Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (14): 2774-2780, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.2774.2780

Research on the Design and Implementation of Online Homework Review and Similarity Comparison System

Xiaoying Wang, Wenyan Yuan, Weili Guo and Xiaojing Liu
Department of Computer Technology and Applications,
Qinghai Computer Science University, 810016, Xining, Qinghai, China

Abstract: Since electronic homework is used more and more popularly and widely in universities, it becomes important and meaningful to manage and analyze the students' homework in an online way. In this study, we have conducted relevant research on the design and implementation of an online homework review and similarity comparison system based on the B/S architecture. This system could help teachers to publish homework, collect homework, review homework and automatically detect the similarity of several copies of different homework. Plagiarism detection results could be given out as suggestions which could help teachers a lot to pick out the cheating students. The design of the system architecture and the specific detection algorithms are described and analyzed in detail. Experiments show that our system is feasible and convenient to use which will be valuable to spread and apply into the education practice.

Key words: Online homework review, similarity comparison, plagiarism detection, group classification, sequence alignment

INTRODUCTION

As the rapid development of computer and Internet technology, electronic homework gradually becomes more and more widely-used instead of traditional homework written on the paper, especially in the universities (Hong *et al*., 2010). Many teachers choose to assign and collect electronic homework in and out of class, including various kinds of documents, such as experiment report, essays, papers and also program codes. The wide usage of electronic homework enables the paperless management of the course which consequently improve the efficiency of both the assignment and collection. However, from another point of view, it also increases the possibility of plagiarism (Zhu and Song, 2002) since the electronic documents are too easy to copy and paste. It takes plenty of time and labor to review the students' homework one by one and detect possible plagiarism manually. Hence, it will be of great use to implement an online homework review and similarity comparison system, in order to offload the teachers' heavy work.

On the other hand, leveraging online detection system to check the homework could also have psychological impact back on the students. Recently, we have conducted a questionnaire survey among the students studying in our department and collected 82 valid survey papers, in which there are 50 papers from boys and 32 papers from girls. The percentages of freshman, sophomore, junior and senior students are 20.73, 31.71, 17.07 and 30.49%, respectively. According to the survey results, more than 60% students declaim to resist plagiarism. However, only 13% students said that they never copy others' homework.

Furthermore, about 79% of the students believe that most of the time, plagiarism could not be discovered by the teachers which becomes the main reason that they are prone to "refer to" others' homework. At last, we asked a question that "would you consider to complete your homework independently if your teacher uses anti-plagiarism software to compare the homework?". Only one person choose "I'll still try to copy others' work" which all of the other people choose "I'll consider completing the homework all by myself". This shows that most of the students are prone to copy is due to the lack of proper supervisory mechanism. From this point of view, the application of online similarity comparison system will greatly improve the overall quality of students' homework.

Plagiarism detection systems and algorithms have been concerned by researchers for decades. The first typical anti-plagiarism software, WordCheck, was developed in 1991 (Dailey Paulson, 2002). Then,

Corresponding Author: Xiaoying Wang, Department of Computer Technology and Applications, Qinghai University, Xining, Qinghai Province, China

Manber (1994) developed SIF tool which could search in large file based on string match. In 1996, Wise. (1996) developed YAP1 and YAP2 which were used for detecting the similarity of program codes. Then, Monostori *et al.* (2000) proposed the MDR model with a text comparison algorithm based on suffix tree, in 2000. Later, Schleimer *et al.* (2003) proposed to use the Winnowing algorithm to accurately recognize the document replication problem based on digital fingerprints.

At the same time, because of the difference between Chinese and English, the above research achievements could not be directly applied to Chinese documents. Consequently, many Chinese researchers have made efforts on Chinese document similarity detection studies. Jin *et al.* (2005) combined the digital fingerprint technique and word frequency count technique to analyze the structure of the entire document. Later, Wuhan University presented ROST anti-plagiarism system which adopted soft match basic on information fingerprint and thus improved the efficiency and the applicability.

In this study, we focused on the similarity comparison of electronic homework. Since the homework is usually assigned with a template or a guide book, the similarity detection system should be aware of the possible reduplication parts among different students' work. Furthermore, the teachers often wonder the copying and imitating relationship among multiple students, since a student might pieces several other ones' work together and makes up a complete document. Thus, we also considered group classification analysis in our system design and implementation. In the remainder of this study, we'll introduce the system design, the detailed algorithm and the illustration of our system. Finally, the conclusion is given and some future work will be discussed.

OVERVIEW OF SYSTEM DESIGN

The entire system is comprised of three main parts: document preprocessing, text similarity comparison and group classification analysis. The data flow diagram of the designed system is shown in Fig. 1.

Specifically, the teachers send the document set and the homework template to the preprocessing module. Then, the strings in the documents and the template are read out which are further used to do word segmentation. After that, digital fingerprints could be generated and the documents will be filtered using the template.

In the next step, sequence alignments are conducted on the fingerprints to compare and compute their similarity value. Then, the similarity matrix could be calculated out which will further be normalized to [0,1] interval.

In the last stage, group classification is done to the similarity matrix and the grouping information will be generated for visualization. The visualized results together will the 2-dimension matrix will be sent back to the teacher, illustrated in a web page.

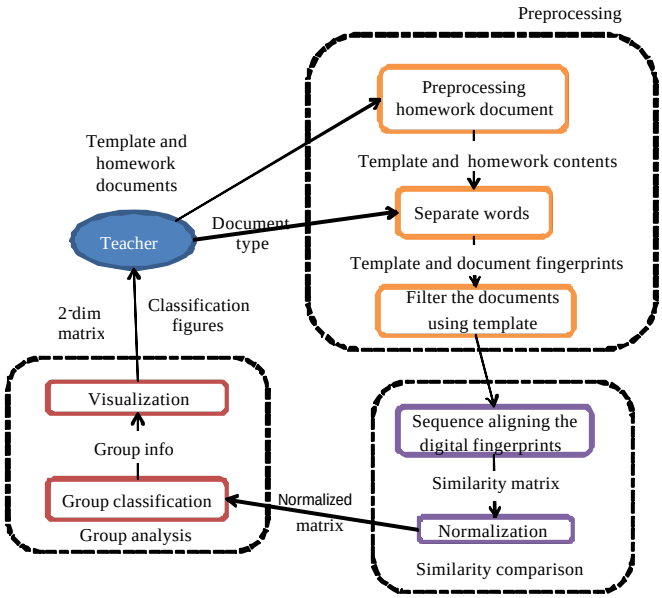

Fig. 1: System data flow diagram

SIMILARITY COMPUTATION PROCEDURE AND ALGORITHMS

Preprocessing: Document preprocessing is necessary for later comparison and similarity calculation which involves match unit segmentation of the template and other homework documents, the string structure design and template filtering methods. Thus, it's important to carefully design the document preprocessing approach.

In order to make the plagiarism detection more flexible, we separate the documents into text blocks. Blocks are regarded as the basic units for similarity comparison. The separation could be based on keywords, sentences or paragraphs. Here, we divide the target documents into multiple sentences according to appearance of punctuation marks such as the comma symbol, period symbol and carriage return characters. The generated sentences are called digital fingerprints in this study. The detailed procedure is described as in Fig. 2.

As shown, after the document preprocessing, the sentence fingerprints of the homework set could be generated for further computation.

Sequential alignment and similarity calculation: The procedure of similarity calculation is mainly comprised of following steps:

Step 1: Select fingerprints s_1 and s_2 from the fingerprint sets of homework documents A and B, respectively

Step 2: Figure out the longest common sequence of s_1 and s_2, denoted as:

$$s_{common}(s_1, s_2) = b_1 b_2 \cdots b_r \quad (1)$$

Document preprocessing
- Define N as the number of homework documents
- Construct string S[0] by reading out all of the characters from template
- Construct string S[1]~S[N] from other homework documents
- for i=0 to S.length
- Segment S[i] according to carriage return characters, and obtain the paragraph fingerprints S[i][]
- end for
- for i=1 to S.length
- compare S[i] with S[0] and filter S[i][] to S'[i][]
- end for
- for j=0 to S'[i].length
- segment S'[i][j] according to predefined separation marks and obtain the sentence fingerprints S'[i][j][]
- end for

Fig. 2: Pseudo code of document preprocessing

where, $b_i (1 = i = r)$ denotes a semantics element of the sequence. The Smith-Waterman algorithm (Smith and Waterman, 1981) is adopted here which is based on the concept of dynamic programming methods, filling a dynamic matrix D one step at a time. Denote the subsequence of s_1 and s_2 as:

$$sub^r_1 = a^1_1 a^2_1 \cdots a^r_1 \quad (2)$$

$$sub^l_2 = a^1_2 a^2_2 \cdots a^l_2 \quad (3)$$

Then, each element D(r,l) in the matrix denotes the maximum similarity of the common sequences of sub^r_1 and sub^l_2. If tracing back the matrix D, we can get the optimal match result of the best common sequences and the similarity could be calculated.

Each element D(r,l) could be calculated according to the left, up and left-up elements. However, there might be three kinds of different situations which need to discuss.

- b_t corresponds to both a^r_1 and a^l_2
- b_t doesn't correspond to a^r_1
- b_t doesn't correspond to a^l_2

We call the last two status as "in gap" states. Hence, in each step, the three different states should be recorded into each element to guarantee the correctness of the algorithm.

Next, the matrix D will be filled step by step, according the following equation:

$$\begin{aligned} H(r,l) &= \max\begin{cases} 0 \\ E(r,l) \\ F(r,l) \\ H(r-1,l-1) + sub(a^r_1, a^l_2) \end{cases} \\ E(r,l) &= \max\begin{cases} H(r,l-1) + \alpha + \beta \\ E(r,l-1) + \beta \end{cases} \\ F(r,l) &= \max\begin{cases} H(r-1,l) + \alpha + \beta \\ F(r-1,l) + \beta \end{cases} \end{aligned} \quad (4)$$

where H(r,l) denotes the maximum score of the three states, E(r,l) denotes the score when s_1 is in gap and F(r,l) denotes the score when s_2 is in gap; α is the penalty factor for each unaligned element and á is the extra penalty factor for the entire sequence; sub(x, y) is a function to calculate the score of matching two single elements x and y which is defined as follows:

$$sub(x,y) = \begin{cases} 1 & if(x = y) \\ p & otherwise \end{cases} \quad (5)$$

where, p is a negative integer with a relatively large absolute value.

Step 3: Calculate the similarity between s_1 and s_2 according to the following equation:

$$sim(s_1,s_2)=\frac{2k}{L_1+L_2} \quad (6)$$

where, k is the length of longest common sequence of s_1 and s_2, L_1 and L_2 are the length of s_1 and s_2 respectively

This equation is applicable since the similarity value will be between [0, 1] interval in the general cases. The value will be 1 only if s_1 is same as s_2 and 0 only if s_1 and s_2 has no common characters. Furthermore, the calculation is symmetric since sim (s_1, s_2) = sim (s_2, s_1).

Step 4: Traverse the fingerprint sets A and B to do the sequence alignment of all the fingerprints and get the 2-dimension matrix storing the similarity value of each pair of fingerprints. If a similarity value is figured out to be 1, we regard it as a replicated fingerprint.

Step 5: Calculate the similarity between the two homework documents as:

$$sim(A,B)=\frac{n_r^A+n_r^B}{N_A+N_B} \quad (7)$$

where n_r^A and n_r^B are the number of replicate fingerprints in document A and B respectively and N_A and N_B are the total number of sentence fingerprints of A and B, respectively.

Grouping and classifying: Till now, we can get the similarity of each pair the two homework documents which are recorded in the matrix M. In order to facilitate the plagiarism detection, we should further exploit the matrix and do deeper analysis. Here, in the design of our system, the teacher could specify a predefined threshold t between 0~1 which is used to process the matrix M. Given threshold t, M could be turned into an incidence matrix M', according to the following equation:

$$M'(t)_{ij}=\begin{cases}1 & \text{if } (M_{ij}>=t)\\ 0 & \text{otherwise}\end{cases} \quad (8)$$

from which we could draw the corresponding graph G. Then, the grouping and classifying problem is turned into the problem that how to figure out the set of connected sub-graphs of the original graph G. If a certain sub-graph contains more than one vertex, the corresponding homework should be considered as possible plagiarism and the sub-graph will be regarded as a plagiarizing group.

COMPLEXITY ANALYSIS

In this section, we attempt to analyze the time and space complexity of the whole algorithm described in previous sections. Since the algorithm is constructed by multiple modules, we'll analyze it step by step.

Document preprocessing: As mentioned above, the preprocessing stage is mainly comprised of three parts, including reading document contents, getting strings from documents and filtering the strings with the given template and getting the fingerprint set.

First, it can be observed that the time complexity of importing all the homework and getting the strings is O(k), wherein k is the number of homework documents.

Secondly, the time spent on splitting the documents based on some punctuations will increase not only as the number of target documents becomes larger but also as the length of the documents increases. Hence, the time complexity of getting the fingerprints should be O(kn) (usually k<<n), wherein n is the average length of target homework documents (i.e. the total number of characters).

Finally, the last step of preprocessing is to compare the fingerprints in each document with the given template one by one. Assume the average number of fingerprints in one documents is m. Then, the time complexity of the filtering step should be $O(m^2)$.

Similarity calculation: The time complexity of comparing a single pair of fingerprints is $O(s^2)$, where s is the average length is the fingerprints. Then, traversing all of the fingerprint pairs of two documents is $O(m^2s^2)$ (usually s<<m). Finally, calculating out the 2-dimension similarity matrix should consume $O(k^2m^2s^2)$ time.

Grouping and classification: After computing out the k¡Ák similarity matrix, flood filling algorithm is adopted to classify the pairs into different groups based on a predefined threshold. The time complexity is mainly determined by k which is $O(k^3)$.

From the above analysis, we could get the summarized time complexity of the entire procedure which is $O(kn+k^2m^2s^2+k^3)$.

ILLUSTRATION AND DISCUSSION

In this section, we'll illustrate the implemented web-based system by showing different parts of functions and discuss about them.

Online review: In order to facilitate the homework review process, we have implemented the online review components which read out the contents of homework documents and turn them into web pages. Then, on the

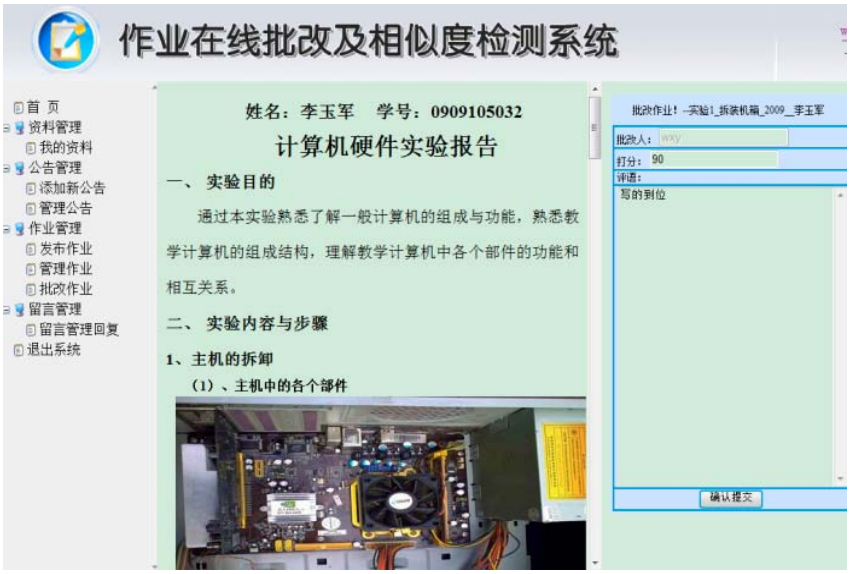

Fig. 3: Snapshot of online homework review

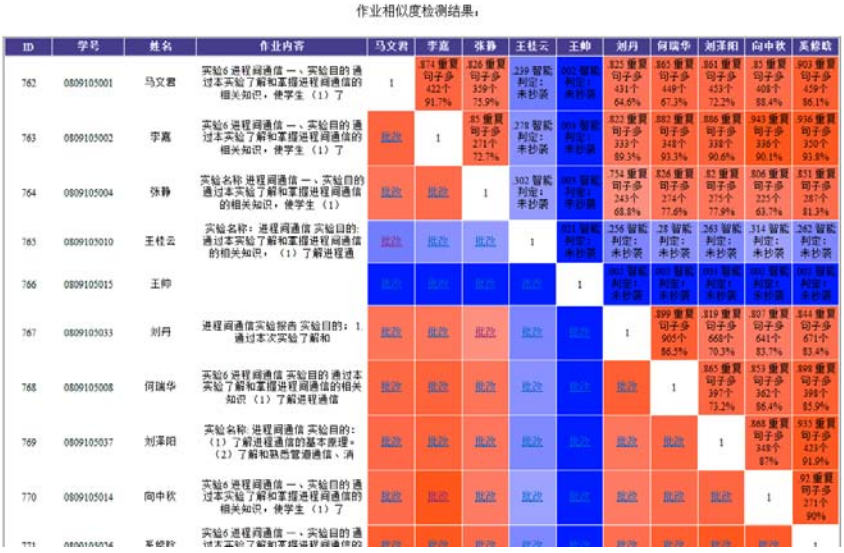

Fig. 4: Snapshot of showing the similarity matrix

browser side, the users can view the contents without installing any extra software or plugins. The teachers could give corresponding scores and comments after viewing the documents online, without download the files to their local hard disks. Fig. 3 shows a snapshot of the online review page, where the right side is the review table for teachers to fill and submit.

Online similarity comparison: After some students submit their homework, the teacher can choose to review and let the system help compare their homework. The comparison is done by using the algorithm described in the previous section. The obtained results will be shown in a table, as illustrated by Fig. 4. The backgrounds of the cells in the table are painted by different colors to imply the similarity situations. The closer to red means higher similarity and the closer to blue means lower similarity. In this way, the teacher can quickly find the possible plagiarism pairs. Furthermore, the system also provide information about how much fingerprints are same in a

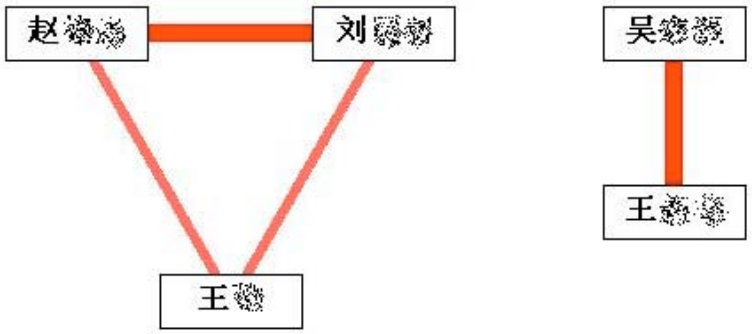

Fig. 5: Snapshot of showing the similarity matrix

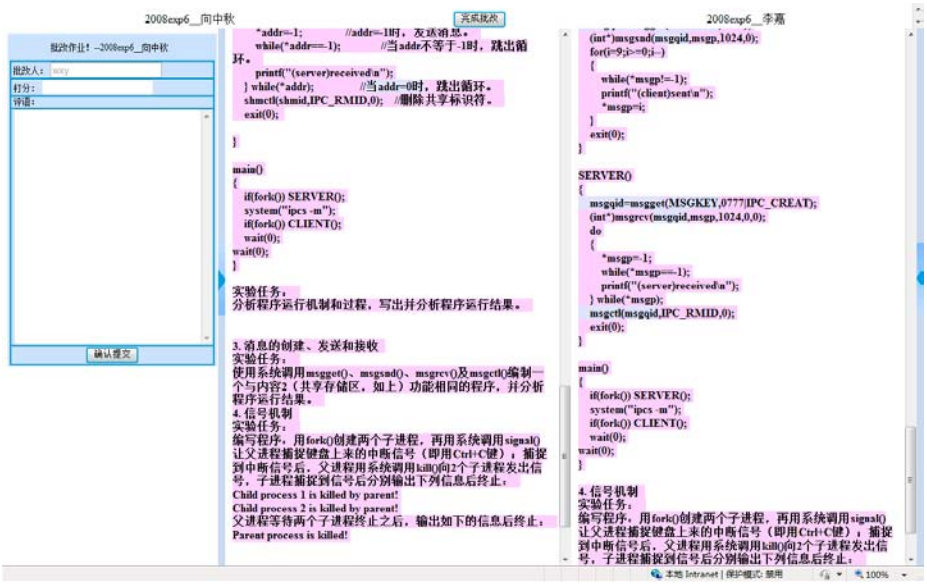

Fig. 6: Highlight the similar contents

pair of documents and then draw a conclusion intelligently to instruct the teachers which are both shown inside a single cell.

Group analysis and visualization: If the number of homework documents becomes large, it's still not enough clear if the similarity values are shown in the 2-dimension table, since the scale of the table will be too large to locate possible plagiarizing pairs. Thus, we attempted to classify the homework pairs into groups and visualize them in a graph for further analysis. The classifying method has been described in previous sections. Here, an example visualized result graph is illustrated in Fig. 5, wherein the vertices represent the students' names and the border width of the edge between two students is determined by the similarity value. The higher the value is, the thicker the edge is and vice versa. In this way, the teacher could specify a threshold value and quickly find some possible plagiarism relationship between multiple students.

Content highlight: To make the web-page-based homework review more clearly for the teacher, the same sentences and paragraphs are marked in highlight color when the teacher enter the comparative correction page. Fig. 6 shows the highlight effect where two homework documents are read out and listed in vertical frames. Besides, it can also be observed there is that another table

shown in the left frame which is provided for the teacher to give comments and scores at the same time as they review and compare the two documents. The teacher can choose to show or hide the table. As shown in Fig. 6, the right side table is now hidden. In this way, the teacher can clearly find similar contents in a pair of target documents which is practically convenient for homework review.

CONCLUSION

In this study, we have presented the design and implementation of a web-based homework similarity calculation system. The overall system design is outlined, including the modules and their relationship shown in a data flow diagram. Then, we presented the similarity comparison procedure and elaborate the algorithms in detail, including document processing, fingerprint computing and the sequence alignment process. The time complexity of the whole algorithm is analyzed to show the feasibility of our approach. At last, we illustrate the function and interfaces of our system including the 2-dimension table, the visualization of group analysis and the content highlights. By applying such a system, the students can submit their homework online and the teachers could review, correct, compare, give scores and comments for the homework online. Most importantly, they can let the system the compute the similarity in the back end and give analysis visualization graphs in the web page which greatly facilitates the plagiarism detection and recognition.

ACKNOWLEDGEMENT

This study is partly supported by the High Education Research and Reform Project in Qinghai Province ("Implementation and Application of Homework Similarity Detection Tools"), Education Teaching and Research Project of the Qinghai University in 2012 (No.2012201301) and also the Course Construction Project in Qinghai University (No. KC-11-3-7, KC-12-2-3).

REFERENCES

Dailey Paulson, L., 2002. Professors use technology to fight plagiarism. Computer, 35: 24-25.

Hong, L., W. Jirong and Z. Longyi, 2010. Application research of intelligent management information system for electronic homework. Res. Explorat. Lab., 29: 311-313.

Jin, B., Y. Shi and H. Teng, 2005. Similarity algorithm of text based on semantic understanding. J. Dalian Univ. Technol., 02: 291-297.

Manber, U., 1994. Finding similar files in a large file system. Proceedings of the 1994 USENIX Conference, pp: 1-10.

Monostori, K., A. Zaslavsky and H. Schmidt, 2000. Document overlap detection system for distributed digital libraries. Proceedings of the 5th ACM Conference on Digital Libraries, June 2-7, 2000, San Antonio, Texas, USA., pp: 226-227.

Schleimer, S., D.S. Wilkerson and A. Aiken, 2003. Winnowing: Local algorithms for document fingerprinting. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 9-12, 2003, San Diego, California, USA., pp: 76-85..

Smith, T.F. and M.S. Waterman, 1981. Identification of common molecular subsequences. J. Mol. Biol., 147: 195-197.

Song, Q.B. and J.Y. Shen, 2001. On illegal copying and distributing detection mechanism for digital goods. J. Comput. Res. Dev., 38: 121-125.

Wise, M.J., 1996. YAP3: Improved detection of similarities in computer program and other texts. ACM SIGCSE Bull., 28: 130-134.

Zhu, G. and Q. Song, 2002. Design and implementation of student's papers handle system on network based on web. Comput. Eng., 28: 251-253.