Original Article

On Rectifying the Mapping between Articles and Institutions in Bibliometric Databases

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

VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Received 04 February 2019
Revised 17 February 2020; Accepted 18 February 2020

Abstract: Today, bibliometric databases are indispensable sources for researchers and research institutions. The main role of these databases is to find research articles and estimate the performance of researchers and institutions. Regarding the evaluation of the research performance of an organization, the 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 efficiently. To this end, we propose an approach to revise the authors’ affiliation information of articles in bibliometric databases. We build a model to classify articles to institutions with high accuracy by assembling the bag of words and n-grams techniques for extracting 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 sequences. 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-iiif, 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,
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 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 the 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, these problem may come from author’s writing mistakes, they may unclearly or incompletely provide their institution information. As a result, WoS or Scopus incorrectly maps the article to authors’ institutions. In addition to the missteps 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 identifying institutions of articles. The proposed tool only exploits basic techniques in Natural 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 so 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 verifying the mapping as a classification problem. We restate the problem as follows. Given a set $S=\{(s_i,y_i)\}_{i \in \{1...n\}}$ 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 a new affiliation string $s$ 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 string extraction, data preprocessing, affiliation string 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 these 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 corresponding 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 transiterate 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. Separate stuck words: By observing affiliation strings, we found that many affiliation strings contain stuck words. Separating 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 \(?<=[a-z]([-]?)?(?=[0-9A-Z])\)?\([a-z]([-]?)\)?\([A-Z]\)[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 strings 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, 1Dai Co Viet, Hanoi, Vietnam”. In the second step, characters of the affiliation string are converted ASCII. Therefore, the string “Dept of Computer Science, HUST, 1Dai Co Viet, 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 ranging from 1 to 3. Extracted words and phrases are then considered as features for affiliation 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 University”. The phrase “University, Hanoi” is considered as meaningless and is ignored.

| 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, 1Đại Co Viet, Hanoi, Vietnam |
| Separate stuck words |
| Dept of Computer Science, HUST, 1Đại Co Viet, Hanoi, Vietnam |
| Convert to lowercase string |
| dept of computer science, hust, 1 dai co viet, hanoi, vietnam |

Figure 3. An example of the preprocessing steps.

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/importance 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, it is assumed that 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:

$$\text{IF}(t, s) = 1 + \log \left( \text{freq}(t, s) \right)$$  \hspace{1cm} (1)

where \(t\) is a feature represents a word or a phrase, \(\text{freq}(t, s)\) is frequency of \(t\) in \(s\).

The global characteristic is evaluated by the inverse institution frequency (IIF) of the word or the phrase. We assume that each institution is a set of words and phrases which are retrieved from prior feature extraction step, the characteristic shows how common a word or a phrase appears in all institutions.

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, Viet Nam | 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 Minh City, 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    |

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 closer it is to 1.0, the more common a word is. The formulation for global characteristics is showed as follows.

$$\text{IIF}(t, C) = \log \left( \frac{|C|}{|C_t|} \right)$$  \hspace{1cm} (2)

where \(C\) denotes a set of institutions and \(C_t\) is the set of institutions containing \(t\).

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

$$\text{IF} - \text{IIF}(t, s, C) = \text{IF}(t, s) \times \text{IIF}(t, C)$$  \hspace{1cm} (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 efficient to represent the corresponding institution and the classifier model can utilize them to predict accurately.

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 \(k\).

$$f_k(x) = \sum_{i=1}^{N} w_{k,i} \Phi(x, x_i) + w_{k,0}$$  \hspace{1cm} (4)

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

$$\Phi(x, x') = \exp(-\gamma \|x - x'\|^2)$$  \hspace{1cm} (5)
The training step optimizes a convex cost function. The probability that an affiliation string $x$ is classified to an institution $k$ is formulated as follows.

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

(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.

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

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

3. Experimental Evaluation

This section presents the experimental result of our method on a data set of affiliations collected from Scopus. About the dataset, we firstly obtain metadata of articles published in both 2016 and 2017 that belongs to at least one Vietnamese institution. 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 IF-IIF function. The feature space has 24383 dimensions. The data set is then split into training data set and testing data set by 80/20 ratio with 10163 affiliation strings and 2541 affiliation strings, respectively. In the training step, 5-fold 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 $\gamma$ 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 $\gamma$ 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 results are described in the Table 2.

Table 2. Accuracy of models

| Model | Precision | Recall | Macro-F1 |
|-------|-----------|--------|----------|
| RF    | 0.6693    | 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   |

The experiment result showed that our SVM model outperforms other models. However, the distribution in Figure 4 produced that the sizes of samples set between classes were imbalanced, especially number of the “OTHER” samples was significant compared to the rest, which may lead to inaccurate evaluation. Accordingly, we further assessed each label accuracy instead of bringing all together. The empirical result revealed that the F1-score of each label ranges between 0.96 and 1.00, and mostly at the highest score 1.00. These statistics rationally pointed out that there were no class had lagged in F1-score, all of them had the values very close to the Macro-F1 after measuring overall result on multiple classes. Besides, comparing to the model proposed by Pascal Cuxac and his colleagues [19] (trained on their own data set), the Macro-F1 score 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 also 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 mapped them to institutions correctly. The result is shown in table 3. The first column indicates institutions. The second one is the number of articles published by the corresponding institution, which purely obtained as cardinality of Scopus’s article set. The third column is the number of articles of each institution as the result of A2I tool based on our approach. The fourth column is the number of articles that Scopus counts for the corresponding institution but our tool decided contrarily, this one is calculated by the cardinality of set difference of Scopus and A2I sets. In contrast, the values in the fifth column is the number of articles of the corresponding institution miscounted by Scopus but were found by A2I. The number in the parentheses is the result after checking manually each difference set, represents number of remaining articles is correctly assigned to the institution. 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 were wrongly counted by Scopus. Meanwhile, our tool found 593 more articles (in Scopus) that belong to the institution. The result of the manual check shows that only 592 (out of 593) actually belong to the institution. Our tool failed 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. Conclusions

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. An implication of the result is that 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. https://doi.org/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. http://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. https://doi.org/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. https://doi.org/10.1016/j.joi.2014.07.003.
[8] R. Buchanan, Accuracy of cited references: The role of citation databases, College Research Libraries 67. http://doi.org/10.5860/crl.67.4.292.
[9] J. Valderrama-Zarián, R. Aguilar-Moya, D. Melero-Fuentes, R. Alexiandre-Benavent, A systematic analysis of duplicate records in scopus, Journal of Informetrics 9 (2015) 570–576. http://doi.org/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. http://doi.org/10.1007/s11192-018-2980-7.

Table 3. The number of affiliation strings of each institution

| Institution                        | [Scopus] | [A2I] | [Scopus\A2I] (manual check) | [A2I\Scopus] (manual check) |
|------------------------------------|----------|-------|----------------------------|-----------------------------|
| Ton Duc Thang Univ.                | 3955     | 3995  | 0 (0)                      | 40 (37)                     |
| Vietnam Aca. of Sci. & Tech.       | 3931     | 4519  | 5 (0)                      | 593 (592)                   |
| Vietnam Natl. Univ. Hanoi          | 2639     | 3132  | 599 (0)                    | 1092 (1092)                 |
| Hanoi Univ. of Sci. & Tech.        | 3052     | 2530  | 572 (0)                    | 50 (48)                     |
| Vietnam Natl. Univ. HCM            | 1839     | 4734  | 154 (0)                    | 3049 (3038)                 |
| Duy Tan Univ.                      | 1789     | 1789  | 2 (1)                      | 2 (2)                       |
| Hue Univ.                          | 624      | 923   | 1 (0)                      | 300 (295)                   |
| Hanoi Univ. of Edu.                | 744      | 774   | 1 (0)                      | 31 (31)                     |
| Can Tho Univ.                      | 964      | 941   | 55 (0)                     | 32 (26)                     |
| Univ. of Da Nang                    | 790      | 868   | 19 (3)                     | 97 (96)                     |
[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. http://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. 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. https://doi.org/10.1016/j.joi.2018.07.008.

[14] E. Krauskopf, Standardization of the institutional address, Scientometrics 94(3) (2013) 1313-1315. http://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. http://doi.org/10.1002/asi.23809.

[16] M. Awad, R. Khanna, Support Vector Machines for Classification, Apress, Berkeley, CA, 2015, pp. 39-66. http://doi.org/10.1007/978-1-4302-5990-9-3.

[17] L. Breiman, Random forests, Machine Learning 45(1) (2001) 5-32. 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. http://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.