VNU Journal of Science: Comp. Science & Com. Eng, Vol. 34, No. 2 (2019) 1–10

On Rectifying the Mapping between Articles and Institutions in Bibliometric Databases

Ngo Kien Tuan, Vo Dinh Hieu*, Bui Ngoc Thang, Le Viet Anh, Pham Khanh Ly

Faculty of Information Technology, VNU University of Engineering and Technology,
No. 144 Xuan Thuy Street, Dich Vong Ward, Cau Giay District, Hanoi, Vietnam

Abstract

Today, bibliometric databases are indispensable for researchers and research institutions. The main functions of these databases are finding research articles, estimating the performance of researchers and organizations. Regarding evaluation of research performance of an organization, accuracy in determining institutions of authors of articles is decisive. However, current popular bibliometric databases such as Scopus and Web of Science have not addressed this point sufficiently. To this end, we propose an approach to revise authors' affiliation information of articles in bibliometric databases. We build a model to classify articles to institutions with high accuracy. To build the model, bag of words and n-grams techniques are employed to extract features of affiliation strings. After that, these features are weighted to determine their importance to each institution. Affiliation strings of articles are transformed into the new feature space by integrating weights of features and local characteristics of words and phrases contributing to the affiliations. Finally, on the feature space, the support vector classifier method is applied to learn a predictive model. Our experimental result shows that the proposed model's accuracy is about 99.1%.

Keywords: Affiliation, Disambiguation, Data cleaning, Classification, Supervised learning, if-iif, Support vector machine, Support vector classifier.

1. Introduction

Bibliometric databases play an important role in academic and research communities. These databases are used by scientists to find relevant research papers and proper journals to publish their research results. In addition, people may use these databases to assert the research performance of a scientist, a research group, an institution or even a country. Many university ranking systems such as THE [1], QS [2], and ARWU [3] rely on data from these bibliometric databases for their ranking methodologies. Today, beside PubMed, a bibliometric database for biomedical and life sciences researches, WoS [4] and Scopus [5] are considered as well known databases.

However, in recent years, some research works have shown that popular bibliometric databases are not accurate as expected. Franceschini and colleagues [6, 7] analysed and showed that many articles in these databases have lost their citations. More concretely, many papers are actually cited by some articles but these citations are not acknowledged by the databases. Some studies researched on the accuracy of citations [8]. Buchanan's work shows that there are many errors in mapping the cited articles to actual articles. Besides, the inaccuracy of authors' names in reference lists is remarkable. Some researchers analysed and pointed out that many papers are duplicated in these

[*] Corresponding author. Email.: hieudv@vnu.edu.vn

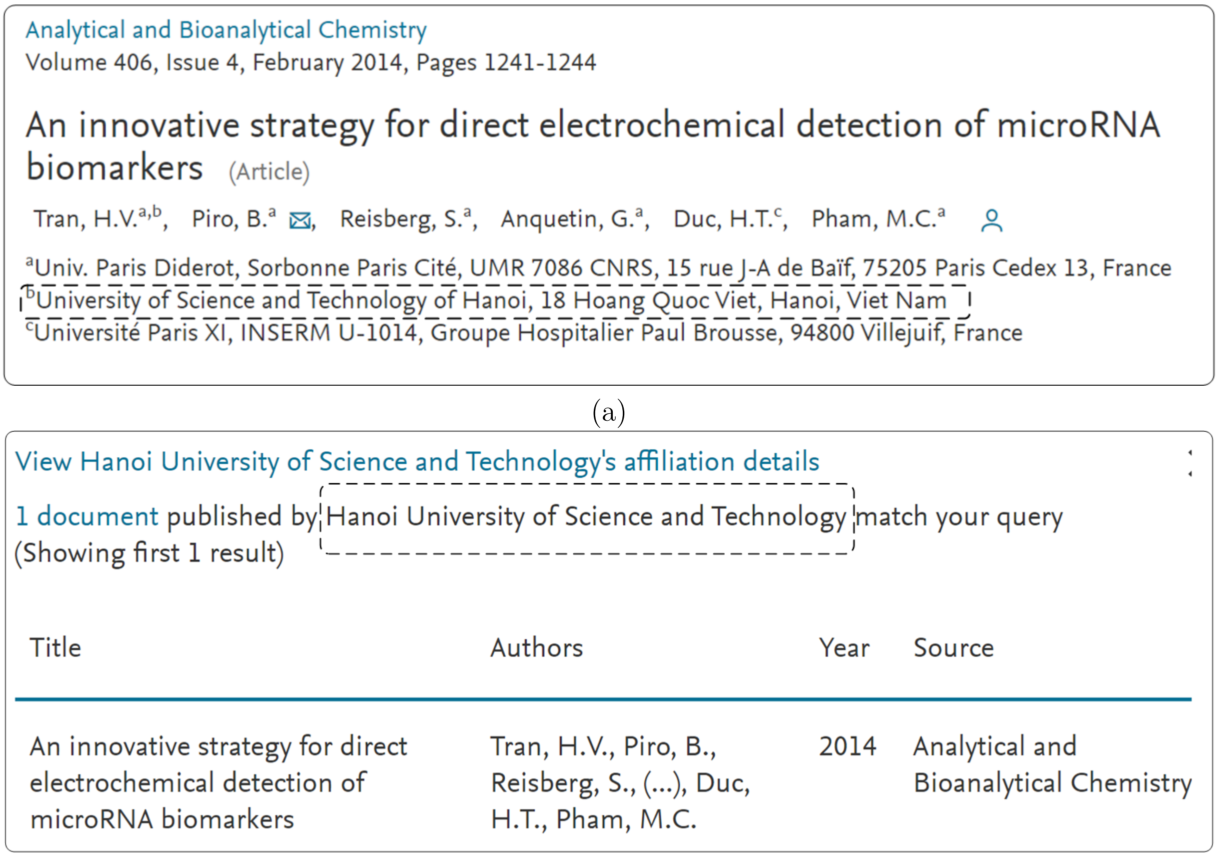
databases, .i.e. one paper is counted twice [9]. Junwen Zhu [10] and Shuo Xu [11] discovered errors related to DOI in WoS meanwhile Erwin Krauskopf [12] showed that Scopus missed a noticeable number of papers of some journals.

While there are several aspects related to the inaccuracy in bibliometric databases, in this work we only focus on affiliation information. The study of Weishu Liu and colleagues [13] pointed out that the lack of author address information in WoS is a significant problem. This problem was also presented in Krauskopf's research [14] [15]. It is common that the affiliation information written in research papers contains name of authors' faculties and universities. However, authors may provide their affiliation information in different manners depending on institutional policy and their habit. Some authors write detail information such as department, research group, address, and so on. In order to indicate research performance of institutions, WoS and Scopus map these written affiliations to the corresponding institutions. For example, in Scopus, the affiliation string "Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam" is mapped to Vietnam National University Hanoi. Examining a number of articles published by authors working at institutions in Vietnam, we found that both databases (Scopus and WoS) have remarkable mistakes in identifying institutions of authors. In some cases, it is authors are responsible for these mistakes. Authors may unclearly and incompletely provide their institution information. As a result, WoS or Scopus incorrectly maps the article to authors' institutions. In addition to the mistakes of authors, mistakes may be originated from algorithms for mapping between articles and institutions of Scopus and WoS. We have discovered that, in many cases, authors provide clear and complete institutional information but Scopus and WoS cannot accurately classify their articles to their right institutions. For example, the article "An innovative strategy for direct electrochemical detection of microRNA biomarkers" (DOI: 10.1007/s00216-013-7292-4) belongs to University of Sciences and Technology of Hanoi (USTH) but Scopus wrongly indicates that the paper belongs to Hanoi University of Sciences and Technology (HUST), an absolutely different institution (Fig.1).

In this paper, we propose a tool (named A2I) to help us to verify the mapping of articles to institutions in bibliometric databases. While most of the existing research works only focus on pointing out the problems with the quality of data in these databases, our research takes a further step. We provide a solution for automatic identification institutions of articles. The proposed tool only exploits basic techniques in Nature Language Processing and Machine Learning fields but works effectively. Our tool helps institutions confidently count the number of publications in Scopus and WoS. It also provides useful information that institutions can send to Scopus and WoS to claim their publications (which wrongly classified). The rest of the paper is organized as follows. The next part presents our method consisting of preprocessing, feature weighting and extracting, and learning a classification model stages. After that, we experiment with the proposed method and discuss the results before drawing up the conclusion.

2. Methodology

In this part, we present a method to verify the mapping articles to institutions. We consider the problem of verification the mapping as a classification problem. We restate the problem as follows. **G**iven a set



Analytical and Bioanalytical Chemistry
Volume 406, Issue 4, February 2014, Pages 1241-1244

An innovative strategy for direct electrochemical detection of microRNA biomarkers (Article)

Tran, H.V.[a,b], Piro, B.[a], Reisberg, S.[a], Anquetin, G.[a], Duc, H.T.[c], Pham, M.C.[a]

[a]Univ. Paris Diderot, Sorbonne Paris Cité, UMR 7086 CNRS, 15 rue J-A de Baïf, 75205 Paris Cedex 13, France
[b]University of Science and Technology of Hanoi, 18 Hoang Quoc Viet, Hanoi, Viet Nam
[c]Université Paris XI, INSERM U-1014, Groupe Hospitalier Paul Brousse, 94800 Villejuif, France

(a)

View Hanoi University of Science and Technology's affiliation details

1 document published by Hanoi University of Science and Technology match your query
(Showing first 1 result)

Title	Authors	Year	Source
An innovative strategy for direct electrochemical detection of microRNA biomarkers	Tran, H.V., Piro, B., Reisberg, S., (...), Duc, H.T., Pham, M.C.	2014	Analytical and Bioanalytical Chemistry

(b)

Hình 1. An example of error in Scopus (a) Affiliation information provided by authors; (b) Institution regconized by Scopus

$S = \{(s_i, y_i)\}_{i\in\{1...n\}}$ where s_i are affiliation strings and y_i are class labels. Each label represents an institution. We need to find a classifier f that can correctly map new affiliation string x to a corresponding label y. In other words, the classifier helps to correctly map affiliation strings to institutions and we can use this result to verify the current mapping between articles and institutions of bibliometric databases.

Our approach consists of two stages namely learning a classifier model and predicting institutions of articles. As shown in Figure 2, the main steps of the learning classifier model stage include affiliation strings extraction, data preprocessing, affiliation strings labeling, feature extraction and affiliation representation, and classifier model learning. The first step is to obtain affiliation data set including affiliation strings from bibliometric databases. The second step is to preprocess affiliation strings by removing noises, correcting missing data, and converting to strings encoded by American Standard Code (ASCII). After that, affiliation strings are manually labelled with institutions. In the fourth step, affiliation strings are secondly represented by significant statistical values of meaningful words and phrases that are extracted from affiliation strings by applying *Bag of Words* and $n - gram$ models. Statistical values of words and phrases for each affiliation string capture the local characteristics and the contribution level of the affiliation string to institutions.

On the feature space, we finally employ the support vector classifier method to train a model that can accurately classify affiliation strings to institutions. In the second stage, we use the learned classifier model to predict institutions of articles. In this stage, affiliation strings of articles are also transformed into the feature space by applying the steps mentioned in the first stage except for the labeling step. In the remaining part, the proposed approach is described in more detail.

2.1. Preprocessing affiliation strings

In order to learn a good representation of data, we remove noises and handle missing data from affiliation data. The preprocessing process consists of the following steps.

Step 1. Remove meaningless substrings: In this step, substrings playing no role in recognizing authors' institutions are removed from affiliation strings. Meaningless substrings are dots, ampersands, and newlines.

Step 2. Convert to ASCII: Affiliation strings may contain Unicode characters. In our approach, we convert affiliation strings to ASCII. Latin alphabet is used for building a character dictionary in purpose to transliterate character-by-character, and it generally produces satisfying results. For example, a Vietnamese affiliation string "Dept of Computer Science, HUST, 1Đại Cồ Việt, Hanoi, Vietnam" is converted to "Dept of Computer Science, HUST, 1Dai Co Viet, Hanoi, Vietnam".

Step 3. Seperate stuck words: By observing affiliation strings, we found that many affiliation strings contain stuck words. Seperating these words will help us build a better model. Regular expressions are used in this step. For example, the regular expressions of institutions' name and address are $(? <= [a - z])[-]?(? = [0 - 9A - Z])$ and $(? <= [0 - 9])(? = [A - Z][a - z]+)$, respectively. These fields must follow their regular expressions. If a character in a field does not match its regular expression, a space is inserted right after the character.

Step 4. Normalize to lower-case: Our approach does not take the style and format of affiliation strings into account. All affiliation string are converted into lower-case for further processing.

Figure 3 demonstrates these steps for the affiliation string "Dept. of Computer Science, HUST, 1Đại Cồ Việt, Hanoi, Vietnam". In the first step, the dot in the affiliation string is removed. The result of this step is "Dept of Computer Science, HUST, 1Đại Cồ Việt, Hanoi, Vietnam". In the second step, characters of the affiliation string are converted ASCII. Therefore, the string "Dept of Computer Science, HUST, 1Đại Cồ Việt, Hanoi, Vietnam" is transformed to "Dept of Computer Science, HUST, 1Dai Co Viet, Hanoi, Vietnam". In the next step, the stuck words "1Dai" is separated. In the final step, upper-case characters are converted to lower-case ones. After these steps, the original affiliation string is transformed to "dept of computer science, hust, 1 dai co viet, hanoi, vietnam".

2.2. Feature Extraction and affiliation representation

In this part, words and phrases are employed as features to represent affiliations of articles. Words and phrases of affiliation strings are extracted by applying two basic models. The first model, Bag of Words, is used to extract all the words in each affiliation string. The second model, $n-grams$, is used to get phrases, with n

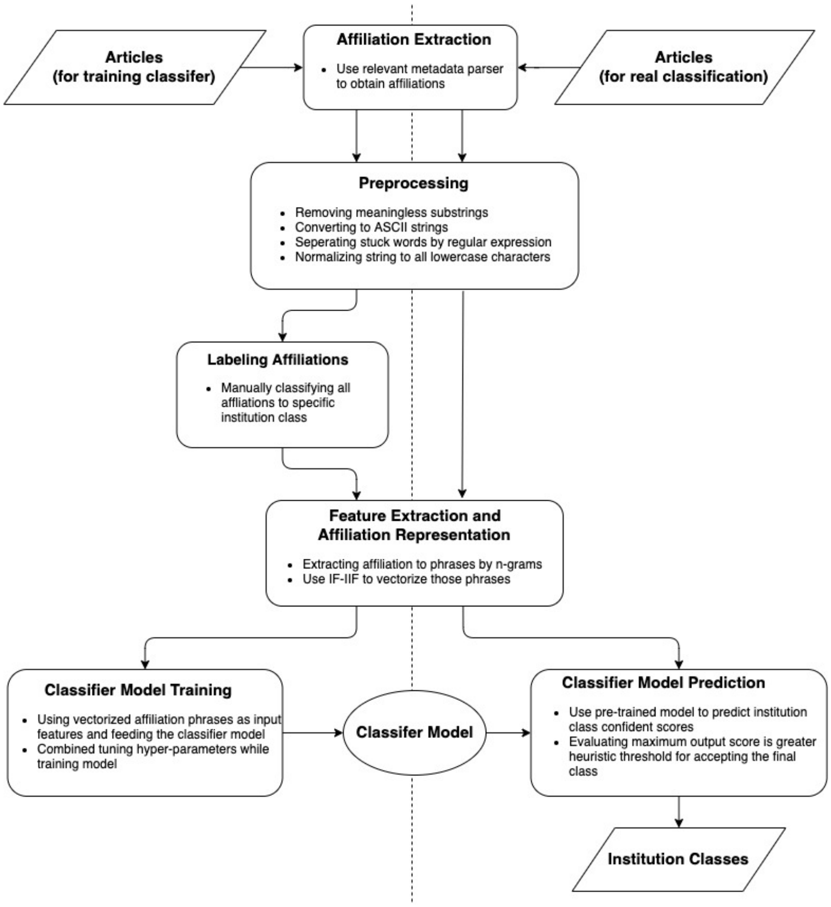
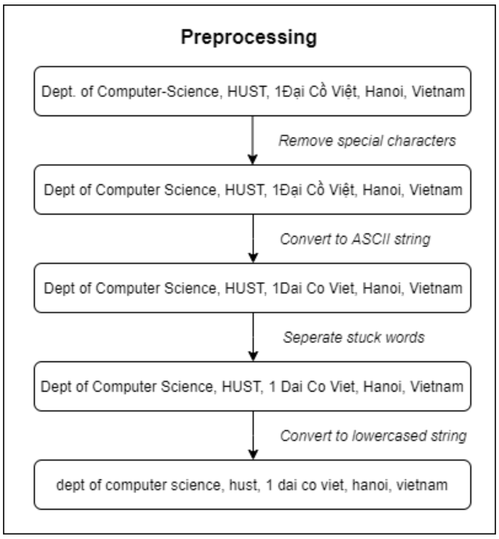

Articles
(for training classifer)
Affiliation Extraction
Use relevant metadata parser to obtain affiliations
Articles
(for real classification)
Preprocessing
Removing meaningless substrings
Converting to ASCII strings
Seperating stuck words by regular expression
Normalizing string to all lowercase characters
Labeling Affiliations
Manually classifying all affliations to specific institution class
Feature Extraction and Affiliation Representation
Extracting affiliation to phrases by n-grams
Use IF-IIF to vectorize those phrases
Classifier Model Training
Using vectorized affiliation phrases as input features and feeding the classifier model
Combined tuning hyper-parameters while training model
Classifer Model
Classifier Model Prediction
Use pre-trained model to predict institution class confident scores
Evaluating maximum output score is greater heuristic threshold for accepting the final class
Institution Classes

Figure 2. The proposed method to detect institutions of articles


Preprocessing
Dept. of Computer-Science, HUST, 1Đại Cồ Việt, Hanoi, Vietnam
Remove special characters
Dept of Computer Science, HUST, 1Đại Cồ Việt, Hanoi, Vietnam
Convert to ASCII string
Dept of Computer Science, HUST, 1Dai Co Viet, Hanoi, Vietnam
Seperate stuck words
Dept of Computer Science, HUST, 1 Dai Co Viet, Hanoi, Vietnam
Convert to lowercased string
dept of computer science, hust, 1 dai co viet, hanoi, vietnam

Hinh 3. An example of the preprocessing steps.

ranging from 1 to 3. Extracted words and phrases are then considered as features for affilation representation. To make a better representation, phrases containing commas are not taken in account. For example, with the affiliation string "Vietnam National University, Hanoi", $2 - grams$ based phrases are "Vietnam National", and "National Unviversity". The phrase "Unviversity, Hanoi" is considered as meaningless and is ignored.

When transforming affiliation strings into the new feature space, we try to capture both local and global characteristics. With the local characteristic of an affiliation string s, we estimate how 'important' extracted words or phrases contribute to s.

Meanwhile, with the global characteristic, we may obtain the contribution/important of extracted words or phrases to the institution in the set of institutions.

The local characteristic is quantified by frequency of the word or phrase appearing in an affiliation string. The importance of a word or a phrase is proportional to the frequency of the word or the phrase. The higher the frequency of the word (phrase) is, the more the importance of the word (phrase) to the institution. The local characteristic is determined by IF:

$$IF(t,s) = 1 + log(freq(t,s)) \qquad (1)$$

where t is a feature representing a word or a phrase. $freq(t,s)$ is frequency of t in s.

The global characteristic is evaluated by the inverse institution frequency (IIF) of the word or the phrase. This characteristic shows how common or rare a word or a phrase is in all institutions. The closer it is to 1.0, the more common a word is. This metric can be calculated by taking the total number of institutions, dividing it by the number of institutions that contain a word or a phrase. The formulation for global characteristics is showed as follows.

$$IIF(t,C) = \frac{|C|}{|C_t|} \qquad (2)$$

where C denotes a set of institutions and C_t is the set of institutions containg t.

We see that an affiliation string is represented by a feature vector receiving values that can capture local and global characteristics of words and phrases decomposed from the affiliation string. These feature values are obtained as follows.

$$IF - IIF(t,s,C) = IF(t,s) * IIF(t,C) \qquad (3)$$

Table 1 shows words or phrases with high IF-IIF for three institutions including Vietnam National University in Hanoi, Vietnam Academy of Science and Technology, and Ton Duc Thang University. The results show that important words or phrases of the affiliation strings have high IF-IIF values. Therefore, these words or phrases can be useful to represent the corresponding institution and we can utilize to classify institutions.

2.3. A SVM model for affiliation string classification

To learn a predictive model, in our approach, we use support vector classifier (SVC)[16]. In addition, the Radial Basic Function (RBF) kernel is used to map data to higher-dimension space before learning the classifier f_k of class k.

$$f_k(x) = \Sigma_{i=1}^{n} w_{k,i} * \Phi(x, x_i) + w_{k0} \qquad (4)$$

where w_k is the weight vector and $\Phi(x, x')$ is the RBF function defined as follows.

$$\Phi(x, x') = exp(-\gamma * ||x - x'||^2) \qquad (5)$$

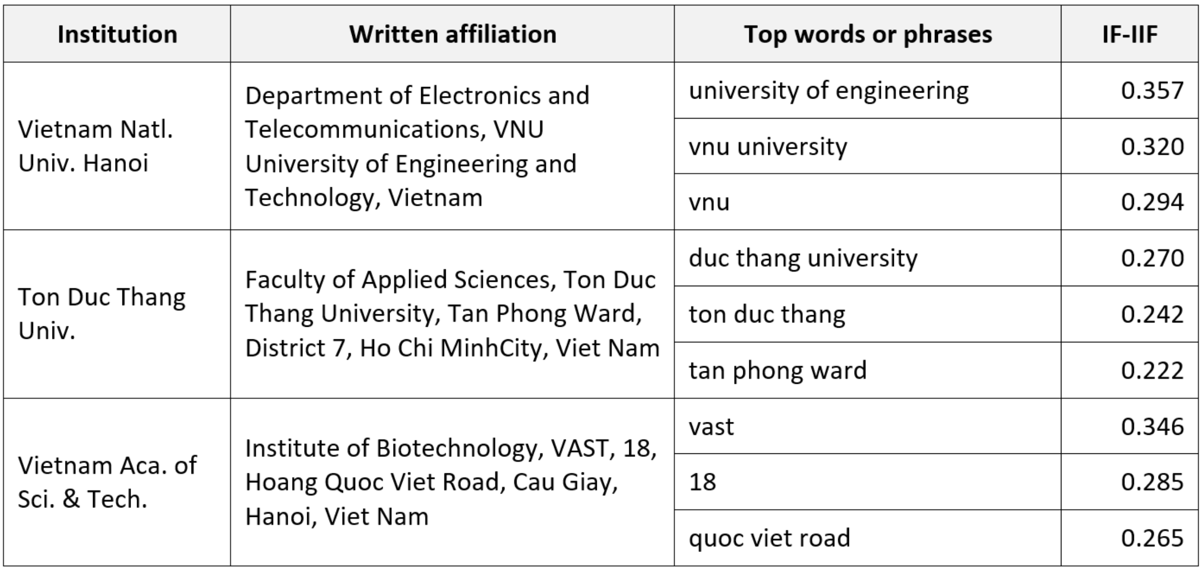
The training step optimises a convex cost function. The probability that an affiliation string x is classified to an institution k is formulated as follows.

$$p(k|x) = \frac{1}{1 + e^{A * f_k(x) + B}} \qquad (6)$$

where A and B are estimated by minimizing the negative log likelihood of training data (using their labels and decision values).

The approach has many benefits. First, the model only depends on the most informative patterns (the support vectors). Second, the learning process is not complicated because there are no false local minima.

Table 1. Examples of IF-IIF of words and phrases



Institution	Written affiliation	Top words or phrases	IF-IIF
Vietnam Natl. Univ. Hanoi	Department of Electronics and Telecommunications, VNU University of Engineering and Technology, Vietnam	university of engineering	0.357
		vnu university	0.320
		vnu	0.294
Ton Duc Thang Univ.	Faculty of Applied Sciences, Ton Duc Thang University, Tan Phong Ward, District 7, Ho Chi MinhCity, Viet Nam	duc thang university	0.270
		ton duc thang	0.242
		tan phong ward	0.222
Vietnam Aca. of Sci. & Tech.	Institute of Biotechnology, VAST, 18, Hoang Quoc Viet Road, Cau Giay, Hanoi, Viet Nam	vast	0.346
		18	0.285
		quoc viet road	0.265

After learning the model using SVC with RBF kernel, we set the threshold 0.6 in classifying affiliation strings to institutions. In equation (6), x is classified as k only if $p(k|x) \geq 0.6$.

3. Experimental Evaluation

This section presents the experimental result of our method on a data set of affiliations collected from Scopus. For the dataset, first, we obtain metadata of articles belonging to at least one Vietnamese institution and published in both 2016 and 2017. After that we extract affiliation strings of Vietnamese institutions. The data set consists of 12704 affiliation strings labeled to 36 classes. 35 classes represent 35 predetermined institutions and one class (OTHER) is for other institutions. Figure 4 shows the distribution of affiliation strings in each institution. It can be seen that the data set is unbalanced.

The data set of affiliations is preprocessed by the steps mentioned above. Features represented by Bag of Words and 1-3 grams are weighted by using IIF function. The feature space has 24383 dimensions. The data set divided into training data set and testing data set with 18605 affiliation strings and 4652 affiliation strings, respectively. In the training step, 5-folds cross validation is used to obtain a fit model. In addition, we tried to tune the hyper-parameters of SVC model with 4 different kernels including Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid. The parameter γ is experimented from 10^{-5} to 10^{-2} while the parameter C, the penalty for misclassifying a data point, changes from 10^{-3} to 10^{3}. Finally, we decided on the SVC model with RBF kernel, 10^{-2} for γ and 10^{2} for C.

The testing data set is used to measure the performance of our model and other models based on other well-known classification methods including Random Forest (RF) [17], Logistic Regression (LR), and K-Nearest Neighbor (KNN)[18]. The

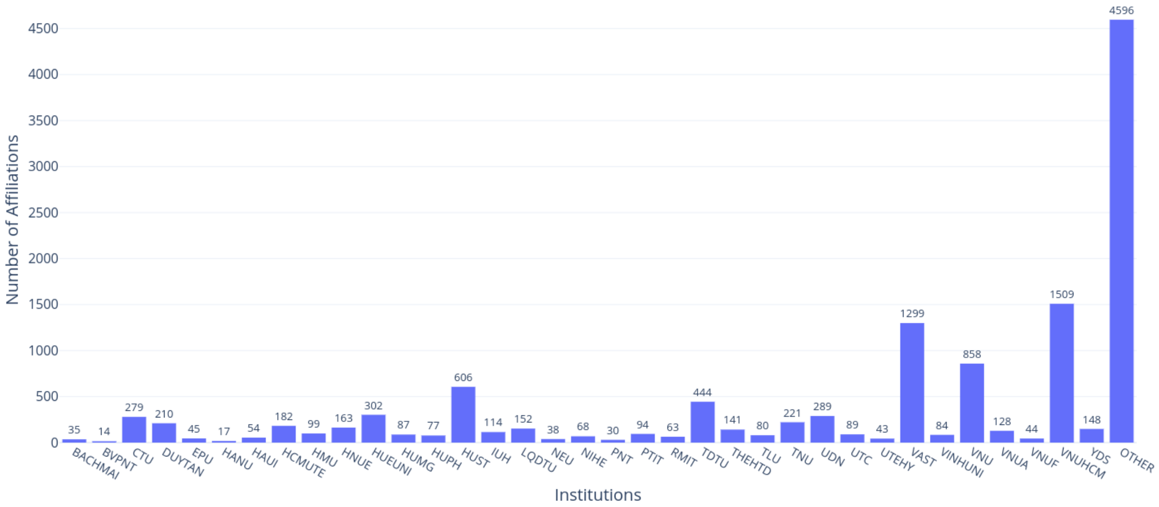


Figure 4. The number of affiliation strings of each institution.

Table 2. Accuracies of models

Model	Precision	Recall	F1
RF	0.6993	0.7665	0.7152
LR	0.9589	0.9595	0.9591
KNN	0.9601	0.9551	0.9575
SVM	**0.9914**	**0.9913**	**0.9913**

results are described in the Table 2.

The experiment result shows that our model outperforms other models. Besides, compared to the model proposed by Pascal Cuxac and his colleagues [19] (trained on their own data set), the accuracy of our model (0.99) is better than that of their model (0.93). The accuracy of our model is very high (approximate 1.0) in three accuracy measures on the testing data set. This result prompts us to apply the model to a practical problem.

We applied our model to verify the mapping articles to institutions in Scopus. From Scopus, we collected metadata of all articles published by at least one Vietnamese institution during the period from 1/2014 to 6/2019. By classifying affiliation strings of each article we can check whether Scopus classifies articles to institutions correctly. Table 3 is the result. The first column indicates institutions. The second one is the number of articles published by the corresponding institution. These numbers are from Scopus. The third column is the number of articles of each institution as the result of our approach. The fourth column is the number of articles that Scopus counts for the corresponding institution but our tool decided contrarily. In contrast, the value in the fifth column is the number of articles of the corresponding institution miscounted by Scopus. The number in the parentheses is the result of our manually check. For example, with the Vietnam Academy of Science and Technology, the number of articles recognized by Scopus is 3931. Our tool shows that this number should be 4519. The tool also indicates that 5 articles which not actually belong to this institution but still being counted by Scopus. By checking manually (i.e. looking at the affiliation strings of articles) we confirm that all these 5 articles are wrongly counted by Scopus. Meanwhile, our tool found 593 more articles (in Scopus) that belong to

Table 3. Result of rectifying affiliation information for Vietnamese institutions.

Institution	**Scopus**	**A2I**	**Scopus-A2I** (manually check)	**A2I-Scopus** (manually check)
Ton Duc Thang Univ.	3955	3995	0 (0)	40 (37)
Vietnam Aca. of Sci. & Tech.	3931	4519	5 (5)	593 (592)
Vietnam Natl. Univ. Hanoi	2639	3132	599 (599)	1092 (1092)
Hanoi Univ. of Sci. & Tech.	3052	2530	572 (572)	50 (48)
Vietnam Natl. Univ. HCM	1839	4734	154 (154)	3049 (3038)
Duy Tan Univ.	1789	1789	2 (1)	2 (2)
Hue Univ.	624	923	1 (1)	300 (295)
Hanoi Univ. of Edu.	744	774	1 (1)	31 (31)
Can Tho Univ.	964	941	55 (55)	32 (26)
Univ. of Da Nang	790	868	19 (16)	97 (96)

the institution. The result of the manual check shows that only 592 (out of 593) actually belong to the institution. Our tool fails to detect one article. Regarding Ton Duc Thang University, 3955 papers indicated by Scopus actually belong to this university (i.e. there is no false positive). Our tool hints that 40 articles are miscounted. Although the correct number is 37 (obtained by manual check), our tool shows its effectiveness, especially in finding miscounted articles for Vietnam National University Hanoi and Vietnam National University HCM.

4. Conclusion

In this work, we study the issue of bibliometric databases such as Scopus and Web of Science in identifying authors' institutions. We propose a method for mapping affiliation strings (written in papers) to authors' institutions. Our method exploits only basic techniques in NLP and machine learning. We experimented the method with papers of Vietnamese institutions in Scopus. The experiment result shows the effectiveness of our method and the current approach of mapping papers to institutions of Scopus needs improving.

Acknowledgments

This work has been supported by Vietnam National University, Hanoi (VNU) under project QG.18.62.

References

[1] S. B. Shereen Hanafi, Discover the data behind the times higher education world university rankings, Elsevier Connect.

[2] M. Dobrota, M. Bulajic, L. Bornmann, V. Jeremic, A new approach to the qs university ranking using the composite i-distance indicator: Uncertainty and sensitivity analyses, JASIST 67 (2016) 200–211.

[3] A.-P. Pavel, Global university rankings - a comparative analysis, Procedia Economics and Finance 26 (2015) 54–63. doi:10.1016/S2212-5671(15)00838-2.

[4] Web of science databases, Clarivate Analytics.

[5] J. F. Burnham, Scopus database: a review, Biomedical Digital Libraries 3. doi:https://doi.org/10.1186/1742-5581-3-1.

[6] F. Franceschini, D. Maisano, L. Mastrogiacomo, A novel approach for estimating the omitted-citation rate of bibliometric databases with an application to the field of bibliometrics, JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 64 (2013) 2149–2156. doi:10.1002/asi.22898.

[7] F. Franceschini, D. Maisano, L. Mastrogiacomo, Scientific journal publishers and omitted citations

in bibliometric databases: Any relationship?, Journal of Informetrics 8 (3) (2014) 751 – 765. doi:https://doi.org/10.1016/j.joi.2014.07.003. URL http://www.sciencedirect.com/science/article/pii/S1751157714000637

[8] R. Buchanan, Accuracy of cited references: The role of citation databases, College Research Libraries 67. doi:10.5860/crl.67.4.292.

[9] J. Valderrama-Zurián, R. Aguilar-Moya, D. Melero-Fuentes, R. Aleixandre-Benavent, A systematic analysis of duplicate records in scopus, Journal of Informetrics 9 (2015) 570–576. doi:10.1016/j.joi.2015.05.002.

[10] J. Zhu, G. Hu, W. Liu, Doi errors and possible solutions for web of science, Scientometrics 118 (2) (2019) 709–718. doi:10.1007/s11192-018-2980-7. URL https://doi.org/10.1007/s11192-018-2980-7

[11] S. Xu, L. Hao, X. An, D. Zhai, H. Pang, Types of doi errors of cited references in web of science with a cleaning method, Scientometrics 120 (3) (2019) 1427–1437. doi:10.1007/s11192-019-03162-4. URL https://doi.org/10.1007/s11192-019-03162-4

[12] E. Krauskopf, Missing documents in scopus: the case of the journal enfermeria nefrologica, Scientometrics 119 (1) (2019) 543–547. doi:10.1007/s11192-019-03040-z. URL https://doi.org/10.1007/s11192-019-03040-z

[13] W. Liu, G. Hu, L. Tang, Missing author address information in web of science—an explorative study, Journal of Informetrics 12 (3) (2018) 985 – 997. doi:https://doi.org/10.1016/j.joi.2018.07.008. URL http://www.sciencedirect.com/science/article/pii/S175115771730353X

[14] E. Krauskopf, Standardization of the institutional address, Scientometrics 94 (3) (2013) 1313–1315. doi:10.1007/s11192-012-0852-0. URL http://dx.doi.org/10.1007/s11192-012-0852-0

[15] E. Krauskopf, Call for caution in the use of bibliometric data, J. Assoc. Inf. Sci. Technol. 68 (8) (2017) 2029–2032. doi:10.1002/asi.23809. URL https://doi.org/10.1002/asi.23809

[16] M. Awad, R. Khanna, Support Vector Machines for Classification, Apress, Berkeley, CA, 2015, pp. 39–66. doi:10.1007/978-1-4302-5990-9_3. URL https://doi.org/10.1007/978-1-4302-5990-9_3

[17] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324

[18] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1) (2006) 21–27. doi:10.1109/TIT.1967.1053964. URL https://doi.org/10.1109/TIT.1967.1053964

[19] L. J.-C. B. Cuxac, P., Efficient supervised and semi-supervised approaches for affiliations disambiguation, Scientometrics 97(1) (2013) 47–58.