Detecting Methods of the Plagiarism for Student Reports Using Text Processing

Yi-Ching Tsai[1], Gendo Kumoi[3], Makoto Suzuki[2,] Takashi Ishida[3], Shigeichi Hirasawa[3]

[1]Leader University, Tainan City 70901, Taiwan
(tsaiyiching@mail.leader.edu.tw)
[2]Shonan Institute of Technology, Fujisawa, Kanagawa 251-8511, Japan
(msuzuki@fork.ocn.ne.jp)
[3]Waseda University, Shinjuku, Tokyo 169-8555, Japan
(motories@ruri.waseda.jp)

ABSTRACT

With the popularization of internet in recent years, almost all the information could be found on Web pages. It becomes very easy for students to copy articles from internet and paste them in their assigned reports. To avoid the behavior of plagiarism and detect the violation of copyright, we use not only the Web search engine to discover articles with the doubt of students' illegal plagiarism, but also propose the following two automatic classification methods to inspect the supposition. 1. CKIP Chinese Word Segmentation System, 2.Smith-waterman Algorithm.

Keywords: Plagiarism, Student Reports, CKIP Chinese Word Segmentation System, Smith-waterman Algorithm

1. INTRODUCTION AND BACKGROUND

Plagiarisms in students' assessed homework are issues of increasing concern to the academic community as a whole. In many disciplines, there is concern over web-based plagiarism, whereby students use material, unattributed from one or more sources on the World Wide Web. Especially with the popularization of internet in recent years, almost all the information could be found by clicking in few keywords. It becomes very easy for students to copy articles from internet and paste them in their assigned reports.

To discover all these web-based plagiarism manually is very difficult nowadays. Because we not only need to collect different origin- plagiarized articles from the web pages as much as possible (in case that web articles plagiarized to each other), but also need to judge whether student reports of part or all just copied (or thinly disguised by changing some words or replacing the sentence orders) the origin-plagiarized web materials by reading both of them one by one. Furthermore, it is necessary to show the appropriate plagiarism part as evidence to make the judgment of students' plagiarism act.

In our research, we devise the similar documents discovery technique by dividing sentences of the student reports and internet articles into word units to discover the continuous word units that co-occurred between both documents. By using this discovery technique, we aim to see if the retrievals could correctly detect the student reports with the doubt of plagiarism by comparing with the reports we manually judged as plagiarizing from the web pages articles.

2. EXPERIMENT MATERIAL

In this experiment, we use 101 leader university students' terminal reports from the foundation course "Movie Technique"(電影科技). Due to some of student reports could be predicted as plagiarizing the web page, we use YAHOO! 奇摩[1] search engine to retrieve 172 articles by human-hand, and try to compare with all the student reports in advance to find out the doubtful reports for the assessment experiment.

2.1 Student reports

Course name: "Movie Technique" (電影科技)
Homework: Write the following three items.1.Outline, 2. Impressions, 3. Technique analysis in the report. Student is free to select one from the four movies to write their report. "The godfather"（教父）, "The lord of the rings"（魔戒）, "Pirates of the Caribbean"（加勒比海盜）, "無米樂".
Submission: Hand out the report in the form of the Word of Microsoft's office 97 and 2000 by email.
The structure and numbers of colleting student reports: "The godfather" for 9. "The lord of the rings" for 34. "Pirates of the Caribbean" for 35. "無米樂" for 23.
(Hereafter, student reports of these four movies are abbreviated respectively as "G","R","C" and "U".)

2.2 Internet articles

The next step, we picked up and downloaded articles from different sources on the web pages in the form of Word of Microsoft's office 97 as the origin-plagiarized materials to check the above-mentioned four contents of reports by human-hand.
The structure and numbers of Internet download articles: "The godfather" for 23. "The lord of the

rings” for 45. “Pirates of the Caribbean” for 72. “無米樂” for 32.

(Hereafter, internet downloaded articles of these four movies are abbreviated respectively as "IG","IR","IC" and "IU".)

2.3 Comparison by human-hand

In order to evaluate if our computer discovery technique could correctly achieve the detection just as the text processing made by human-hand, we managed to read all student reports concerning each movie to pick out and classify which were plagiarizing internet articles in advance. And the manual comparing results are as following.

- Student reports about "G":
  3 in 9 plagiarized internet articles.
- Student reports about "R":
  14 in 34 plagiarized internet articles.
- Student reports about "C":
  21 in 35 plagiarized internet articles.
- Student reports about "U":
  12 in 23 plagiarized internet articles.

Furthermore, to examine if the detecting methods we suggest later could find out the plagiarized parts exactly, we also marked out the plagiarism part of student reports where students somehow obtained a copy of internet articles as evidence to judge students' plagiarism act. The compassion results by human-hand are arranged as following Table 1.

Table 1: Plagiarism found between student reports and internet articles by human-hand

Movie	Student reports : internet plagiarized articles
“G”(3 cases detected)	G1 : IG21（1 corresponded）G2 : IG22（1 corresponded）G8 : IG23(1 corresponded）
“R”(14 cases detected)	R1 : IR21 （1 corresponded） R2 : IR22、IR23、IR24 （3 corresponded）R3 : IR25、IR26、IR27（3 corresponded）R5 : IR28（1 corresponded）R6 : IR29（1 corresponded）R11: IR30 （1corresponded ） R15 : IR31、IR32、IR33、IR34 （4 corresponded） R20： IR35、IR36、IR37 （3 corresponded） R21 : IR38 （1 corresponded） R22 : IR39 （1 corresponded） R23 : IR40 （1 corresponded） R28 : IR41 （1 corresponded） R34 : IR44 （1 corresponded） R35 : IR45 （1 corresponded）
“C” (21 cases detected)	C2 : IC38、IC39、IC40（3 corresponded）C3 : IC41 （1 corresponded） C4 : IC42 （1 corresponded）C5 : IC43、IC44 （2 corresponded） C11 : IC45、IC46、IC47、IC48 （4 corresponded）C12 : IC49、IC50、IC51、IC52（4 corresponded） C13 : IC50、IC53、IC54、IC55（4 corresponded） C14 : IC38、IC56 （2 corresponded） C15 : IC43、IC44 （2 corresponded） C18 : IC57 （1 corresponded） C21 : IC61 （1 corresponded） C22 : IC49、IC62、IC63（3 corresponded）C23 : IC46、IC53、IC62、IC64、IC65、IC66、IC67（7 corresponded）C25 : IC72 （1 corresponded）C26 : IC68（1 corresponded）C28 : IC68、IC69、IC70、IC71 （4 corresponded） C30:IC21,IC22,IC23,IC24,IC25,IC26 ,IC27 ,IC28 ,IC29 ,IC30 （10 corresponded） C31: IC31 （1 corresponded） C33: IC32 ,IC33 （2 corresponded） C34: IC34, IC35 （2 corresponded） C36:IC37 （1 corresponded）
“U” (12 cases detected)	U1 : IU21、IU22、IU23、IU24 （4 corresponded） U4 : IU25 （1 corresponded）U5 : IU26、IU27、IU28（3 corresponded） U6 : IU27（1 corresponded）U7 : IU27（1 corresponded） U8 : IU26 （1 corresponded）U11 : IU29、IU30、IU23（3 corresponded）U13 : IU27（1 corresponded）U14 : IU25、IU27 （2 corresponded）U18 : IU31、IU26（2 corresponded）U20 : IU27 （1 corresponded） U23 : IU32、IU26（2 corresponded）

- The table shows the contrast between student report and internet articles. For example, when student 1 submits report about “G” wholly or partly plagiarized from internet article 1 about “G”, we show the comparison as G1:IG1. And the corresponded numbers means the sauces where student reports plagiarized from.

3. CKIP CHINESE WORD SEGMENTATION SYSTEM

To divide sentences of 101 students reports and 172 internet articles into the smallest word units efficiently for the detecting method later, we make use of “CKIP Chinese Word Segmentation System” (中文斷詞系統)[2] developed by Taiwan Academia Sinica to do the morphological analysis.

4. SMITH-WATERMAN ALGORITHM

To discover the continuous word units that co-occurred between 101 student reports and 172 internet articles and to examine the whole materials, we use the Smith-waterman algorithm proposed by Robert W. Irving [3] in our research. To formulate the classical Smith-Waterman dynamic programming scheme [4], Robert W. Irving defined S_{ij} to be the maximum score obtainable by aligning a substring of X ending at position i with a substring of Y ending at position j. The standard recurrence relation for S_{ij} is

$$S_{ij} = \begin{cases} S_{i-1,\, j-1} + h \text{ if } X(i) = Y(j) \\ \max(0,\, S_{i-1,\ j} - d,\, S_{i,\, j-1} - d,\, S_{i-1,j-1} - r) \end{cases}$$

otherwise, subject to the initial conditions $S_{i,0} = S_{0,j} = 0$ for all i, j.

Notice that a negative score is impossible, since aligning the empty substrings ending at positions i and j yields a score of zero. Application of this recurrence relation leads to a dynamic programming algorithm enabling the computation of the elements of the array S, for example in row by row order. As is standard with dynamic programming schemes of this kind, he use the idea of a traceback path to construct an optimal local alignment ending at position i in X and position j in Y. For a given cell (i, j) he defines a parent cell as follows:

- if $S_{ij} = 0$ then (i, j) has no parent;
- if $X(i) = Y(j)$ then (i, j) has the parent $(i - 1, j - 1)$;
- in addition (i, j) has as a parent any cell $(p, q) \in \{(i - 1, j), (i, j - 1)\}$ such that $S_{ij} = S_{pq} - d$, and/or cell $(i - 1, j - 1)$ if $S_{ij} = S_{i-1,j-1} - r$.

So each cell containing a non-zero value has at least one parent, and may have as many as three. For any cell *(i, j)* for which $S_{ij} > 0$, he defines a traceback path in the array to be any path obtained by starting from cell *(i, j)*, stepping successively from a cell to a parent cell, and terminating as soon as the next step in the path would reach a cell with a zero entry. Let $O_{ij} = (x_{ij}, y_{ij})$ be the final cell in a traceback path for cell *(i, j)* (so that the value in this cell is necessarily equal to *h*). Robert W. Irving called O_{ij} an origin for cell *(i, j)*. Because parents need not be unique, a cell may have more than one traceback path and more than one origin. Any origin for cell *(i, j)* specifies the starting points in *X* and *Y* respectively of a highest scoring local alignment ending at *X (i)* and *Y (j)*.

In our thesis, we use the expression of *S_{ij}* of the algorithm in detail and we assume the parameter *d* and *r* as 1 in the expression to fit our research. With the assumption, the search begins counting backward between *S_{i-1,j-1}*, *S_{i-1,j}*, and *S_{i,j-1}*for the maximum values. When the repetition of moving destination becomes 0 in *S_{i,j}* , the search should stop. And in-between the numerical value before the search stops and the maximum value should be assumed as the plagiarism part. In addition, when the continuous word units that co-occurred inside student reports and internet articles are over ten words, we can judge the certain part as plagiarism sentence. Moreover, the word unit targeted in this research is assumed to be a noun or a verb.

5. EXPERIMENTS RESULTS

5.1 Comparison between human-hand and automatic-detecting data

At first, we examined the auto-searching data of "G" by checking all student reports classified as plagiarizing or not-plagiarizing internet articles by human's judgment. It ends up that not only all the manually detected plagiarized-articles were all checked out but also the plagiarized-parts of the sentences that had been overlooked through human judgment were neatly confirmed. (Table 2)

Table 2: Plagiarism found by automatic-detecting methods between student reports and internet articles

Movie	Student reports：web-based plagiarism articles
"G"	G1 ：IG21,IG2,IG5,IG7,IG9,IG22(6 corresponded) G2 ：IG22,IG4, IG8,IG10, IG11,IG13,IG14, IG16,IG17, IG21, IG23 （11 corresponded） G8 ：IG23, IG4 , IG13, IG17, IG22 (5 corresponded)

- The underlined part means the extra web-based plagiarism articles detected inside student reports by using our detecting processing.

Furthermore, student reports G5 and G6 were judged as not plagiarism by human-hand; meanwhile, the data shown by our automatic detecting method judged the 2 reports as plagiarism. By double-checking G5, G6 and internet articles, we find out that the automatic-detecting processing made a right judgment. Thus the number of student reports about "G" plagiarizing internet articles number should be corrected from 3 to 5.

Secondly, we examined the data of Table1 by comparing to the automatic-detecting data. It is interesting to find out that not only all the human-judged plagiarism cases of student reports were proved to be as they are, but also human-judged correspondence (C12 : IC49) was proved to be a mistake. This means that human-judged results could not be precise as our automatic-detecting data.

In short, using our proposing method, student reports concerning "G", "R", "C" and "U" judged as plagiarism by human can be exactly detected out near 100 %.

And then, just as above Table 2 shows, the plagiarized-sentences not being found out manually were neatly detected by our detecting data.（Table3）

Table 3: The correct numerical values of the web-based articles sauces found by automatic detecting data inside student reports

Movie	Student reports (the total web-based plagiarism articles)
"R"	R1 （1+2） R2 （3+ 15） R3 （3+ 13） R5 （1+14） R6 （1+12） R11 （1+2） R15 （4+17） R20 （3+16） R21 （1+1） R22 （1 +1） R23 （1+2） R28 （1+2） R34 （1+7） R35 （1+0）
"C"	C2 （3+1） C3 （1+15） C4 （1+2） C5 （2+0） C11 （4+ 4） C12 （3+7） C13 （4+5） C14 （2+2） C15 （2+0） C18 （1+1） C21 （1+4） C22 （3+15） C23 （7+14） C25 （1+1） C26 （1 +0） C28 （4+4） C30 （10+7） C31 （1+2） C33 （2+7） C34 （2+5） C36 （1+1）
"U"	U1 （4+8） U4 （1+1） U5 （3+1） U6 （1+ 0） U7 （1+0） U8 （1+ 0） U11 （3 +7） U13 （1+0） U14 （2+0） U18 （2+ 6） U20 （1+0） U23 （2+2）

- The +number part means the figure of extra web-based articles found inside student reports detected by the automatic-detecting data.

On the other hand, 14 student reports about "R", 1 student report about "C" and 3 student reports about "U" judged as not plagiarized by human-hand are judged as plagiarized ones according to the automatic-detecting data. (Table 4)

Table 4: Plagiarism found by automatic-detecting data between student reports and internet articles

Movie	Student reports judged as not plagiarized by human-hand but judged as plagiarized according to the automatic-detecting data
"R"	R4、R7、R8、R9、R10、R12、R13、R14、R25、R26、R27、R31、R32、R33　(14 more student reports detected)
"C"	C29　(1 more student report detected)
"U"	U3、U17、U22　(3 more student reports detected)

By following the automatic-detecting data to check the correspondence of internet articles and student reports, we proved again that the automatic-detecting data give the right judgment.

Thus student reports about "R" plagiarized internet articles numbers should be corrected from 14 to 28, reports about "C" plagiarized number should be corrected from 21 to 22, reports about "U" plagiarized number should be corrected from 12 to 15.

5.2 Extra experiments

5.2.1 Automatic-detecting data about the comparison of student reports

In this part, we make extra experiment to compare student reports. To see the difference from the results we got in last section, we choose to skip over student reports judged as plagiarism by human-hand and automatic-detecting data. Instead, we pay special attention to student reports judged as not-plagiarizing by human-hand to see their correspondence. According to the automatic-detecting data about student reports, there are 3 types of correspondences in-between student reports.

1. Judged as not-plagiarizing by both human-hand and automatic-detecting data.

4 student reports (G3、G4、G7、G8) about “G”, 6 student reports (R16、R17、R18、R19、R24、R30) about “R”, 6 student reports (U2、U12、U15、U16、U19、U22) about “U”, and 10 student reports (C1、C6、C7、C8、C9、C10、C16、C17、C24、C32) about “C” don't have the corresponding data about other reports. Therefore, we could judge these 26 reports as the original submissions written by students their own.

2. Judged as not-plagiarizing by human-hand but found the mutual correspondence inside student reports according to the automatic-detecting data.

According to the automatic-detecting data, U9 and U10 corresponded to each other. After checking these two contents, we found out that they are just exactly the same. This shows that human-hand check did not find out they are the similar articles. Consequently, the plagiarism could be speculated as following two situations. 1. Both student reports plagiarized internet articles which we did not find out through web page by human-hand. 2. One student plagiarized another one’s report.

Anyway, by using the automatic-detecting data of student reports is quite helpful for checking the plagiarism. That is to say, when we only use the automatic-detecting data of comparing student reports and internet articles, we can not compensate the shortage of quantity about only downloading 32 internet articles for “U”. Thus, we could use this extra experiment to detect the plagiarism inside student reports. At the same time, we could take the comparing result and above-mentioned result together as a reference while we evaluate students’ grades.

3. Judged as not-plagiarizing by human-hand but found the one-way correspondence inside student reports according to the automatic-detecting data.

For example, there is no correspondence data of C27 in C1 though there is a corresponded data of C1 in C27. When we check the sentences between student reports, they show as the next quoting descriptions.

(C27) 貝克特公爵,打斷伊莉莎白和威爾即將舉行的婚禮,以協助海盜脫逃的罪名,要將伊和威 2 人處死!

(C1) 在第一部最後,互相告白的伊麗莎白和威爾終於舉行婚禮了。但是海軍卻要以協助海盜脫逃的罪名逮捕他們。

- The fence part means the correspondence parts shown by our automatic-detecting data.
- The words in shadow mean the different words between two student reports inside the correspondence part.

According to the contexts, we can consider both narrations as original because they are not writing the same thing. Thus we could judge they are not-plagiarized reports.

As for the reason why the data only shows no correspondence of C27 in C1, we concluded that the sentence of C1 is longer than C27 by expanding the sentence with more words. In other words, C1 sentence just incidentally connoted same keywords internally as C27 having (see the fence part). That means we need to check this kind automatic-detecting data again by human-hand to confirm if our automatic-detecting method made a mistake.

However, this type of correspondence could also be detected inside the automatic-detecting data about the comparison of internet articles as follow, so we might be able to recognize the same connoting, too.

5.2.2 Automatic-detecting data about the comparison of internet articles

When we use the automatic-detecting data to cancel the internet articles which correspondent to each other mutually, almost all the data could be deleted. For example, the mutually correspondent data about 23 “IG” internet articles are agree with each other. Only few exceptions are just like the above-mentioned 3 item.

For instance, there is correspondent data of IR25 in IR2, IR4, IR5, IR10 and IR44. However, there is no correspondent data of IR2, IR4, IR5, IR10 and IR44 in IR25. When we check the sentences between these internet articles, they show as the next quoting descriptions.

(IR2),(IR4),(IR5),(IR10),(IR44) 他在不知情的狀況下繼承了一枚戒指,卻發現這枚戒指是魔王遲遲不能統治世界的關鍵。

(IR25) 佛羅多發現這只戒指的製造者是黑暗君王索倫，而索倫正急著要把戒指找回去。因爲這只戒指正是魔王遲遲不能統治世界的關鍵,如果索倫得到這代表偉大邪惡勢力的魔戒，將使人民永遠在黑暗君王索倫的統治之下，而他統治的這片土地就是俗稱的中土世界.

- The fence part means the correspondence parts shown by our automatic-detecting data.
- The words in shadow mean the different words between two internet articles inside the correspondence part. (Hereafter, the same marks should be applied to the following quotation.)

By enumerating the sentences of these two groups, we could find out that IR25 just changed a word and increased a word. This increasing word might be the reason why IR2, IR4, IR5, IR10 and IR44 showed no corresponding data inside IR25, though there are data corresponding to IR25 inside these 4 articles respectively. The same contrast could also be recognized in the following two groups below.

(IR37) 這只戒指擁有無窮的神秘 力量，戒指原來是黑暗君王索倫所有的，卻意外地到了佛羅多手 裏。

(IR25) 這只魔戒具有無與倫比的力量，足以讓黑暗君王索倫席捲全世界並奴役所有生靈。戒指原來是黑暗君王索倫所有的，卻意外地到了佛羅多手裡。

(IC28),(IC49),(IC62) 他的夥伴紛紛到世界各地尋找他的下落，而最終來到了亞洲的麻六甲海峽，這次來自四面八方的海盜將史無前例的團結起來，砲口一致。

(IC33) 他們到世界各地尋找他的下落，途中他們到了新加坡，遇到精明幹練的中國海盜 - 嘯風船長，最後來到了亞洲的麻六甲海峽，這次來自四面八方的海盜將史無前例的團結起來，砲口一致。

With the enumeration of sentences, we found out that IR25 are longer than IR37 and IC33 are longer than IC28, IC49 and IC62. The quotations above proved that why there is no IR37 automatic-detecting data in IR25 and there are no IC28, IC49, IC62 data in IC33 due to the former sentences are connoted inside the later sentences according to the context.

Besides, it is obvious that internet articles plagiarized to each other. But it is hard for to make a conclusion about which one is the original one because we could not judge only from the length of different sentences. (A short sentence might delete some words from a long sentence. On the other hand, a long sentence might add some words from a short sentence. So it is not easy to judge the plagiarism unless we know the announced date of each internet article.)

Anyway, when we use the automatic-detecting data about the comparison of internet articles to confirm their corresponding situations, we found out that 21 “IG” articles, 37 “IR” articles, 36 “IC” articles and 17 “IU” articles are definitely plagiarizing to each other. That is according to the automatic-detecting data show that only 2 "IG" articles, 8 "IR" articles, 36 "IC" articles, and 15 "IU" articles have no corresponding data inside their own movie-category articles.

Also, we could tell from the auto-detecting data that IG22 is the most plagiarized one or the most plagiarizing one due to it was detected out for having most correspondences of other 11 internet articles. The same situation could be applied to IR26 , IR28(detected out for having most correspondences of other 15 internet articles), IC62, IC69 (detected out for having most correspondences of other 6 internet articles), and IU30 (detected out for having most correspondences of other 11 internet articles) as well.

All these correspondences between internet articles explain what caused the statistics of Table 3. That means when we use the automatic-detecting data about the comparison of student reports and internet articles to check the numerical values of the web-based articles sauces, the numerical values are far more than the results checked by human-hand due to internet articles plagiarized to each other. Thus, with this extra experiment, we might be able to collect lot more internet articles with different contents in advance without reading them one by one.

6. FUTURE ASSIGNMENT

By making use of the CKIP Chinese Word Segmentation System and Smith-waterman Algorithm, our research precisely discovered the plagiarized parts between Taiwanese student reports and Chinese internet articles just as human-judgment did. However, the same automatic-detecting method had been applied to detect copyright violation in our last Japanese essays [5] [6]. However, we indented to expand the direction of this essay into comparing Japanese student reports and Taiwanese student reports in the future. It is meaningful to work on with examining if the propriety of compatibility would be the same using our proposed automatic- detecting methods between two different languages.

REFERENCES

[1] Yahoo!奇摩,http://tw.yahoo.com/

[2] 中央研究院資訊科學所詞庫小組“CKIP Chinese Word Segmentation System” (中文斷詞系統) http://ckipsvr.iis.sinica.edu.tw/

[3] Robert W. Irving, “Plagiarism and Collusion Detection using the Smith-Waterman Algorithm”, *Technical Report 164*,Dept of Computing Science, University of Glasgow, pp1-21, 2004

[4] T.F.Smith, M.S.Waterman, “Identification of commonmolecular subsequences”, *Journal of Molecular Biology147*, pp19597, 1981

[5] 高島秀佳，坂口朋章，長尾壮史，石田崇，平澤茂一，“著作権侵害検出を目的とした類似文書発見手法”，経営情報学会，2006 年度秋季全国研究発表大会予稿集，pp58-61 ，2006

[6] 坂口朋章，雲居玄道，石田崇，平澤茂一，“著作権侵害検出のための剽窃 Web ページ発見システム”，経営情報学会，2007 年度秋季全国研究発表大会予稿集，pp454-457 ，2007