O-051

剽窃レポートの検出システムの提案

A Proposal for Detecting Plagiarism in Academic Reports

森　祐貴†　吉田　博哉†
MoriYuuki　YoshidaHiroya

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，n-gram 手法[2]や文章の係り受け関係から判定する手法[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，類似度が高くなる事が考えられる．そこで本研究では，引用箇所を判別し，それらの文章を除いた上でレポート間を比較する剽窃レポート検出システムを提案する．

2. 提案システムの概要

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提案する．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．本システムでは，以下に示す処理によって構成される．

- レポート登録処理
- 特徴語抽出処理
- Web 情報取得処理

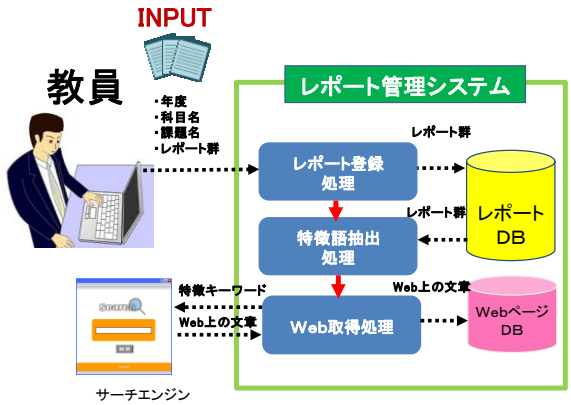

図 1　レポート管理システムの全体像

2.1.1 レポート登録処理

指定されたレポート群に含まれる文章をレポート DB に登録する．

2.1.2 特徴語抽出処理

登録したレポート群から特徴語を抽出する．

2.1.3 Web 情報取得処理

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

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．本システムでは，以下に示す処理によって構成される．

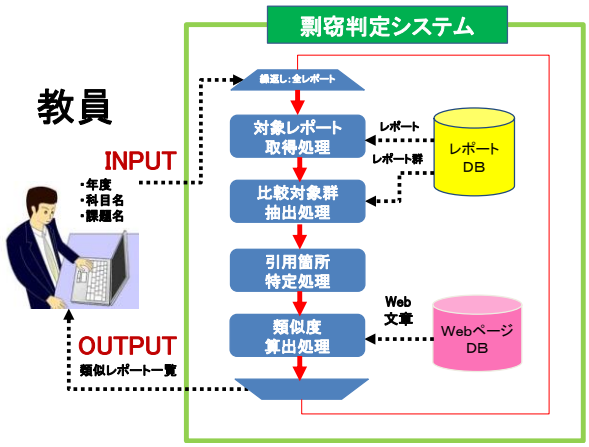

図 2　剽窃判定システムの全体像

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

- 対象レポート取得処理
- 比較対象群抽出処理
- 引用箇所特定処理
- 類似度算出処理

2.2.1 対象レポート抽出処理

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．

2.2.2 比較対象群抽出処理

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

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

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）文書間類似度の算出，といった流れで実施する．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．その後，取り除かれた文書間の類似度を算出する．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

本システムの有効性を確認するために，引用箇所特定処理を実行し，本来の剽窃行為が明らかとなるような類似度を算出しているかどうかを確認する．

3.2 実験方法

本実験では，引用箇所の文章を取り除いた結果をもとに，類似度を算出した場合と，引用箇所を考慮しない提出レポートの類似度を算出した結果を比較する．本実験で使用するデータは，ある授業で課されたレポート課題に対し，提出のあった 30 人分のデータを利用した．また，提出されたレポートは，引用ルールに則って記載されていないものが多いため，本実験では Web ページに記載されている文章と一致した場合，引用箇所であると判定した．なお，引用判定に利用した Web ページは，当該レポートのテーマから任意でキーワードを抽出し，サーチエンジンで検索した上位 3 ページを利用した．

3.3 実験結果と考察

提出レポートをもとに算出した類似度，および引用箇所の文章を取り除いた文書をもとに算出した類似度の差を**図 3** に示す．また，比較レポートの例を**表 1** に示す．

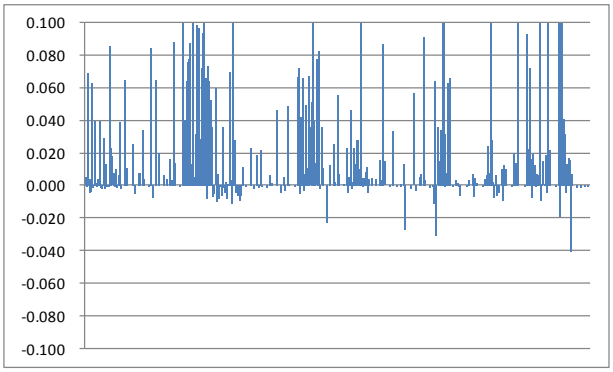

図 3　算出した類似度の差

表 1　比較レポートの例

レポート番号	文章の数	削除後の文章数	文章を削除した割合	類似度(削除なし)	類似度(削除あり)	類似差
5	97	92	5%	0.494387837	0.274143106	0.220244731
23	97	37	62%			
24	104	85	18%	0.404597482	0.445506751	-0.04090927
29	75	74	1%			
2	42	42	0%	0.282521657	0.282521657	0
3	89	89	0%			

図 3 および**表 1** に示す通り，類似度の差が正の数値を示している場合は，引用と判定された部分がレポート間の類似度を求める際に大きく影響していることがわかる．一方，類似度の差が負の数値を示している場合は，引用とみなした文章を削除したことにより，類似度が高くなっていることがわかる．このように，引用と判定した箇所を取り除いた上でも，文書間類似度が高い場合は，学生間で剽窃行為が行われた可能性があると考えられる．そのため，本来の剽窃行為が明らかとなる事から，本システムが有効であると言える．

4. おわりに

本研究では，引用箇所を考慮した剽窃レポート検出を目的に，引用箇所特定処理を実行し，本来の剽窃行為が明らかとなるような類似度を算出しているかどうかを確認するための実証実験を行った．その結果，引用箇所の文章を取り除いた結果をもとに，類似度を算出した場合と，引用箇所を考慮しない提出レポートの類似度を算出した結果には差が見られたことから，引用箇所を考慮し，剽窃レポートを検出することが有効であるということがわかった．

今後は，引用箇所を特定する方法について検討するとともに，Web ページのサイト数，検索ワードの妥当性を検討し引用箇所を考慮した剽窃レポート検出システムの開発を進める．

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
[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