Storage and Allocation of English Teaching Resources Based on k-Nearest Neighbor Algorithm

https://doi.org/10.3991/ijet.v14i17.11188

Yi Lou
Zhengzhou Preschool Education College, Zhengzhou, China
louyiweoe@163.com

Abstract—The boom of Internet technology gives a boost to the informatization of education in China. Internet resources serve as a new carrier of knowledge, offering teachers and students an alternative to books. However, the exponential growth of Internet resources has greatly complicated the storage and allocation of resources. This paper attempts to fully utilize English teaching resources through effective resource management and allocation. Specifically, the features of English teaching resources were analyzed, and then the term frequency-inverse document frequency (TF-IDF) weight method and k-nearest neighbor (kNN) algorithm were improved to make resource allocation more efficient and effective. The improved methods were then verified through a case analysis. The results show that the improved kNN provides a feasible way to allocate English teaching resources. The research findings provide reference to the storage and allocation of teaching resources.

Keywords—Teaching resources, k-nearest neighbor (kNN) algorithm, erm frequency-inverse document frequency (TF-IDF) weight, storage and allocation

1 Introduction

In recent years, the boom of network technology has ushered in the information age [1-2]. In the meanwhile, with the improvement of people's living standards, the education industry has aroused an unprecedented concern of the whole society. Computerization has also been prevalent in various universities [3]. The surge of Internet technology has spawned digital education to make sure that knowledge transmission will be no longer limited to books. However, no matter where you are, as long as the Internet exists, we can accept good quality education whenever and wherever possible [4]. According to data statistics, more than 80 % of schools across the country have accessed the Internet [5], and most of them also have independent multimedia education resources. There is also the surge of in-school education platforms, greatly enriching the ways students receive education [6]. However, there are so many online education resources that a great challenge educational scholars face now is how to store, distribute and manage the complex education resources in the network [7-8].

In the past, the schools usually classified educational resources manually since the Internet was not popular. Only those with professional knowledge were allowed to operate storage and distribution of teaching resources [9]. This classification method can yield accurate results only when there are less resources [10]. Manual allocation is inefficient and inaccurate as the number of resources continues to increase [11]. In this case, how to effectively store and classify the teaching resources in the network is a staggering problem that needs to be solved urgently [12].

For this purpose, in view of the fact that the general method is used for classification and storage of English teaching resources [13], this paper outspreads the study on the storage and classification algorithm based on the properties of English teaching resources, but not combine the practices of it [14]. Based on the illustration of resource classification process, this paper analyzes how to choose feature resources, how to determine the weight based on TF-IDF and how to evaluate resultant resource allocation based on value F1. A teaching resource storage and allocation strategy is also proposed based on improved KNN algorithm. The findings show that the improved KNN algorithm is feasible for storage and allocation of English teaching resources; it has high operation efficiency and reliable classification results.

2 Basic Theory for Resource Storage and Allocation

It is usually required to pre-process resources for storage and allocation, select feature items that can represent resource properties to perform weight calculation, and select and train resource allocator to achieve the purposes of classifying, outputting, and evaluating new resources. The process of resource storage and allocation is shown in Figure 1.

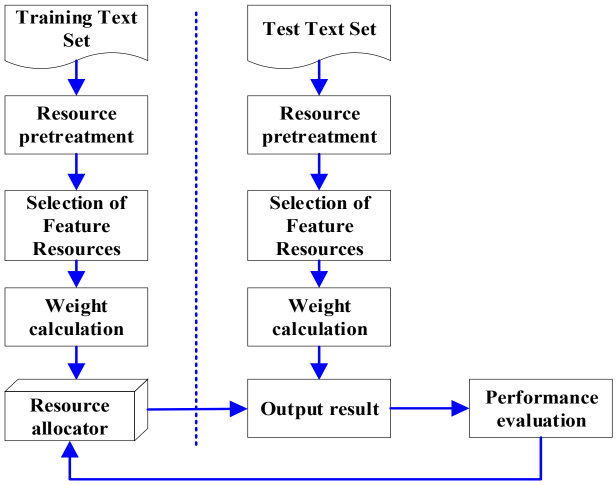

Fig. 1. Flow chart of resource classification

2.1 Selection of unique resources

To filter the resources with strong properties from the mass of raw data for storage to reduce the dimensionality, the frequency of resources, mutual information, information increase and chi-square statistics can be used, expressed using the DF、MI、IG and x^2 statistics. Assume that there are n types of resources in the corpus, W_j represents j type of resources; TZ_i represents i type of feature terms.

DF represents a certain type of resources as percentage of the total. It can be calculated by Formula (1).

$$DF(T_i) = \frac{F(T_i)}{N} \tag{1}$$

Where, $F(T)_i$ is the number of resources containing feature item TZ_i.

In actual operation, the threshold *M* of resources reserved is usually set in advance. When the frequency of a certain type of stable resources is greater than *M*, it is reserved; otherwise, it is scrapped. Although this method is operated simply, it is likely to accidentally remove resources containing important features, so is rarely used alone in practice.

MI refers to the extent to which the resource text is correlated to the feature item, as shown in Formula (2).

$$MI_i(TZ_i, W_j) = P(W_j)\sum\nolimits_{i=1}^{N} \log\frac{P(TZ_i \mid W_j)}{P(TZ_i)} \tag{2}$$

Where, $\mathrm{P}(\mathrm{TZ}_i|W_j)$ represents the frequency that TZ_i appears in W_j; $\mathrm{P}(W_j)$ represents the frequency of resources W_j as percentage of the total resources; $\mathrm{P}(TZ_i)$ represents the frequency of resources that contain feature items.

IG represents the importance of the feature item in the corpus by calculating the information gap of the corpus in the case when it owns or does not own a feature item, as shown in Formula (3).

$$IG(TZ_i) = -\sum_{j=1}^{m} P(W_j)\log_2 P(W_j) + P(TZ)\sum_{j=1}^{m} P(W_j \mid TZ_i)\log_2 P(W_j \mid TZ_i)$$
$$+ P(\overline{TZ_i})\sum_{j=1}^{m} P(W_j \mid \overline{TZ_i})\log_2 P(W_j \mid \overline{TZ_i}) \tag{3}$$

Where, $\mathrm{P}(W_j|\mathrm{TZ}_i)$ represents the probability that W_j contains TZ_i; $\mathrm{P}(W_j|\overline{\mathrm{TZ}_\iota})$ represents the probability that W_j does not contain TZ_i.

In practice, although information increase is an effective method, there are still phenomena that may interfere with the allocation results.

x^2 statistics evaluate the extent to which both are correlated to each other by calculating the chi-square values of TZ_i and W_j, as shown in Formula (4).

$$x^2(TZ_i, W_j) = \frac{N(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)} \tag{4}$$

Where, A represents the number of resources that belong to $W_j$and contains TZ_i; B represents the number of resources that belong to W_j but do not contain TZ_i; C represents the number of resources that do not belong to W_j but contain TZ_i; D represents the number of resources that neither belong to W_j nor contain TZ_i.

Since the chi-square statistics can well reduce the dimension of corpus, here uses this method as the option of feature items.

2.2 Weight determination

After selecting the feature resources, in order to differentiate contribution values of resources with different features to the total corpus, the weight should be assigned to resources with different features. The term frequency-inverse document frequency, referred to as TF-IDF, is introduced.

The core idea of the method is to recognize the unique resource that appears at a higher frequency in a certain type of resources but does at a lower frequency in other resources, and given a higher weight, as shown in Formulas (5), (6), and (7).

$$TF_{ik} = \frac{N_{ik}}{\sum_{j=1}^{n} N_{ij}} \tag{5}$$

$$IWF_k = \log \frac{N}{N_k} \tag{6}$$

$$w_{ik} = TF_{ik} * IWF_k \tag{7}$$

Where, TF_{ik} represents the frequency at which feature item TZ_k appears in a resource W_i ; IWF_k represents the frequency of TZ_k in the reverse resource in W_i; N_{ik} represents the number of occurrences of TZ_k in W_i; N_k represents the number of resources containing TZ_k.

Although TF-IDF has a better effect than other weight algorithms, there are still gap that it has not considered whether the feature items TZ_k are evenly distributed in all resources or concentrate in one resource. Therefore, Formula (7) can be improved as follows:

$$w_{ik} = TF_{ik} * IWF_k * \left(\frac{TF_{ik} N_{ik}}{TF_{ik} N_i} * \frac{\sum_{j=1}^{N_{ik}} (TF_{jk} - \overline{TF_i})^2}{N_{ik}} \right) \tag{8}$$

Formula (8) can calculate the weight TZ_k more accurately by remeasuring the distribution of $\mathrm{T}Z_k$ across the resources, in order to improve the allocation result.

2.3 Resource allocation evaluation index

To evaluate the allocation of teaching resources, the precision rate (as shown in Formula (9)), recall rate (as shown in Formula (10)), F-metric value (as shown in Formula (11)), macro value F1 (as shown in Formula (12)) and other indicators are used, of which the last two are more common.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

Where, TP, TP represent the numbers of correctly and incorrectly allocated resources, respectively; FN represents the number of texts that belongs but are not allocated to a type of resources.

$$F1 = \frac{2 * P * R}{P + R} \tag{11}$$

$$MacroF1 = \frac{\sum_{k=1}^{n} F1_k}{k} \tag{12}$$

Where, the macro value F1 is the average of all F-metric values.

3 Resource Allocation Based on KNN Algorithm

3.1 Algorithm introduction

The core idea of Proximity algorithm ($\mathrm{K-NearestNeighbor}$), referred to as KNN algorithm, is [15]: In the feature space, if the majority of the *K* samples most adjacent to one sample belong to a certain type, the test sample also belongs to it and has attributes as this type of samples has [16].

As shown in Figure 2, the training set has two classes, i.e. red five-pointed star and green square, and blue circle is a sample to be classified. When $\mathrm{K} = 3$, there are two red and one green samples among the three closest to the blue one, the sample is classified into red class. When $\mathrm{K} = 5$, there are three red and two green ones among the 5 samples closest to the blue one, the sample is classified into green class. It is obvious that the results are subject to different value *K*.

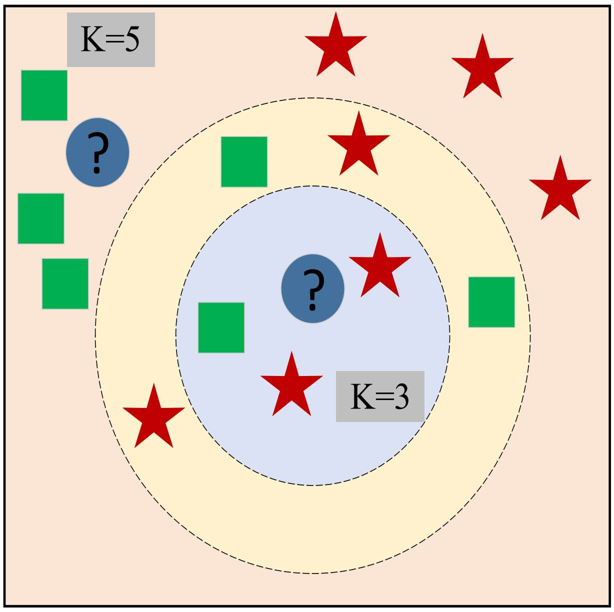

Fig. 2. Sketch map of the KNN algorithm

In general, the Euclidean distance method (Formula (13) and the included angle cosine method (Formula (14) can be used to calculate the distance between two resources.

$$D(d_i,d_j)=\sqrt{\sum_{k=1}^{n}\left(w_{ik}-w_{jk}\right)^2} \tag{13}$$

$$sim(d_i,d_j)=\frac{\sum_{k=1}^{n}w_{ik}*w_{jk}}{\sqrt{\sum_{k=1}^{n}w_{ik}{}^2\sum_{k=1}^{n}w_{jk}{}^2}} \tag{14}$$

Where, w_{ik}, w_{jk} represent the weights of the feature items the texts d_i, d_j correspond to.

3.2 Implementation of algorithm

As there are many training sets in the teaching resources, the computation load of the KNN algorithm is larges. There is also serious uneven distribution between different types of resources. In this case, the traditional KNN algorithm should be improved. The computation load of KNN algorithm can usually be reduced by two ways, one is to optimize the algorithm itself, and the other is to optimize the training set.

Here is the second way, that is, a density tailoring strategy is used for the original training samples to minimize the dimensions of the training set and further reduce the time complexity of the algorithm. The KNN algorithm can be improved by the following procedures, as shown in Figure 3.

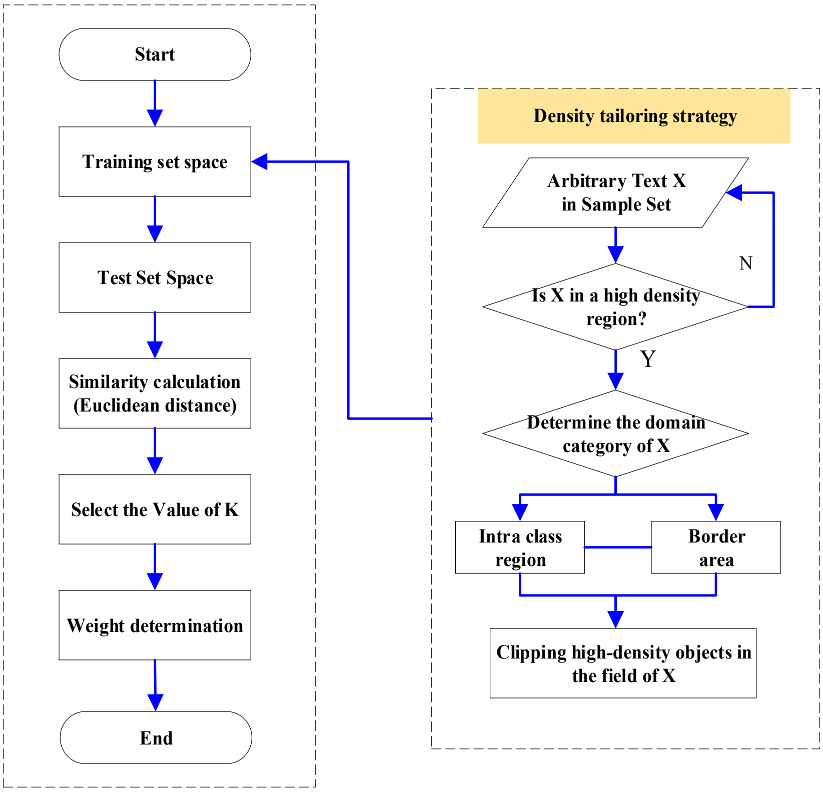

Fig. 3. Basic procedure for improving KNN algorithm

Step 1: Express the training and the test sets with a vector space model;

Step 2: Calculate the spatial distance between the resource to be allocated in the test set and the resource available in the training set by Formula (13) or (14);

Step 3: determine value K, and select k training set resources most similar to the resources to be classified;

Step 4: determine weight, and count and classify the resources to be allocated in the test set.

For the key parameters in the improved algorithm, Hpts usually takes 6 %-8 % of the average number of samples. Minpts usually takes the integer between $[0, \text{Hpts}]$; ε is determined by the Formula (15).

$$\varepsilon = \frac{1}{D}\sum_{i}^{|D|} Dist_{Hpts}(X_i) \tag{15}$$

Where, $Dist_{Hpts}(X_i)$ is the closest distance between sample k and sample (X_i).

4 Case Analysis

The constructed corpus contains 5,000 types of teaching resources, including 3,600 resources in the training set and 1,400 resources in the test set. *K* is changed from 5 to 40, the change trend of the macro value F1 is shown in Figure 4.

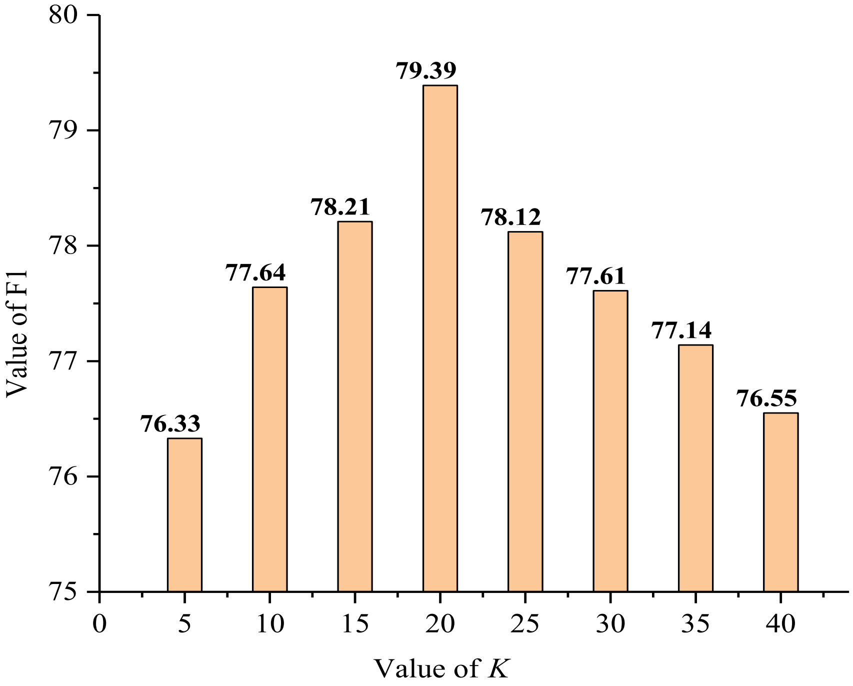

Fig. 4. Macro value F1 as a function of value *k*

As shown in Figure. 4, when K changes from 5 to 20, the macro value F1 gradually increases, and when K increases from 20 to 40, the macro value F1 gradually decreases. Therefore, when value K takes 20, the classification effect comes the best.

Further, the allocation effects of teaching resources before and after KNN is improved are compared. As described above, Hpts takes 10 and Minpts takes 1~10. The training samples are tailored using a density tailoring strategy. The results are shown in Table 1.

Table 1. Tailoring of training set

Minpts	Number of clippings	Cutting ratio
1	156	0.05
2	169	0.06
3	172	0.06
4	272	0.09
5	321	0.11
6	365	0.12
7	368	0.12
8	371	0.12
9	382	0.13
10	424	0.14

Tailoring scales correspond to respective effects shown in Figure 5.

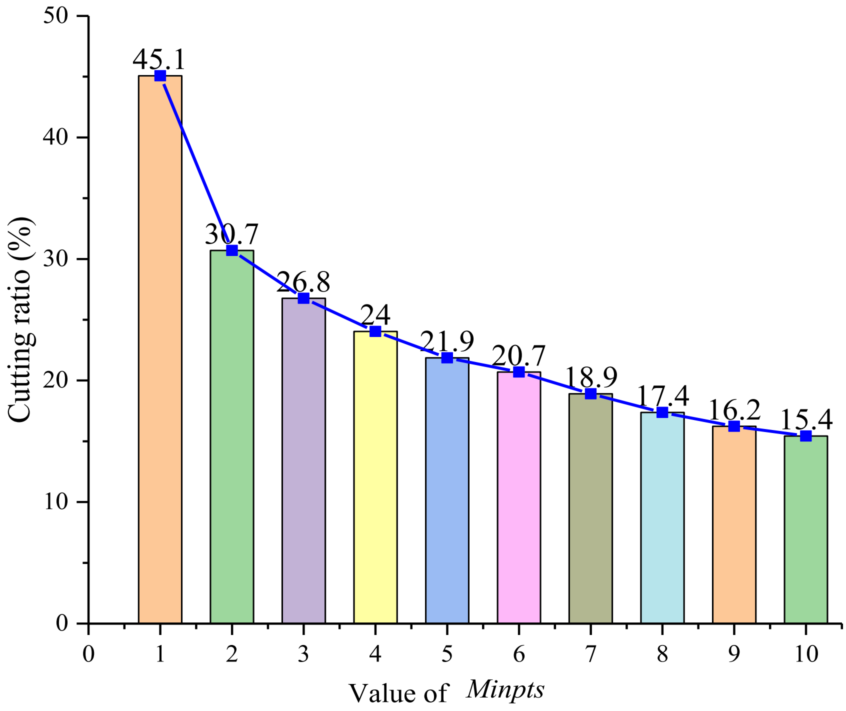

Fig. 5. Curve of tailoring scale as a function of values Minpts

As shown in Table 6-4, the tailoring scales decreases as the value Minpts increases. The classified macro values F1 are compared, as shown in Figure 6.

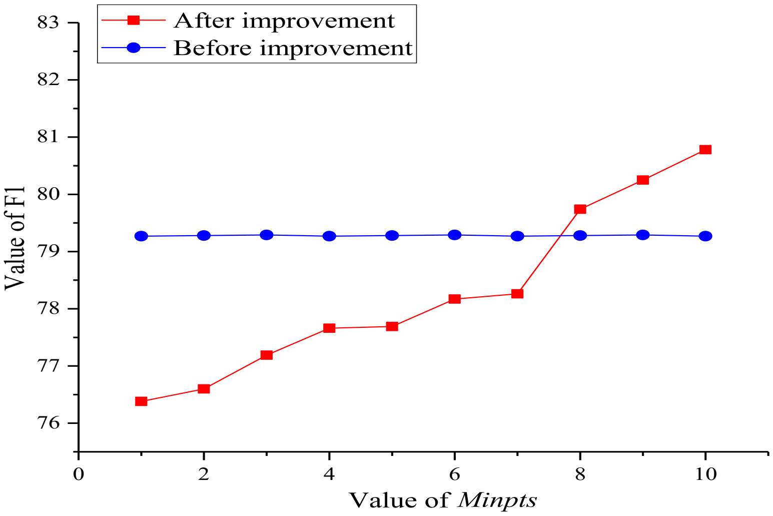

Fig. 6. Curve of value Minpts before and after algorithm is improved

As shown in Figure 6, the improved KNN algorithm can achieve better allocation of teaching resources when Minpts takes 8~10. It is further explained that the improved algorithm is effective, not only can save the time of resource allocation, but also improve the effect of resource allocation to a certain extent.

5 Conclusion

This paper discusses the storage and allocation strategies for English teaching resources. Main conclusions are drawn as follows:

- The basic theories involved in the process of resource allocation are deeply analyzed, and feature extraction, weight calculation and evaluation on allocation effect are expounded;
- The KNN algorithm has been improved. Aiming at the gaps of traditional KNN algorithm in teaching resources, an improved KNN algorithm is proposed based on the density tailoring strategy.
- Application case analysis compares different values K and pre- and post-tailoring values Minpts . It is proved that the improved KNN algorithm can effectively reduce the time complexity of the algorithm, and the resource allocation effect is also improved.

6 Acknowledgement

Department-Level Project Name: The Research of the Construction and Practice of Specialty Course in English Education Major.

7 References

[1] Kalochristianakis, M. N., Paraskevas, M., Varvarigos, E. A., Xypolitos, N. (2007). The greek school network: a paradigm of successful educational services based on the dynamics of open-source technology. IEEE Transactions on Education, 50(4): 321-330. https://doi.org/10.1109/te.2007.904574

[2] Geissbuhler, A., Bagayoko, C. O., Ly, O. (2007). The raft network: 5 years of distance continuing medical education and tele-consultations over the internet in french-speaking africa. International Journal of Medical Informatics, 76(5-6): 351-356. https://doi.org/10.1016/j.ijmedinf.2007.01.012

[3] Abler, R. T., Contis, D., Grizzard, J. B., Owen, H. L. (2006). Georgia tech information security center hands-on network security laboratory. IEEE Transactions on Education, 49(1): 82-87. https://doi.org/10.1109/te.2005.858403

[4] Kalochristianakis, M. N., Paraskevas, M., Varvarigos, E. (2008). Asynchronous tele-education and computer-enhanced learning services in the greek school network. Lecture Notes in Computer Science, 5288: 234-242. https://doi.org/10.1007/978-3-540-87781-3_26

[5] Paris, D. C. (2018). More evolution than revolution? the work and workplaces in higher education. Change The Magazine of Higher Learning, 50(3-4): 44-45. https://doi.org/10.1080/00091383.2018.1509588

[6] Snow, C., Pullen, J. M., Mcandrews, P. (2005). Network educationware: an open-source web-based system for synchronous distance education. IEEE Transactions on Education, 48(4): 705-712. https://doi.org/10.1109/te.2005.854577

[7] Marlino, M., Sumner, T., Fulker, D., Manduca, C., Mogk, D. (2001). The digital library for earth system education: building community, building the library. Communications of the ACM, 44(5): 80-81. https://doi.org/10.1145/374308.374356

[8] Li, S. G., Liu, Q. (2003). Interactive groundwater (igw): an innovative digital laboratory for groundwater education and research. Computer Applications in Engineering Education, 11(4): 179-202. https://doi.org/10.1002/cae.10052

[9] Greenberger, M., Aronofsky, J., Mckenney, J. L., Massy, W. F. (1973). Computer and information networks: the movement in research and education toward national resource sharing via networks is accelerating. Science, 182(4107): 29-35. https://doi.org/10.1126/science.182.4107.29

[10] Allegra, E., Pietro, R.D., La Noce, M., Ruocco, V., Verde, N. V. (2012). Cross-border co-operation and education in digital investigations: a european perspective. Digital Investigation, 8(2): 106-113. https://doi.org/10.1016/j.diin.2011.09.001

[11] Mash, B., Marais, D., Van, D. W. S., Van, D. I., Steyn, M., Labadarios, D. (2006). Assessment of the quality of interaction in distance learning programmes utilizing the internet or interactive television: perceptions of students and lecturers. Medical Teacher, 28(1): e1-e9. https://doi.org/10.1080/01421590600568439

[12] Franco-Lopez, H., Ek, A. R., Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sensing of Environment, 77(3): 251-274. https://doi.org/10.1016/s0034-4257(01)00209-7

[13] Chou, K. C., Shen, H. B. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic k-nearest neighbor classifiers. Journal of Proteome Research, 5(8): 1888-1897. https://doi.org/10.1021/pr060167c

[14] Shen, M., Xiao, Y., Golbraikh, A., Gombar, V. K., Tropsha, A. (2003). Development and validation of\r, k\r, -nearest-neighbor qspr models of metabolic stability of drug candi-

dates. Journal of Medicinal Chemistry, 46(14): 3013-3020. https://doi.org/10.1021/jm020491t

[15] Maltamo, M., Kangas, A. (1998). Methods based on k -nearest neighbor regression in the prediction of basal area diameter distribution. Canadian Journal of Forest Research, 28(8): 1107-1115. https://doi.org/10.1139/x98-085

[16] Wang, J., Neskovic, P., Cooper, L. N. (2006). Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. Pattern Recognition, 39(3): 417-423. https://doi.org/10.1016/j.patcog.2005.08.009

8 Author

Yi Lou, female, was born on August 23,1967, majored in English language and literature. She is currently working as an English teacher in Zhengzhou Pre-school Education College, Henna, China. She has been teaching English courses for 31 years, including College English course and English Education course. Currently, she is serving as the Director of English Teaching and Research. She has published five books and several papers on English teaching. In 2013, she is excellent in annual assessment. In 2016, she was rated as an expert of professional education in Henan Province.

Article submitted 2019-06-13. Resubmitted 2019-07-17. Final acceptance 2019-07-18. Final version published as submitted by the authors.