disambiguation (Yarowsky, 1995), object detection (Rosenberg et al., 2005), sentence subjectivity classification (Wang et al., 2008), sentiment classification (He et al., 2011), and so on. Furthermore, it has been shown the effectiveness in improving the predictive performance of base classifiers (Li et al., 2008; Tanha et al., 2017; Zhang et al., 2021). Therefore, to make full use of the unlabeled instances, the semi-supervised self-training method is preferred to identify important citations in this paper.

Methodology

Figure 1 depicts the sketch of our research framework on important citations identification, which is based on the full-text articles. After the preprocessing steps, six groups of features are extracted in the *feature engineering* module. Then the whole dataset is divided into labeled dataset and unlabeled dataset, which are fed to the SVM and RF models with self-training strategy to identify important citations. The self-training strategy, preprocessing, feature engineering, and statistics of datasets will be described in more details in the following sections in turn.

Self-training strategy

The main idea of self-training strategy is to train a base classifier on a small volume of labeled instances and make predictions on large amount of unlabeled instances. Then, pseudo-labeled instances with high level of confidence are selected to expand the scale of labeled dataset. After that, the model is retrained on newly synthesized labeled dataset. This process is iterated until no new instance meets the condition. A significant advantage of this method is that any supervised model can be used as a base classifier in theory (van Engelen & Hoos, 2020).

Figure 2 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 dataset with fivefold cross validation procedure. After learning the training set of each fold, the labels of the unlabeled dataset are predicted respectively. We select the samples with 95%, 90%, 85%, 80%, 75%, and 70% confidence levels as the pseudo-labeled dataset to rejoin the training set. For each fold, the model is retrained on the new combined dataset and evaluated on the test set. The involved parameters are optimized correspondingly. The areas under the curve of PR and ROC are

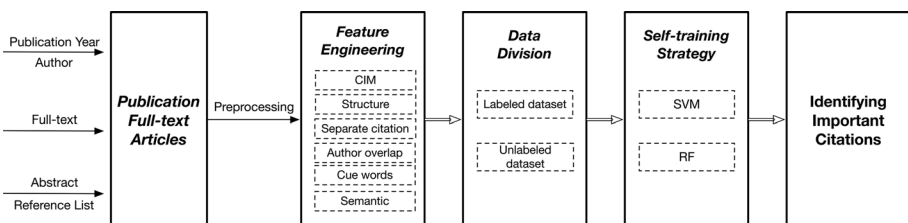


Fig. 1 Research framework on identifying important citations

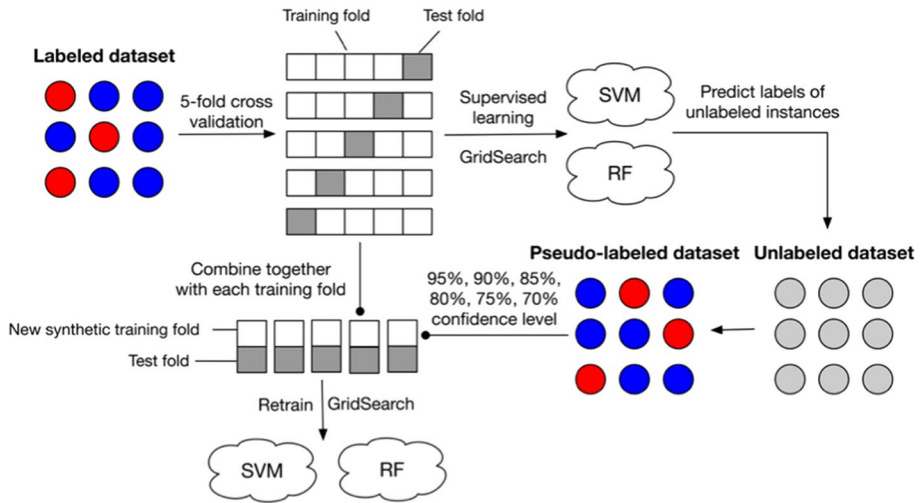


Fig. 2 Framework of semi-supervised self-training strategy for identifying important citations

used as indicators for evaluating the performance. Please refer to the pseudo code in Algorithm 1 for more details on our methodology for identifying important citations (in our case, $V = 5$).

Algorithm 1 Algorithm for identifying important citations with semi-supervised self-training strategy

Input: L is the labeled dataset, U is the unlabeled dataset, and V is the number of folds.

Output: The performance of the classifier in term of AUC-PR and AUC-ROC;

- 1: To calculate the features G1 to G6 (cf. Subsection *Feature Engineering*);
- 2: To randomly split the labeled dataset L into V folds with nearly equal size and similar category distribution: L_1, L_2, \dots, L_V ;
- 3: To initialize $L_{real} = []$, $L_{prob} = []$;
- 4: **for** $i \leftarrow 1$ to V **do**
- 5: To train a base classifier (such as SVM, RF) Classifier = Classifier ($L - L_i$) after optimizing the parameters with grid search (Xu et al., 2007);
- 6: To predict labels for the unlabeled dataset U : Classifier.predict_prob(U);
- 7: To select instances above a certain confidence level as the pseudo-labeled dataset U' to rejoin the training set $L'_i \leftarrow (L - L_i) \cup U'$;
- 8: To retrain the classifier: Classifier = Classifier (L') after optimizing the parameters with grid search (Xu et al., 2007);
- 9: To estimate the posterior probability of each instance in the test set: $Probs = \text{Classifier.predict_prob}(L_i)$;
- 10: $L_{real}.append(L_i)$, $L_{prob}.append(Probs)$;
- 11: **end for**
- 12: To output the AUC-PR (L_{real}, L_{prob}) and AUC-ROC (L_{real}, L_{prob})

Preprocessing

The preprocessing includes the following steps: (1) The citing papers are collected in the format of PDF and then converted to text format with the Xpdf toolkit (<http://xpdfreader.com>). (2) The textual data is parsed by the ParsCit software (Councill et al., 2008) to extract title, author list, abstract, main body, and references of each citing paper. It is worth noting that the ParsCit software can normalize each section in a citing publication into a generic section header (*introduction, related work, method, experiment, discussion, and conclusion*). To avoid parsing mistakes, each parsed document is checked carefully and corrected manually. (3) The citation contexts are extracted on the basis of regular expressions. (4) All textual information including citation contexts and abstract is cleaned with Natural Language Toolkit (NLTK).

Feature engineering

As for the feature engineering, the following six groups of features from our previous study (An et al., 2021a) are utilized here, as shown in Table 1. The effectiveness of these features has been verified on identifying important citations. $G1$ contains two generative features extracted from the CIM model (Citation Influence Model) (Dietz et al., 2007; Xu et al., 2019), which incorporates the topical innovation and topical inheritance via citations on the basis of the first-order Markovian assumption. One is a multinomial distribution of references, which reflects the importance degree of a cited publication to a citing publication. The other is the symmetrized Kullback–Leibler divergence between multinomial distribution of topics specific to a pair of citing and cited publications, which represent the similarity in the topic space between citing and cited papers.

According to the structure of a paper, $G2$ counts the number of each cited paper mentioned in each section (e.g., *introduction*, *related work* and so on) of a citing paper. In actual situations, a cited publication may be mentioned separately in the text of a citing publication, or mentioned together with other cited ones. Abu-Jbara et al. (2013) argued that a separate cited publication was usually more important to a citing publication than a group of cited ones. Therefore, $G3$ devotes to calculating the proportion of a cited publication mentioned separately in a citing paper.

Valenzuela et al. (2015) observed that if a pair of citing-cited articles was simultaneously authored by one or more scholars, this might indicate that the citing article extends the cited one. Hence, the Jaccard similarity coefficient between the citing authors and cited authors is generated as $G4$. This feature can reflect the development of self-work. Indeed, it is rare that a pair of citing-cited articles share common authors. That is to say, in most cases, this feature is assumed to zero. But according to our observation (An et al., 2021a), this feature with non-zero provides a signal on distinguishing important citations.

$G5$ counts the number of important/incidental cue words appearing in citation contexts, which may reveal the citing intention of authors. Specific words can hint different intentions of citing paper and reveal different importance thereby. For example, “according to” or “use” may indicate that a method/technique/theory from a cited paper is utilized in a citing paper. From over 80 papers pertinent to citation behavior, Hassan et al. (2017) compiled 81 important cue words and 51 incidental ones. The calculation of $G5$ is based on these cue words. Finally, $G6$ is derived by calculating the cosine similarity between the citation contexts and the abstract of a focal cited paper. The similarities are averaged if the cited paper is mentioned several times in a citing article. Please refer to An et al. (2021a) for more details.

Datasets

Two types of datasets are used to evaluate the performance of the proposed methodology in this paper.

Dataset I: This dataset was collected from a collection of 20,527 papers in ACL anthology with 106,509 citations, and 465 citations were randomly selected (Valenzuela et al., 2015). The citations were manually annotated by one expert with the label 0 (related work), 1 (comparison), 2 (using the work), and 3 (extending the work) according to the citation contexts in the citing papers. The inner-annotator agreement was verified between

Table 1 Features utilized for important citation identification

Group	Feature	#of features
G1: CIM model based feature	The importance degree of a cited publication to the citing publication and the similarity in the discovered topic space between citing and cited publications extracted from the CIM model	2
G2: Structure based feature	Number of citations mentioned in the each section (<i>introduction, related work, method, experiment, discussion, conclusion</i>) and the total number of unique sections that citations appear	7
G3: Separate citation based feature	Proportion of a cited publication mentioned separately in a citing paper	1
G4: Author overlap based feature	Jaccard similarity coefficient between author lists of citing and cited papers	1
G5: Cue words based feature	Number of important/incidental cue words appearing in citation contexts	2
G6: Semantic relevancy based feature	Cosine similarity between the citation contexts and cited paper's abstract on the basis of the TF-IDF	1

Table 2 Statistics of *Dataset I* and *Dataset II*

Dataset	#of unique citing papers	#of unique cited papers	#of labeled pairs of citing and cited papers	#of unlabeled pairs of citing and cited papers
<i>Dataset I</i>	434	4,589	456	8,085
<i>Dataset II</i>	112	2,579	2,685	0

Table 3 Summary of labeled *Dataset I* and *Dataset II*

Dataset	Label	Class	Number of Samples
<i>Dataset I</i>	0	Incidental	389 (85.3%)
	1	Important	67 (14.7%)
<i>Dataset II</i>	0	Incidental	2,374 (88.4%)
	1	Important	311 (11.6%)

two experts for a subset of the dataset to reduce the bias raised by human annotation and reached 93.9% (Valenzuela et al., 2015). In this study, we combine the “related work” and “comparison” categories into incidental class with the label 0 and “using the work” and “extending the work” categories into important class with the label 1.

Table 2 lists the statistics of this dataset. Note that during the preprocessing steps, several citing-cited paper pairs with missing information, non-English language or non-article document type are removed. From 465 labeled pairs, 434 citing papers and 4,589 unique cited papers are collected after the preprocessing steps. In the end, this yields 8,541 pairs of citing-cited papers totally with 456 annotated pairs and 8,085 un-annotated ones. The summary of labeled dataset is listed in Table 3, of which 14.7% are important citations. Similar to the labeled dataset, the feature engineering and preprocessing are also conducted on the unlabeled dataset.

Dataset II: This dataset is very different from *Dataset I*. This is an author-labeled dataset, which was collected in Zhu et al. (2015). Through the questionnaire survey, about 40 researchers were required to indicate the most essential references for them in each paper they provided. According to previous classification approach (An et al., 2021a), the references marked as “essential” are viewed as important citations, and the others as incidental citations. After the same preprocessing steps, 112 citing papers and 2,579 unique cited papers are collected. This yields 2,685 labeled citing-cited paper pairs in total, of which only 11.6% are important citations. Table 2 shows the statistics of *Dataset II*, and the summary of labeled dataset is showed in Table 3. It is worth mentioning that *Dataset II* involves 10 different disciplines, as illustrated in Table 4. About 70% of citing papers are from the discipline of *Computer Science*.

Experimental results and discussion

As shown in Fig. 3, two experiments based on two different types of datasets are conducted. *Experiment I* is based on *Dataset I*, which is an expert-labeled dataset from one single domain (Valenzuela et al., 2015). *Experiment II* is based on *Dataset II*, which is an author-labeled dataset from multiple domains (Zhu et al., 2015). In *Dataset I*, only a part

Table 4 Statistics of disciplines in *Dataset II*

Discipline	#of authors	#of papers	Discipline	#of authors	#of papers
Computer Science	29	82	Signal Processing	1	4
Genetics	2	6	Mathematics	1	2
Biophysics	2	6	Physics	1	1
Ecology	1	5	Chemistry	1	1
Geophysics	1	4	Translation	1	1

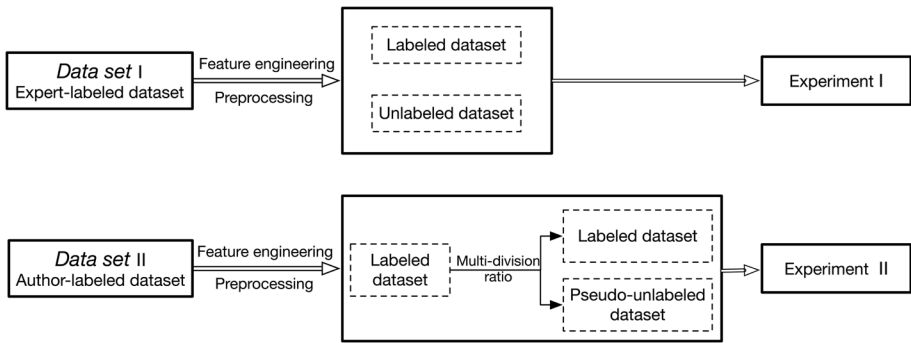


Fig. 3 Experimental framework on identifying important citations

of citations was actually annotated. Thus, one can readily obtain a labeled dataset and an unlabeled dataset. However, all citations in *Dataset II* were labeled manually by the resulting authors. Therefore, to reserve a data subset as unlabeled one, multi-division ratio is implemented on *Dataset II* for the follow-up *Experiment II*.

Experiment I

The first experiment is based on *Dataset I*, which contains 456 labeled citing-cited paper pairs and 8,085 unlabeled ones. As two state-of-the-art discriminative models, SVM and RF are utilized here as our base classifiers. First of all, these two models are trained on the labeled dataset. To tune the parameters of these two classifiers, grid search with fivefold cross-validation (Xu et al., 2007) is used in this study. Figure 4 shows the PR curves and ROC curves of SVM and RF models. 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 (An et al., 2021a; Hassan et al., 2017, 2018a; 2018b; Qayyum & Afzal, 2019; Valenzuela et al., 2015; Wang et al., 2020; Zhu et al., 2015).

Then, a semi-supervised self-training on the unlabeled dataset is conducted. After learning the training set of each fold, the labels of the unlabeled dataset are predicted. We select the samples with 95%, 90%, 85%, 80%, 75%, and 70% confidence levels to rejoin the training set. Table 5 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 dataset and

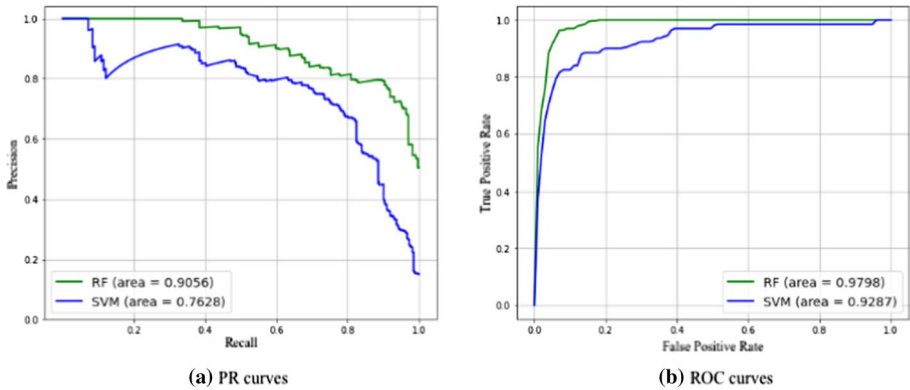


Fig. 4 The PR curves (a) and ROC curves (b) of SVM and RF models on labeled dataset with supervised learning strategy

Table 5 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 6 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

evaluated on the test set. Similarly, grid search is also used to tune the involved parameters. Table 6 reports the results of mean AUC-ROC and AUC-PR of 5 folds under different confidence level. It can be seen that the AUC-PR and AUC-ROC for the SVM model reach the maximum at the 75% confidence level, which are 0.8102 and 0.9622 respectively. The

Table 7 Performance of benchmark methods on the dataset in Valenzuela et al. (2015)

Benchmark methods	SVM		RF	
	PR	ROC	PR	ROC
Valenzuela et al. (2015)	0.80	–	0.80	–
Hassan et al. (2017)	0.78	0.85	0.84	0.91
Hassan et al. (2018a)	0.88	0.91	0.89	0.95
Wang et al. (2020)	–	0.94	–	0.92

Table 8 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)	3.50
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)	3.50

RF model has the highest AUC-PR and AUC-ROC at 95% confidence level (0.9248 and 0.9841). These two models outperform the supervised learning counterparts and several benchmark methods in Table 7, including Valenzuela et al. (2015), Hassan et al. (2017), Hassan et al. (2018a) and Wang et al. (2020). That is, our semi-supervised methodology with the SVM and RF as base classifiers performs best in terms of PR and ROC.

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 8 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 fivefold cross validation by controlling for structure features (G2). For each combination, the resulting parameters are optimized separately. As we can observe, the baseline model based on the structure features can achieve the mean AUC-PR of about 0.7600 and 0.7903, and AUC-ROC of about 0.8906 and 0.4743. The author-overlap based features (G4) ranks first, which promotes respectively the AUC-PR to 0.9462 and 0.8145, AUC-ROC to 0.8145 and 0.4798. The CIM (Citation Influence Model) (Xu et al., 2019) 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 previous work (An et al., 2021a).

Experiment II

For purpose of checking the generalizability of the proposed method, the second experiment is conducted on *Dataset II*, which consists of 2,685 author-labeled citing-cited paper

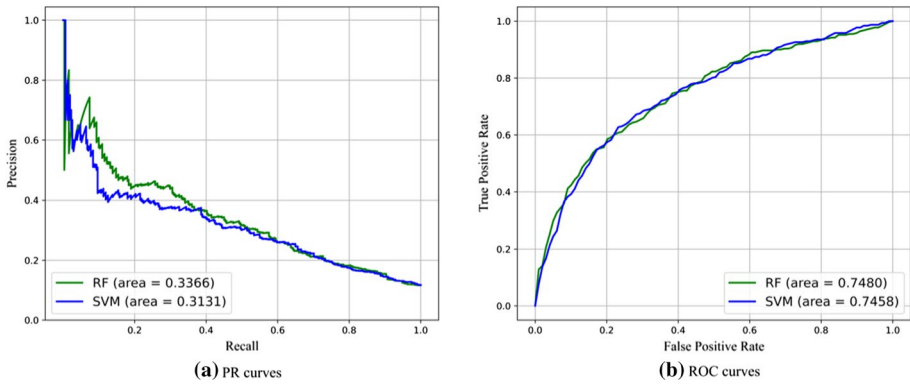


Fig. 5 The PR curves (a) and ROC curves (b) of SVM and RF models on the *Dataset II* with supervised learning strategy

Table 9 Statistics of labeled data and pseudo-unlabeled data under different division ratios in *Dataset II*

Division ratio	Number of labeled citations		Number of pseudo-unlabeled citations	
	Total	Important Class	Total	Important Class
10%	268	31	2,417	280
15%	402	47	2,283	264
20%	537	62	2,148	249
25%	671	78	2,014	233
30%	805	93	1,880	218

pairs. First of all, similar to Experiment I, the SVM and RF models with supervised learning strategy are conducted on all 2,685 labeled instances with 6 groups of features under fivefold cross-validation procedure. Figure 5 shows the PR curves (a) and ROC curves (b) of the SVM and RF models on *Dataset II* with supervised learning strategy. The SVM and RF models can obtain 0.7458 and 0.7480 in term of AUC-ROC respectively, and 0.3131 and 0.3366 in term of AUC-PR respectively. It is very easy to see that these two models perform worse on *Dataset II* than on *Dataset I* (cf. Subsection *Experiment I*). Main reasons can be attributed to two points: (1) different disciplines of *Dataset II* follow different citation patterns; (2) it is very possible that the authors in *Dataset II* only annotated the most essential references, but ignored the less essential ones. The annotation guideline for *Dataset II* is very different from that for *Dataset I*. Hence, the patterns for important citations may vary between in *Dataset II* and in *Dataset I*. Further verification will be discussed in the following subsection "[Discussion](#)".

To conduct semi-supervised learning task, 10%, 15%, 20%, 25%, 30% of *Dataset II* are randomly divided as labeled dataset and the rest as pseudo-unlabeled dataset. In order to ensure the consistency of the data characteristics, the labeled dataset and unlabeled dataset follow approximately the same category distribution. Table 9 lists the statistics of labeled dataset and pseudo-unlabeled dataset under different division ratios.

According to Experiment I, SVM model reaches the maximum at the 75% confidence level and RF model performs best at 95% confidence level. Therefore, SVM model under

75% confidence level and RF model under 95% confidence level are deployed in this experiment. That is to say, after learning the training set of labeled instances of each fold and predicting the unlabeled instances, samples above 75% confidence level of SVM and above 95% confidence level of RF are rejoined to the training set. It is noteworthy that we also try other confidence levels in our experiments, but no performance improvement is observed. Figure 6 shows the trends of PR and ROC of SVM and RF with semi-supervised strategy under different division ratios. The red lines denote the supervised learning counterparts.

From the perspective of PR curves, similar trends can be observed for the SVM and RF models. In more details, the overall trend rises first, reaches the peak at 15% division ratio, and then falls below the corresponding curve of supervised learning counterpart at 30% division ratio. To say it in another way, the models with semi-supervised strategy are slightly better than their resulting counterparts in term of AUC-PR when the division ratio is less than 20%. However, as the proportion of labeled instances further increases, the performance of semi-supervised learning shows a downward trend. We ascribe the trend to the mixed disciplines in *Dataset II*. From Section "Datasets", it is not difficult to see that *Dataset II* covers about 10 different disciplines, and the distribution of the number of documents in each discipline is extremely skewed. When the proportion of labeled dataset is expanded, more and more citations from other disciplines would be added to the training set. This means that more outlier instances would appear in the training set. Furthermore, too few instances in a certain discipline may make a machine learning model to be under-fitting. In this situation, the predicted labels for pseudo-labeled instances may differ greatly from the ground-truth labels. Consequently, the performance of classifiers tends to be worse. Therefore, for this type of mixed dataset, it should be better to ensure that there is a sufficient amount of instances in each discipline.

In addition, the ROC curves of both SVM and RF models fluctuate around the resulting curves of the supervised learning counterparts. Since a large change in the number of

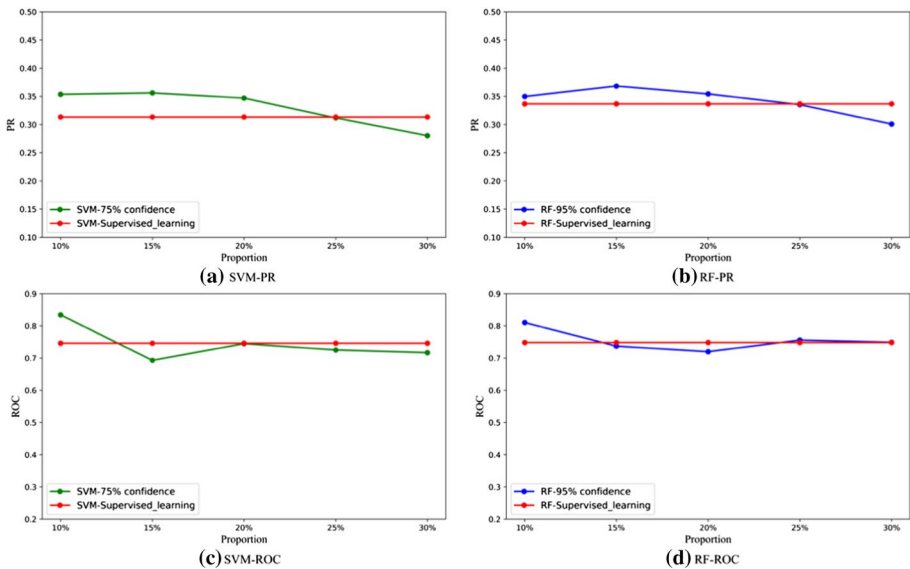


Fig.6 The PR and ROC of SVM under 75% confidence level and RF under 95% confidence level with semi-supervised strategy under different division ratios

Table 10 Performance of SVM model with semi-supervised strategy on the discipline of *Computer Science*

Division ratio	SVM-under 75% confidence level	
	PR	ROC
10%	0.1760	0.6264
15%	0.1730	0.5958
20%	0.2737	0.7248
25%	0.3514	0.7375
30%	0.3179	0.7267

Table 11 The performance of SVM and RF models with different groups of features in *Dataset II* in terms of mean AUC-PR, AUC-ROC, and their ranks

Feature	SVM		RF		Average_rank
	PR	ROC	PR	ROC	
G2	0.2294(5)	0.7136(6)	0.3010(2)	0.3653(3)	4
G2+G1	0.3093(1)	0.7384(1)	0.2741(6)	0.3685(1)	2.25
G2+G3	0.3082(2)	0.7264(4)	0.3236(1)	0.3651(4)	2.75
G2+G4	0.3013(3)	0.7326(2)	0.2989(4)	0.3661(2)	2.75
G2+G5	0.3010(4)	0.7323(3)	0.2836(5)	0.3537(6)	4.5
G2+G6	0.2218(6)	0.7235(5)	0.2996(3)	0.3650(5)	4.75

false positives can only lead to a small change in the false positive rate when the dataset is highly skewed, the ROC is not sensitive to the highly skewed datasets (Davis & Goadrich, 2006). Therefore, the ROC curve may not be suitable for this highly skewed dataset. That is, the PR curve can provide a more informative picture than the ROC curve for our *Dataset II*.

Then, the semi-supervised learning strategy is conducted separately on the discipline of *Computer Science*. When the RF model serves as the base classifier, too few new samples are rejoined to the training dataset. Hence, the experiment is only implemented with SVM model as the base classifier. Table 10 shows the performance of SVM model under 75% confidence level. The performance is not satisfactory when the division ratio is lower than 25%. As the proportion of training instances increases, the performance tends to get better and exceeds the supervised learning counterpart at 25% division ratio. But overall, the performance is far from our expectation. Different annotation guideline for important citations in *Dataset II* should result in un-satisfactory performance on this dataset.

Finally, similar to Experiment I, the contributions of each group of features are further analyzed. By controlling for structure features (G2), Table 11 shows the performance of SVM and RF models with different groups of features in terms of mean AUC-PR, AUC-ROC and their rankings (in parentheses) as well as the average rank. The baseline model based on the structure features (G2) can achieve the mean AUC-PR of about 0.2294 and 0.3010, and AUC-ROC of about 0.7136 and 0.3653. The CIM model based features (G1) rank first, which confirms again its effectiveness on identifying important citations. Separate citation based feature (G3) and the author-overlap based features (G4) rank the second. Nevertheless, semantic relevancy based feature (G6) performs even worse than the baseline model. Compared with *Dataset I*, the features generated from the generative model have

Table 12 Statistics of citations for disciplines of Computer Science, Genetics, Biophysics and Ecology in *Dataset II*

Discipline	#of important citations	#of incidental citations	#of total citations	#of proportion of important citations
Computer Science	218	1,374	1,592	13.7%
Genetics	21	300	321	6.5%
Biophysics	18	230	248	7.3%
Ecology	12	190	202	5.9%

Table 13 The performance of supervised learning strategy on the disciplines of Computer Science, Genetics, Biophysics and Ecology in *Dataset II*

Discipline	SVM		RF	
	PR	ROC	PR	ROC
Computer Science	0.3261	0.7335	0.3787	0.7493
Genetics	0.1764	0.5689	0.1107	0.6300
Biophysics	0.4067	0.7712	0.4428	0.7675
Ecology	0.1150	0.6542	0.0650	0.5566

more contributions in *Dataset II*, and the contribution of the author-overlap based features drop slightly. As we can observe, the features have different contributions on the classification performance of important citations in the two datasets. Again, this may be caused by the differences in the criteria that determine important citations.

Discussion

To validate our speculation about different citation patterns across different disciplines, top four disciplines are chosen according to data volume (cf. Table 4): *Computer Science*, *Genetics*, *Biophysics*, and *Ecology*. In each discipline, the number of citing papers is over 5 and the number of pairs of citing-cited papers is more than 200. Table 12 lists the distribution of important and incidental citations in these four disciplines. Table 13 shows the PR and ROC values of SVM and RF in each discipline. Two interesting phenomena can be observed as follows. (1) The performance varies between different disciplines, and the performance on *Computer Science* and *Biophysics* disciplines outperforms that on the others as well as the whole *Dataset II*. (2) The larger the ratio of important citations, the better the performance (cf. Tables 12, 13). In more details, these four disciplines can be sorted in term of performance as follows: *Biophysics* > *Computer Science* > *Genetics* > *Ecology*. As a matter of fact, compared to *Computer Science* discipline, the ratios of important citations in the other three disciplines are very close, but their performance varies greatly in term of AUC-PR. Hence, different citation patterns across different disciplines should be very closely related to different performance.

Though, the performance on *Computer Science* discipline is far from that in *Dataset I*. Now, let us further check the different annotation guideline for these two datasets. More specifically, we want to determine whether or not there exists several references that are non-essential for the resulting authors, but are actually very important for knowledge diffusion. For this purpose, an article with id=Z002 is randomly selected from *Dataset II*. For

convenience, this article's title and its partial references are reported in Table 14. In total, this article cited 18 documents, among which [14] is annotated the most essential one by its author. Hence, important class is attached to [14] under our framework, leaving other references into incidental class. However, after analyzing the citation contexts and citation features of other references, [2] and [8] are also very important to the citing paper Z002. Table 15 shows the citation features of references in Z002. [2] and [8] are mentioned several times and appear in more sections. From the citation contexts, it is not difficult to see that the citing paper Z002 actually uses or refers to the method proposed in [2] and [8]. Therefore, [2] and [8] are also very meaningful to the citing paper Z002. To summarize, due to different annotation guideline in *Dataset II* from that in *Dataset I*, non-essential references in *Dataset II* are actually made up of important and non-important citations.

Conclusions

Effectively identifying important citations has great practical significance in the bibliometrics, which can contribute to fair distribution of academic resources and fair evaluation of talents. The supervised learning methods are the most commonly used for this task, which rely on large amount of labeled instances to maintain the performance. However, only a small volume of instances have been annotated manually with the labels in real-world applications, which greatly limit the generalizability of the supervised learning techniques for this task.

In this paper, a semi-supervised self-training strategy is proposed to identify important citations by leveraging labeled instances and unlabeled instances to promote the performance and generalization ability of supervised learning models. To demonstrate the effectiveness of our proposed strategy, two different types of datasets are utilized here: (1) an expert-labeled dataset from one domain (Valenzuela et al., 2015), and (2) an author-labeled dataset from multiple disciplines (Zhu et al., 2015). Through the semi-supervised self-training on the unlabeled dataset of the expert-labeled dataset from one domain, 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 outperforms the benchmark methods proposed by Valenzuela et al. (2015), Hassan et al. (2017), Hassan et al. (2018a) and Wang et al. (2020). This demonstrates the effectiveness of our semi-supervised self-training strategy for important citations identification. Additionally, the CIM model based features, structural based features and author-overlap based features contribute greatly on important citations identification.

As for the author-labeled dataset from multiple disciplines, the semi-supervised learning model can perform better than the resulting supervised learning counterpart in term of AUC-PR when the rate of labeled instances is less than 20%. The performance of semi-supervised learning models may decline as the proportion of labeled instances further increases. The main reason may be ascribed to the various citation patterns in different disciplines and different annotation guideline followed by this dataset. These speculations have been verified by our further experiments on four main disciplines and in-depth analysis on citation contexts of non-essential references. Furthermore, the difference in the criteria that determine important citations causes the different contributions of each group of features in the classification performance in the two datasets. In addition, we argue that the

Table 14 The citing paper Z002 and its partial references in *Dataset II*

Citing paper ID	Title of citing paper	Cited paper ID	Title of cited paper	Class
Z002	slope one predictors for online rating-based collaborative filtering	[14]	item-based collaborative filtering recommendation algorithms	Important
		[2]	empirical analysis of predictive algorithms for collaborative filtering	Incidental
		8]	an algorithmic framework for performing collaborative filtering	Incidental

Table 15 Citation features of several references in the citing paper Z002

Cited paper ID	Number of citations	Citation position	Partial citation contexts
[14]	3	Related work, method	<p>(1) We also use naive weighting. It was observed in [14] that even their regression-based $f(x) = ax + b$ algorithm didn't lead to large improvements over memory-based algorithms</p> <p>(2) We also compare to the item-based approach that is reported to work best [14], which uses the following adjusted cosine similarity measure, given two items i and j</p>
[2]	5	Related work, Method, Experiment	<p>(1) Following [2, 8], we set $\gamma(u, w) = \text{Corr}(u, w) \text{Corr}(u, w) ^{\rho-1}$ with $\rho = 2.5$, where ρ is the Case Amplification power</p> <p>(2) To do so, we have used the All But One Mean Average Error (MAE) [2].</p>
[8]	4	Related work, Method, Experiment	<p>(1) Following [8, 11], we used enough evaluations to have a total of 50,000 ratings as a training set (\mathcal{X}) and an additional set of evaluations with a total of at least 100,000 ratings as the test set (\mathcal{X}')</p> <p>(2) Another simple scheme is known as BIAS FROM MEAN (or sometimes NON PERSONALIZED [8])</p>

AUC-PR curve should be preferred to when the dataset from multiple disciplines has an extremely skewed distribution.

In our next work, the potential of deep learning models with semi-supervised strategy should be exploited for identifying important citations. Additionally, a scientific verification of our semi-supervised framework still needs to be further investigated on the dataset from multiple disciplines in the near future.

Acknowledgements The present study is an extended version of an article (An et al., 2021b) presented at the first Workshop on AI + Informetrics at the iConference 2021, 17 March, 2021. This research received the financial support from the National Natural Science Foundation of China under grant number 72004012 and 72074014.

References

- Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 596–606).
- Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., & Afzal, M. T. (2021). Important citation identification using sentiment analysis of in-text citations. *Telematics and Informatics*, 56, 101492.
- An, X., Sun X., Xu, S. (2021b). Important citations identification with semi-supervised classification model. *The first Workshop on AI + Informetrics at the iConference 2021*.
- An, X., Sun, X., Xu, S., Hao, L., & Li, J. (2021a). Important citations identification by exploiting generative model into discriminative model. *Journal of Information Science*. <https://doi.org/10.1177/0165551521991034>
- Bennett, K., & Demiriz, A. (1999). Semi-supervised support vector machines. *Advances in Neural Information Processing Systems*, 368–374.
- Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory* (pp. 92–100).
- Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 19–26).
- Chapelle, O., Sindhwani, V., & Keerthi, S. S. (2008). Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9(2), 203–233.
- Councill, I. G., Giles, C. L., & Kan, M. Y. (2008). ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 661–667).
- Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240).
- Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 233–240). ACM.
- Dong, C., & Schäfer, U. (2011). Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 623–631).
- Garfield E. (1965). Can citation indexing be automated. In *Proceedings of the Symposium on Statistical Association Methods for Mechanized documentation* (pp. 189–192).
- Garfield, E. (2006). Citation indexes for science. A new dimension in documentation through association of ideas. *International Journal of Epidemiology*, 35(5), 1123–1127.
- Hassan, S. U., Akram, A., & Haddawy, P. (2017). Identifying important citations using contextual information from full text. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 1–8). IEEE.
- Hassan, S. U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018a). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics*, 117(3), 1645–1662.
- Hassan, S. U., Safder, I., Akram, A., & Kamiran, F. (2018b). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, 116(2), 973–996.
- He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing and Management*, 47(4), 606–616.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572.

- Iqbal, S., Hassan, S. U., Aljohani, N. R., Alelyani, S., Nawaz, R., & Bornmann, L. (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies. *Scientometrics*, 126(8), 6551–6599.
- Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning* (pp. 200–209).
- Lazaridis, T. (2010). Ranking university departments using the mean h-index. *Scientometrics*, 82(2), 211–216.
- Li, X., He, Y., Meyers, A., & Grishman, R. (2013, September). Towards fine-grained citation function classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp. 402–407).
- Li, Y., Guan, C., Li, H., & Chin, Z. (2008). A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system. *Pattern Recognition Letters*, 29(9), 1285–1294.
- Qayyum, F., & Afzal, M. T. (2019). Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics*, 118(1), 21–43.
- Radoulov, R. (2008). *Exploring automatic citation classification*. Master's thesis, University of Waterloo.
- Rosenberg, C., Hebert, M., & Schneiderman, H. (2005). Semi-supervised self-training of object detection models. In *Proceedings of the 7th IEEE Workshop on Applications of Computer Vision* (pp. 29–36).
- Tanha, J., van Someren, M., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, 8(1), 355–370.
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110).
- Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. In *The 2015 AAAI Workshop on Scholarly Big Data: AI Perspectives, Challenges, and Ideas* (pp. 21–26).
- Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2), 373–440.
- Vapnik, V. (1998). *Statistical learning theory*. Springer.
- Wang, B., Spencer, B., Ling, C. X., & Zhang, H. (2008). Semi-supervised self-training for sentence subjectivity classification. In *Proceedings of the 21st Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 344–355). Springer, Berlin, Heidelberg.
- Wang, M., Zhang, J., Jiao, S., Zhang, X., Zhu, N., & Chen, G. (2020). Important citation identification by exploiting the syntactic and contextual information of citations. *Scientometrics*, 125(3), 2109–2129.
- Xu, S., Ma, F., & Tao, L. (2007). Learn from the information contained in the false splice sites as well as in the true splice sites using SVM. In *Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering* (pp. 65–71). Atlantis Press.
- Xu, S., An, X., Qiao, X., Zhu, L., & Li, L. (2011). Semi-supervised least-squares support vector regression machines. *Journal of Information and Computational Science*, 8(6), 885–892.
- Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics*, 13(4), 100983.
- Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (pp. 189–196).
- Zeng, T., & Acuna, D. E. (2020). Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. *Scientometrics*, 124(1), 399–428.
- Zhang, F., Pan, T., & Wang, B. (2021). Semi-supervised object detection with adaptive class-rebalancing self-training. arXiv preprint. arXiv:2107.05031.
- Zhu, X., Ghahramani, Z., & Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine learning* (pp. 912–919).
- Zhu, X., Lafferty, J., & Rosenfeld, R. (2005). *Semi-supervised learning with graphs*. Doctoral dissertation. Carnegie Mellon University.
- Zhu, X. J. (2008). Semi-supervised learning literature survey. Technical Report. University of Wisconsin-Madison.
- Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2), 408–427.

Authors and Affiliations

Xin An¹ · Xin Sun² · Shuo Xu³ 

Xin Sun
sunx@istic.ac.cn

Shuo Xu
xushuo@bjut.edu.cn

¹ School of Economics and Management, Beijing Forestry University, Beijing 100083, People's Republic of China

² Institute of Scientific and Technical Information of China, Beijing 100038, People's Republic of China

³ College of Economics and Management, Beijing University of Technology, Beijing 100124, People's Republic of China