C-05

引用個所を考慮した剽窃レポートの検出システムの開発

Developing a Plagiarism Detection System based on Citations for Academic Reports

森　祐貴†　　谷川　佳延　　吉田　博哉†
Yuuki Mori　Yoshinobu Tanigawa　Hiroya Yoshida

1. はじめに

近年，大学では，授業で課されたレポート課題に対し，第三者が作成した文章をそのまま利用し，あたかも自分の意見のように文章を纏める剽窃行為の横行を問題視している．これらの行為は，著作権侵害に当たるきわめて悪質な行為であると言える．そのため，教員は，学生が提出したレポートに対し，剽窃行為が行われてないか精査すると共に，これらの行為が行われた場合，適切な処置を講ずる必要がある．一方，剽窃行為を特定するには，時間と労力がかかる事から，剽窃レポートの検出に関する研究が進められている．例えば，剽窃チェッカー[1]では，入力した文に対し，同一の文が Web 上に存在するかどうかを確認する機能を有している．他にも，学生間で剽窃行為が行われていないかを確認するために，レポート間の文章を比較し，文書間類似度を算出する方法として，文章の係り受け関係から判定する手法[2]や n-gram 手法[3]が研究されている．

ただし，これらの研究では，複数のレポートで同一の文献を引用した場合を考慮していない．レポートにおける引用とは，他人の著作を自身のレポートで紹介する方法であり，著作権法でも認められている合法な行為である．これらを考慮せずに文章を比較すると，類似度が高くなり，学生間の剽窃行為を判断し難い事が考えられる．そこで本研究では，Web ページ内の引用箇所を判別し，レポート間類似度を算出する際にそれらの箇所を除く事で，学生間の剽窃行為を特定するシステムを開発する．

2 . 本システムの概要

本研究では，引用箇所を考慮した剽窃レポート検出システムを開発する．剽窃レポート検出システムは，レポート管理システムと剽窃判定システムといった 2 つのサブシステムによって構成される．

2.1　レポート管理システムの概要

レポート管理システムは，1)年度，2)科目名，3)課題名，4)レポート群，といった情報を入力する事で，特定の年度で開講された授業における課題レポートとして，レポート群を管理する．**図 1** にレポート管理システムの全体像を示す．

2.1.1　レポート登録処理

レポート登録処理では，教員が指定したレポート群に含まれる文章をレポート DB に登録する．

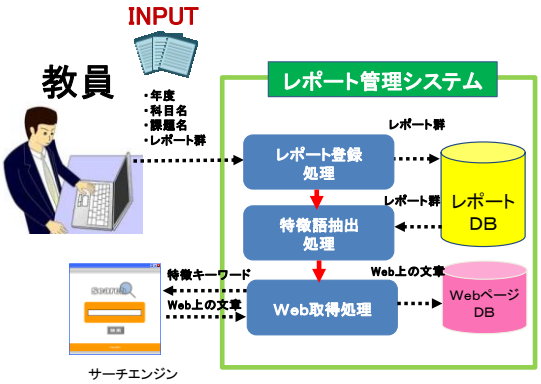

図 1　レポート管理システムの全体像

2.1.2　特徴語抽出処理

学生は，レポートを作成する際， Web ページを参考にする事が考えられる．そのため，レポート内で類似度の高い文章が存在する場合，それらは同一の Web ページを剽窃している可能性が高いと考えられる．そこで，特徴語抽出処理では，登録したレポート群から類似度の高い文章を抽出し，それらの文章群に対し，形態素解析を行い，特徴となる語句を抽出する．なお，文章を比較する方法は太田らの文書間類似度の計算式[2]を用いる．また，特徴語を判別する上で，式(1)の TF 値を算出する．

$$TF(\mathrm{t},\ \mathrm{d}) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}} \qquad (1)$$

$TF(\mathrm{t},\ \mathrm{d})$・・・文書d内のある単語tのTF値
$n_{t,d}$・・・ある単語tの文書d内での出現回数
$\sum_{s \in d} n_{s,d}$文書d内のすべての単語の出現回数の和

本研究では，形態素解析結果のうち，名詞と判定された形態素を特徴語候補として用いる．

2.1.3　Web 取得処理

Web 取得処理では，特徴語抽出処理で得られた特徴語をもとに，サーチエンジンを利用して Web 上の文章を取得し，Web ページ DB に登録する．

2.2　剽窃判定システムの概要

剽窃判定システムは，1)年度，2)科目名，3)課題名，といった情報を入力する事で，特定の課題に対して提出されたレポート群に対する剽窃判定を行う．**図 2** に剽窃判定システムの全体像を示す．

2.2.1 対象レポート取得処理

対象レポート取得処理では，指定された課題に対するレポート群の中から，比較元となる対象レポートを任意に選択する．

† 神戸情報大学院大学　情報技術研究科
Kobe Institute of Computing;
Graduate School of InformationTechnology

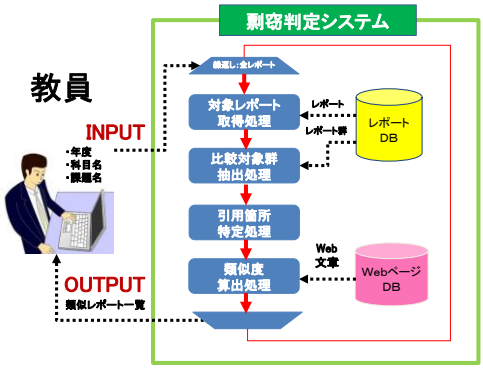

図 2　剽窃判定システムの全体像

2.2.2 比較対象群抽出処理

比較対象群抽出処理では，指定された課題に対するレポート群の中から，対象レポートを除くレポート群を比較レポートとして取得する．

2.2.3 引用箇所特定処理

引用箇所特定処理では，レポート内の文章のうち，引用ルールに則って記載されている文章を引用箇所として扱い，対象レポートおよび比較レポートから取り除く．

2.2.4 類似度算出処理

類似度算出処理では，引用箇所特定処理によって取り除かれた文章をもとに，1)Web ページからの剽窃判定，2)レポート間の剽窃判定，といった流れで処理を行う．まず，Web ページからの剽窃判定は，対象レポートと Web ページ DB に格納された文章の類似度を算出し，閾値以上であれば剽窃であると判定し，対象レポートから当該文章を取り除く．なお，類似度は，文章を構成するキーワードの係り受け関係の一致率から算出する．そして，レポート間の剽窃判定は，前述の処理によって取り除かれたレポート文章間の類似度を算出し，閾値以上であれば剽窃であると判定する．なお，本研究では，太田らの文書間類似度の計算式[2]を用いる．

3. 実証実験

3.1　実験目的

本システムでは，Web ページの剽窃を特定するために，当該レポート群が参照している可能性の高い Web ページをデータベースに蓄積する必要がある．そこで，本実験では，特徴語抽出処理によって抽出された特徴語が，Web ページの剽窃判定に有効であるかどうかを確認する．

3.2　実験方法

本実験では，1) レポート群の全文章から特徴語を抽出する単純抽出方法，と 2) レポート管理システムにおける特徴語抽出処理によって特徴語を抽出する本研究抽出方法，の 2 つを比較した．その際，式(1)に示す TF 値の高い語句を特徴語として抽出する．そして，それらの特徴語全てを利用し，サーチエンジンで検索した結果の上位に現れた Web ページをデータベースに登録した．最後に，剽窃判定システムにおける類似度算出処理によって，Web ページからの剽窃と判定され一致した文章の割合を一致率として求め，比較した．

なお，本実験では，ある授業で課されたレポート課題に対し，提出のあった 125 人分のデータを利用した．また，TF 値の高い語句 5 つを特徴語として抽出すると共に，それらを用いてスペースで区切り，google サーチエンジンで検索した結果，上位 5 件の Web ページの文章をデータベースに登録した．

3.3 実験結果と考察

単純抽出方法によって得られた特徴語と本研究抽出方法によって得られた特徴語の比較を**表 1** に示す．

表 1　得られた特徴語の比較

順位	単純抽出方法		本研究抽出方法	
	特徴語	TF 値	特徴語	TF 値
1 位	GPS	0.038810	衛星	0.050392
2 位	衛星	0.032766	0	0.035645
3 位	こと	0.023002	こと	0.029990
4 位	0	0.021889	距離	0.028795
5 位	1	0.017844	受信	0.028271

また，得られた特徴語をもとに構築した Web ページ DB を用い，剽窃判定システムにおける類似度算出処理によって，Web ページからの剽窃と判定され一致した文章の一致率を比較した結果を**図 3** に示す．

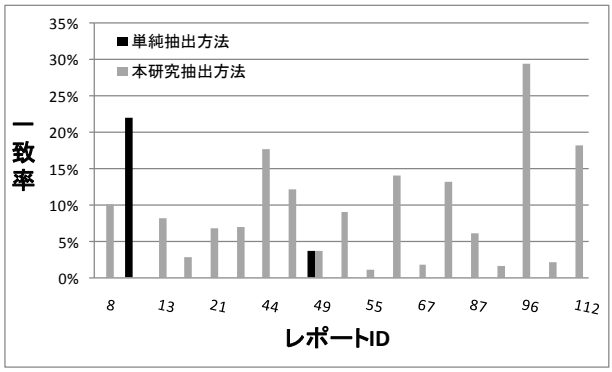

図 3　一致率の比較

図 3 に示す通り，本研究抽出方法により取得した Web ページを利用した方が，一致する文章が存在するレポートの件数が多い結果となった．よって，本研究抽出方法は，レポート作成者の多くが使用した Web ページを特定する上で有効であると言える．一方で，「0」や「こと」といったレポートとは関係無い特徴語が含まれていたため，これらを省く仕組みを検討する必要がある．

4. おわりに

本研究では，引用個所を考慮した剽窃レポートの検出システムの開発し，そのうちの特徴語抽出が Web ページの剽窃判定に有効であるかを実験し，その有効性を確認した．一方，抽出した特徴語は，レポートとは関係の無い語句が確認され事から，より精度の高い特徴語を抽出出来る仕組みを検討する．

参考文献

[1] 論文チェッカー, http://plagiarism.strud.net/
[2] 太田貴久, 増山繁, “模倣レポート判定に用いる文書間類似度の考案”, 言語処理学会年次大会発表論文集, pp.A10B6-03, 2004.
[3] 小高知宏, 村田哲也, 高建斌, 諏訪いずみ, 白井治彦, 高橋勇, 黒岩丈介, 小倉久和, “n-gram を用いた学生レポート評価手法の提案”, 電子情報通信学会論文誌 D, Vol.J86-D1, No.9, pp.702-705, 2003.