A Plagiarism Detection Method Based on Learning Behavior Analysis

Wen-jun TANG¹, Du ZOU²,* and Ling ZHANG³

¹South China University of Technology, Communication & Computer Network Lab of GD, Guangzhou, China
²South China University of Technology, Information and Network Engineering and Research Center, Guangzhou, China
³South China University of Technology, Communication & Computer Network Lab of GD, Guangzhou, China

*Corresponding author

Keywords: Learning behavior analysis, Plagiarism detection, Similarity, Clustering, E-learning.

Abstract. Plagiarist detection is a complicate topic in plagiarism detection area. Most existing algorithms can only compare similarity between assignments, but cannot detect plagiarist. This paper designed a homework assessment model based on the South China University of Technology e-learning plagiarism detection module. By analyzing students' learning behavior data collected from the teaching platform, a ranking of the possibilities of plagiarism in students' work is obtained, which provides the basis for judging the plagiarist. And comparing the determination of plagiarism with the actual investigation results, the accuracy of plagiarist determination using the model is significantly improved. The model has been used on the e-learning platform, which provided an effective way for teachers to evaluate assignments.

Introduction

Modern teaching relies more and more on the network. Online learning system changes the way students learn as well as teachers' teaching mode. The instructional interaction, including lecture note sharing, homework assignment and discussion, is mostly carried out online. Network teaching improves the teaching efficiency, reduces the burden on teachers, and increases the flexibility of students' learning. Network resources are convenient for students to access, but pose difficulty for teachers in evaluating students' learning effects. Plagiarism detection technology and the artificial intelligence start to be adopted to help teachers teach online. In this paper, a novel method is proposed to help detect plagiarism and assist homework grading by collecting students' behavioral data and building statistical models for analysis.

Related Work

Plagiarism detection technology is widely used in the field of intellectual property protection. CNKI academic misconduct detection system relies on its large literature database to detect plagiarism. Wuhan University ROST anti plagiarism system provides similarity evaluation service between the candidate paper and existing literature. Plagiarisma is an open source plagiarism detection platform that provides plagiarism detecting services through search engine support. South China University of Technology developed a study management system in which the homework plagiarism detect module [1,2,3,4] can find the homework plagiarism within a class.

According to Underwood [5]'s research on e-learning, the sources of students' Plagiarism mainly consist of two types: local plagiarism and global plagiarism. If the whole or part of the student's work or thesis is obtained from the study community (university, class, or study group), this is called local plagiarism. In the global plagiarism, the source of plagiarism is relatively broad, including Internet, academic monographs, printed publications, P2P networks, and etc. Against global plagiarism, there
are many commercial or non-commercial systems that can provide plagiarism detection services. One of the most famous systems is the Turnitin system developed by iPlagiarism company of the United States. At present, many mainstream e-learning systems in the world, such as Blackboard, WebCT, ANGEL and Moodle, have adopted the Turnitin system either directly or through the company's plug-in's.

Most of the existing plagiarism detection systems determine the plagiarism relationship based on the publication time [6]. As for two papers with high similarity, plagiarism tends to be the one that was published on a later date. However, time order of student homework submission does not necessarily correlate with plagiarism. In other words, time order relation cannot be used as a factor to determine homework plagiarism. Plagiarism detection module in the existing learning management systems can only detect the existence of plagiarism within a set of homework, and it cannot determine in this set, who is the plagiarist, who is plagiarized. Yannakoudakis [7] analyzed the paper type homework, and found that there exists a certain relation between students' daily study habits and their attitude toward homework. Generally students who study hard tend to treat homework in a relatively serious way. This paper proposes a learning management system that records various learning behavior data from students. And based on the original plagiarism detection module, we develop a supplemental correction model of learning based on the behavior analysis, to solve the determination problem of student plagiarism relationship.

**Plagiarism Detection and Hierarchical Clustering Method**

**Plagiarism Detection**

In the plagiarism detection module, asymmetric similarity algorithm is used to determine the similarity between student’s homework. For a given two documents, a and b, the similarity between document a and document b is represented by the ratio $C_{ab}$: The degree of similarity is expressed by the ratio of the equivalent number of fingerprints in the sampled fingerprint set of document a and document b to total number of fingerprints with document a. By asymmetric similarity, $C_{ab}$ and $C_{ba}$ can represent the degree of inclusion between homework assignments. The correlation degree of homework is expressed as correlation coefficient $\rho_{ab}$. The relationship between the correlation coefficient $\rho_{ab}$ and the asymmetric similarity $C_{ab}$ is as follows:

$$\rho_{ab} = \sqrt{C_{ab}C_{ba}}$$

(1)

In the formula: $0<\rho_{ab}<1$, the closer $\rho_{ab}$ is to 1, the stronger the linear relationship is between the two document vectors; the closer $\rho_{ab}$ is to 0, the weaker the linear correlation between the two vectors. $1-\rho_{ab}$ represents the distance between the document vectors, and the smaller the $1-\rho_{ab}$ value is, the higher the similarity between the two documents is.

**Hierarchical Clustering**

Hierarchical clustering method is used to cluster the similar homework. Depending on the correlation coefficient between the homework, the distance is analyzed from small to large, and the documents with the minimal distance are merged into one cluster. This procedure is reiterated in the same way, until all documents are clustered or a threshold is reached. The method of hierarchical clustering is used to analyze the plagiarism of a course paper in a learning system. The result is shown in Figure 1:
In the diagram, the ordinate represents the correlation distance between the homework, the abscissa represents the number of homework. According to the definition of hierarchical clustering, the smaller the correlation distance, the higher the similarity between the two documents. Assume that 0.1 is used as the threshold for similarity operations. As can be seen from the graph, hierarchical clustering detects 8 groups of plagiarism coming from the same source and bearing the 90% similarity as the source.

**Plagiarism Determination**

Through hierarchical clustering, the homework with similar similarity is classified, but it cannot determine who is the plagiarist and who is plagiarized. In several similar homework assignments, it is relatively subjective to determine who is the plagiarizer. In practice, it is found that there is a certain correlation between the students' daily learning behavior and plagiarism. Students with good learning behavior are generally considered to be less likely to plagiarize. Students who are not active in their studies and who do not have good study habits are more likely to plagiarize.

Learning behavior statistics for a given period of time in the learning system based on data from the count of system log in, course access rate, lecture note access rate, the count of homework hand in, time sequence of homework hand in, section exam score, class attendant rate and other learning behavior. These factors will be considered in evaluating the performance of student study, and their correlation with the final exam score will be investigated. And these factors are further used to determine the correlation between plagiarism and their learning behavior.

In this experiment, we have studied the system data during 2016, and have selected statistical analysis system, the count of system log in, course access rate, lecture note access rate, the count of homework hand in, time sequence of homework hand in as the behavior learning indicators. We have analyzed the actual plagiarism survey results, the students' learning behavior, the final examination scores and actual cases of plagiarism, and the results are shown in Figure 2:
As can be seen from Figure 2, the final exam results of students are positively related to their degree of active study in the learning system. That is, students with more course access, more lecture note access, and more homework hand in are better performed in the final exam. It shows that students' learning behavior on the learning system is positively related to their achievements. There is a correlation between the sequence of homework hand in and final exam scores. As shown in Figure 2, the vertical coordinate values in the figure indicate that the larger the value, the earlier the submission time. The value 0 indicates close to the submission deadline. Students who hand in their homework relatively late consist of both the best and the worst ones, while students who submit homework earlier have a middle grade.

According to the actual survey results, exam results and homework plagiarism are related. Students with poor academic performance commit more homework plagiarism, while students with excellent results rarely plagiarize. One can judge who is a plagiarist from the time sequence of homework hand in and the learning behavior:

Good learning behavior, early submission of homework. This group of students have low probability of plagiarism, and most of them have good exam results.

Good learning behavior, late submission of assignments. Students in this group work are hard working, having low probability of plagiarism, a few of them having outstanding academic performance.

Learning behavior is poor, such as absenteeism, poor grades in unit tests, late delivery and missed delivery. This group of the students tends to be more likely to plagiarize, and exam results are poor.

**Experiments and Results**

In this experiment, the factors, such as the counts of system log in, the course access rate, the lecture notes access rate, the count of homework hand in, time sequence of homework hand in, are used to do the weighted supplemental determination. The establishment of auxiliary operation evaluation model based on the analysis of learning behavior to help exclude non plagiarism. The accuracy of the model is compared with the actual survey results, as shown in the table below:
We can see from the table, the accuracy of plagiarist determination using the model is significantly improved. Interestingly, we also observed that, for the groups with relatively large number of (6 or more students), the accuracy of the model decreases dramatically. Through the analysis, we learned that the plagiarist determination can be negatively affected by the wide scattering of student scores over a large range, as well as the increased number of plagiarism sources.

**Summary**

Experimental results show that the statistical analysis of learning behavior provides good supplemental functionality to the plagiarism detection algorithm, and can improve the accuracy of the determination of plagiarism. In the case that the group contains a large number of people, to maintain the accuracy, it is possible to add the determination of nonplagiarism. In the future work, we will introduce more indicators of learning behavior for evaluation, and improve the artificial intelligence module in order to further improve the system performance.

**Acknowledgement**

The work in this paper represents one of the achievements in the project sponsored by the National Natural Science Foundation of China (61070092).

**References**


