様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 5 月 28 日現在

研究種目：基盤研究(C)(一般)
研究期間：2007—2008
課題番号：19500790
研究課題名（和文）レポート・小論文に対するWWWページからの剽窃の自動検出に関する研究
研究課題名（英文）Research on Plagiarism Detection for Reports and Essays Imitated from WWW pages

研究代表者 湯川 高志（YUKAWA TAKASHI）
長岡技術科学大学・工学部・准教授
研究者番号：70345536

研究成果の概要：電子的に作成されたレポートや小論文において，他者のレポートや WWW 上のページの記述からの剽窃が大きな問題となっている．特に，WWW からの剽窃に対しては，レポート間の模倣度合いの同定に加えて，剽窃の元となった WWW ページをいかに見つけ出すかも課題である．本研究では，剽窃のチェック対象であるレポートや小論文から WWW ページを検索するためのクエリを生成し，全文検索サービスにより検索されたページから剽窃元である可能性の高いページを選択し，文書間剽窃度合を計算する手法を考案し，ｅラーニング向け剽窃検出システムの実現を試みた．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,700,000	510,000	2,210,000
2008 年度	1,700,000	510,000	2,210,000
年度			
年度			
年度			
総　計	3,400,000	1,020,000	3,502,000

研究分野：総合領域
科研費の分科・細目： 科学教育・教育工学，教育工学
キーワード：ｅラーニング，テキスト処理

１．研究開始当初の背景

近年，情報技術の発達により電子的な情報のやり取りが容易になってきている．インターネット上の文章をコピー＆ペーストして簡単に窃用することが可能となってきた．さらにそれら情報をネットワークを介して送受信することや補助記憶装置に格納して持ち運ぶことも容易になったため，学生は手間をかけずに，レポート・小論文を簡単に模倣するようになった．レポート・小論文は大学などの授業の課題の提出方法としてよく用いられているため，模倣は昔より大きな問題となっている．

コンピュータのない時代は，学生が行っていた模倣手法とは，図書館で学術書を抜き書きすることであった．それも許されることではないが，それでも，抜き書きという手作業を通じて勉強にはなった．ちょうど読書ノートを取るようなスピードで本を再読するようなものである．ところが，インターネットの時代になって，学生は web ページからコピー＆ペーストして，簡単にレポート・小論文を作成するため，その知的レベルの衰退は顕著である．さらに，模倣レポートが巧妙化しており，公平な採点を行うことが，教師にとって極めて困難となってきた．

模倣の程度を自動的に判定するシステムを作ることができれば，教師の仕事量を減らすだけでなく，学生が他人の文章を模倣する行為をある程度予防することが可能である．

従来の模倣レポート検出は，過去に他の学生が書いたレポートからの模倣の検出に主眼をおいていた．従来は学生が，一つの元文書から，文を取捨選択したり，１文を分割/結合したり，文節を換言したりするなどの改変技法を用いて，それに自分のオリジナル文を少々加えて模倣レポートを作成していた．

ところが，最近の模倣レポートは

(1) 複数文書からの繋ぎ合せ
(2) 模倣の手法が巧妙化
(3) web ページを模倣元としている

という特徴がある．そのような模倣レポートの模倣関係をすべて発見するのは，従来の模倣検出手法は対応できない．そして，これに対応できる模倣検出システムが望まれる．そのためには，複数 web ページからつなぎ合せて，それに巧妙な改変技法を用いて作成したレポートの検出と web ページから元文書の検出という２つの技術を確立する必要がある．

２．研究の目的

学生が作成する模倣レポートを検出するために，以下の特徴を持つ模倣レポートを検出できる技術を実現する：

(1) web ページを元文書として
(2) 複数の文書からの繋ぎ合せ
(3) 巧妙な文の改変を加えた

レポート．

そのために，システム全体の構成を考案し，元となる web ページの検索技術と巧妙な改変に耐性を持つ文書間模倣検出技術を確立することが必要となる．

３．研究の方法

本研究は次のようにな手順で実施した．

まず，tf・idf を用いたベクトル空間モデルと smith-waterman アルゴリズムを複合した従来の模倣検出手法に対して，検出性能の評価を行う．

次に，誤検出に着目して，改良された模倣検出手法を提案し，その手法について評価を行う．

さらに，文書からキーワードの自動抽出手法を提案し，その手法について評価を行う．

最後に，文書間の模倣検出と web からの元文書の取得との両手法を適用した模倣レポート検出システムの総合性能について評価を行う．

４．研究成果

(1) 模倣レポート検出システムの構成

本研究は，全体システムをまず提案し，個々の要素技術もより高度なものを考案するという流れで研究を進めた．模倣レポート検出システムの構成の概略図を図１に示す．

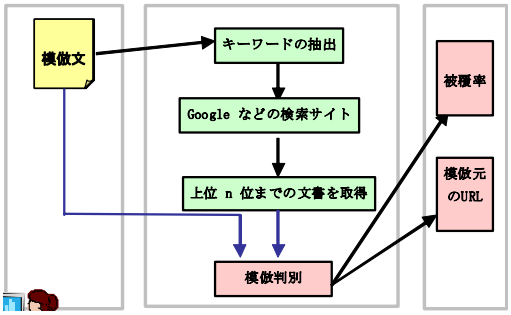

図１　模倣検出システムの構成

提案した模倣レポート検出システムは，模倣文を入力とし，その模倣文から自動的にキーワードを抽出して検索サイトにかけて，上位 n 位までを元文書プールとして取得する．次に，それらの元文書プール内の文書と模倣文とを模倣判別して，被覆率（元文書から切り取った文の割合）と模倣元の URL をユーザに示す．最終的に模倣文がその模倣元となる web ページから実際に模倣されたものであるかを判断するのは人手にまかされることになる．従来の模倣検出システムは，模倣元の URL の代わりに，模倣文と元文書間の類似部分をユーザに提示する．これでは，何通の元文書から模倣したのかが分からない．本システムでは，それに対応できるように，模倣元の URL をユーザに示す．

①文書間の模倣検出

学生が模倣する際，文単位の変換としての文の再構成や文の追加などの技法(レベル３)をよく用いる．このような文の検出は単語を処理単位としては対応しにくいという問題点があるため，本文研究では，文を処理単位として文書間の模倣関係を発見する．このため文書間の模倣検出は文と文の模倣検出をベースとすることになる．また，複数文書からの繋ぎ合せ，巧妙な模倣手法，web ページからの元文書の取得という特徴を持つレポートの模倣検出を目指す．まず，tf・idf を用いたベクトル空間モデルで文間類似度を計算し，表層的に模倣した文を模倣文として検出し，その後全体を解析し巧妙な模倣手法を用いた文を検出する．

模倣レポート中の文と元文書中の文の模倣関係を発見した後，元文書から切り取った文の割合を計算し，あらかじめ設定した閾値に超えたら，文書と文書は模倣していると判断する．

② web からの元文書の検索

学生が作成したレポートは模倣したかどうかを判断するため，模倣元となる web ページを見つける機能が必要となる．現在，web 上には誰でも無料で利用できる非常に高性能な検索サイトがある．模倣レポートから適切な検索キーワードが抽出できれば，検索結果として得られるのはその模倣レポートに関する情報である．

検索キーワードから考えると，学生が web 上の検索サイトを用いて模倣元となるページを入手する場合には，必ず作成するレポートの内容に関連するキーワードを検索に用いている．逆に，学生が模倣を行い作成したレポートの内容に直接関連するキーワードを検索に用いれば，模倣元となる web ページも大量に見つけることができる．このため，模倣レポートから検索キーワードを抽出することが可能であると考えられる．

(2) 誤検出に着目した模倣検出手法

本研究では巧妙に改変を加えた模倣レポートにも正しく対応できる手法を提案した．本手法の処理プロセスを図 2 に示す．

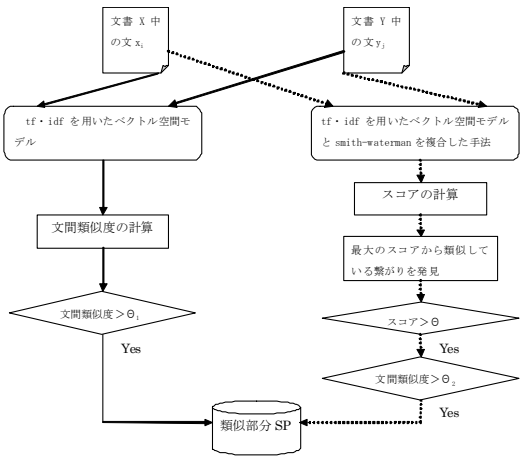

図 2　提案手法のプロセス

ここで，類似部分 SP における文組は最終の模倣文組である．これは二つの模倣検出手法によって得られる．1 つの手法は tf・idf を用いたベクトル空間モデルであり，個々の文の模倣関係を発見することで，文間類似度がある閾値 Θ1 より大きいとき，類似部分 SP に加える．もう一つの手法は tf・idf を用いたベクトル空間モデルと smith-waterman アルゴリズムを複合した手法である．最大のスコアから類似している繋がりを見つけて，その繋がりをチェックするため，スコア＞Θ，かつ文間類似度＞Θ2 の時の文組を類似部分 SP に加える．このチェックによって，文と文の間に独自の文を挿入する文組の模倣関係は独自の文を検出しなく，文と文の模倣関係を正しく検出できることとなる．

ここで重要なのは 3 つの閾値である．Θ は従来手法の閾値と同じである．Θ1 は独立類似している文を検出するため定める．Θ2 は連続類似する文組における類似してない文組を検出しないために定める．Θ2 は Θ1 の設定に応じて変化させる必要がある．

① 個々の文の模倣関係の発見

本手法では tf・idf を用いたベクトル空間モデルで文間類似度を求める．初めに，文書における単語を重みつける．次に，以下の計算式で文間類似度を計算する．

$$s(x_i, y_j) = \frac{V(x_i) \cdot V(y_j)}{|V(x_i)||V(y_j)|}$$

ここで，V(xi) ，V(yj) はそれぞれ文 xi と文 yj に対する tf・idf 値を要素としたベクトルを表す．

以下に，文 xi と文 yj の類似度 s(xi,yj) の満たすべき制約条件を示す．

$$0 \le s(xi,yj) \le 1$$
$$s(xi,yj) = s(yj,xi)$$

ここで，文間類似度 s(xi,yj) は，値が大きいほど文 xi と文 yj が類似しているものとする．

あらかじめ閾値 Θ1 を定めて，文間類似度 s(xi,yj) が Θ1 を超える場合，文 xi と文 yj は模倣の可能性があると判断する．

図 3 は文間類似度の分布図である．横軸は文書 X（模倣レポート）中の文の文番号で，縦軸は文書 Y（模倣元）中の文の文番号である．各枠は文書 X 中の文 xi と文書 Y 中の文 yj の類似度と示す．しかし，tf・idf を用いたベクトル空間モデルは黒い枠が対応する文組のような巧妙化な分割/結合した文などの検出が対応できないので，他の手法と併用する手法を提案した．

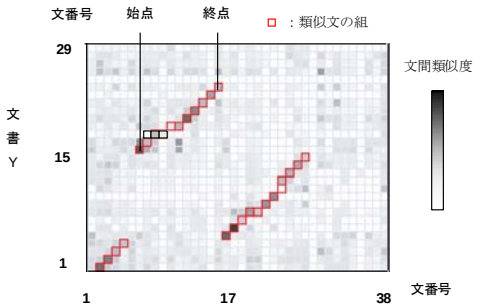

図 3 文間類似度の分布

② 巧妙な改変を施した文組の発見

提案する手法は太田らによって提案されたアルゴリズムをベースとした改良手法である．スコアの求め方は従来の手法と同様であり，スコアから類似部分の発見する部分が従来手法とは異なる．提案手法の処理手順を以下に記す（ Step1～Step4 は従来の手法と同様である）．

【提案手法】

Step1 tf・idf を用いたベクトル空間モデルで文書 X 中の文 xi と文書 Y 中の文 yj の類似度を計算する．

Step2 文間類似度を 2.4 式に代入し，各文組のスコアを計算する．

Step3 スコア Si,j＞0 となる文組（ xi ，yj ）の集合を終点候補集合の初期値とする．

Step4 終点候補集合から最大のスコアを見つけ，そこから終点として追跡を行い類似部分を決定する．(追跡を行う手法は従来手法と同様である．)

Step5 Si,j＝0 または s(xi,yj)- Θ１＞0 の時，終了．

Step6 Si,j＞0 かつ s(xi,yj) ＞Θ2 の時，模倣文組とする ．

Step7 終点から始点までの類似部分の長さが 3 より大きいとき，この類似部分を類似部分集合 SP に加える．

Step4～Step7 により，模倣と見られる連続類似する文組が 1 組得られる．この操作を複数回繰り返すことで 2 文書間のすべての連続類似した文組が得られる．この結果得られた連続類似した文組 (SP) は模倣箇所と見なすことができる．

Step6 において Si,j＞0 かつ s(xi,yj) ＞Θ2 の時，模倣文組としている．この意味は，模倣文組の決め方としてスコアではなく，ある程度文間類似度が高いことも類似した文組と見なす条件としている．このステップは，スコアが高くて文間類似度が非常に低い場合の模倣検出をしないためのものである．

(3)キーワードの自動検出手法

模倣されたレポートからキーワードを抽出し，それを web 検索サービスに入力することで，模倣元 web ページを見つける．適切なキーワードを抽出することができれば，模倣元ページをランキングの上位に出現させることが可能である．逆に不適切なキーワードであれば，模倣元ページは，ランキングの下位出現するか，あるいは，まったく出現しない．そこで，適切なキーワードの抽出が重要となる．

① キーワード

検索キーワードとは，欲しい情報に関連する「単語」「単語と演算子から成る論理式」のことである．「Google」「Yahoo」「goo」などの検索サイトは，Web ページの検索機能を提供している．一般な検索サイトは AND・OR 検索の機能を持っている．ブーリアンモデルでは，web ページに指定したキーワードが含まれており，論理式が真になれば，その web ページは検索結果として表示される．複数のキーワードを用いて検索する場合は，検索サイトが持つ AND・OR 検索の機能による指定ができる．全てのキーワードを含む web ページが欲しい場合，キーワードを「AND」で結合し，複数のキーワードを全て含んでいる web ページを検索することができる．いずれかのキーワードを含む web ページが欲しい場合には，キーワードを「OR」で繋げることにより，複数指定されたキーワードのいずれかを含んでいる web ページを検索することができる．そこで，この AND・OR 検索機能を用いて模倣元となる web ページの検索を行う．

しかしながら，適切な模倣元となる web ページを検索するには，検索結果を絞り込むことを考える必要がある．一般的に検索サイトに単語を一つのみ与えた場合には，模倣元となる web ページ以外のものが数多く検索されてしまうため，なるべく長い検索ワードを抽出するほうがよい．また，模倣レポートに出現する特徴的な単語と出現頻度が高い単語がキーワードとなる可能性が高いと考えられる．そのため，本提案では検索サイトが持つ AND・OR 検索の機能を用いて，模倣レポートに出現する特徴的な単語と出現頻度が高い単語をキーワードとして検索を行う．

② キーワードの自動抽出手法

特徴的な単語と出現頻度が高い単語が内容を表す可能性が高いこと，キーワードが長い方が検索結果を絞り込めることの両者を勘案し，キーワードの文字長と出現頻度との両方を考慮したキーワードの抽出手法を提案する．本手法は，以下に示す抽出法 1 と抽出法 2 によってキーワードを抽出する．また，抽出手法の概略図を図 4 に示す．

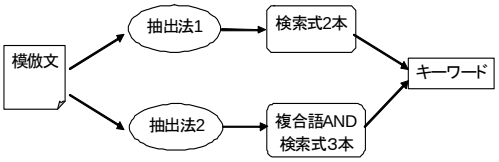

図 4 キーワード抽出の概略図

【抽出法１】

模倣レポートにある単語を出現頻度によってランキングし，上位３位までの単語集合 {w1 , w2 , w3} を論理演算式で結合する．抽出した３つの単語を下記の２種類の論理演

算式に当てはめたものをキーワードとする．
(1) w1 + w2 + w3
(2) w1・w2・w3

【抽出法２】

模倣レポートに出現する特徴がある単語に着目する手法である．下記の文に出現する名詞を抽出する．
(1) 表題文
(2) 見出し
(3) 定義文（「とは」,「では」,「と言う」などのような特定の助詞の前の言葉）
(4) 主題になる文

上記４つの文に出現する名詞を抽出し，これらの単語を出現頻度でランク付けし上位５個を単語集合｛w1 , w2 , w3 , w4 , w5｝として抽出する．その５個の単語から文中左右に出現する名詞を探し，単語を組み合わせる．組合せ方を下記に示す．
(1) 模倣レポートを n の単語列 L=(a0 , a1 , a2 , ・・・ ,an-1) に分割する．
(2) w1 と一致する aj から左右に連続する k 個の要素を「AND」で結合し，左右における単語が名詞ではないとき，終了．
(3) 得られた単語列を，候補キーワードとする．
(4) w2 ～w5 に対し (2) を繰り返す．

次に下記の計算式で各候補キーワードのスコアを計算する．得られたスコアによってランク付けし，その上位３位をキーワードとする．

$$Skeyword = (tf1 + tf2 + \ldots + tfm) \times length(s)$$

ここで，tfi は候補キーワード集合における i 番目の単語の出現頻度である. length(s) は候補キーワードの長さである．
ただし，
tfi ≥ 2 (キーワードの出現頻度 ≥ 2)
length(s) ≥ 2 (キーワードの長さ ≥2，一つの単語はキーワードにならない)

この２つの抽出法の組み合わせにより，キーワードの文字列長と出現頻度との両者を考慮することになるので，適切なキーワードの抽出でき，それにより検索結果を絞り込めると期待できる．

③ 性能評価

キーワード抽出システムがどの程度適切なキーワードを抽出することができるかについて，評価を行った．模倣レポートには，１つの web ページから模倣したレポートもあり，２つの web ページから模倣したレポートもあり，３つ web ページから模倣したレポートもある．それらの模倣レポートには，模倣元となる URL が記録してある．

キーワード抽出システムによって模倣レポートからキーワードを抽出し，それを検索サイトに入れて検索し，得られた結果の上位 100 位を模倣元文書プールとする．模倣文に記録してある URL が模倣元文書プールにある文書の URL と一致すれば, 正解とする．正答率を表１に示す．

表１　キーワード抽出の正答率

模倣元ページ数	正答率	
1	36%	
2	63%	54% (１ページ正解)
		9% （２ページ正解)
3	64%	40% (１ページ正解)
		21% (２ページ正解)
		3% （３ページ正解)

表１から１つのページから模倣したレポートに対する正答率が最も低いことが分かる．3 つのページから模倣したレポートに対して，模倣元ページを少なくとも１ページ正しく検索できた比率は約 64% であり，１ページからの模倣や２ページからの模倣の場合より高くなっている．しかし，３ページとも正しく検索できたのは 3%程度しかない. 2 つのページから模倣したレポートに対しては 2 ページとも正しく検索できたのは 9%程度である．結果から見ると，模倣となるページ数が多いほど，模倣元ページを１ページ以上含む正答率が高いことが分かる．

(4) システムの総合性能の評価

システムが，模倣レポートと模倣していないレポートを正しく判定できるかについて，テストセットにおける模倣レポートと模倣していないレポートを用い，それぞれ評価を行った．

本評価は, web からの元文書取得と模倣検出機能との総合評価である．模倣レポートから抽出したキーワードによって検索された上位 100 位までの web ページの内容を元文書として，それとレポートとの模倣関係を発見する．また，模倣か否かは被覆率で判断する．さらに，レポートと元文書の被覆率が閾値を超え模倣であると判定された場合には，元文書の URL が模倣レポートに記録された URL と一致するかも調べる．

評価結果を表 2 に示す．表 2 から１つのページから模倣したレポートに対する正答率が最も低いことが分かる．3 つのページから模倣したレポートに対して，模倣元ページを少なくとも１ページ正しく検索できた比率は約 56%であり，１ページからの模倣や 2 ページからの模倣の場合より高くなっている．しかし，3 ページとも正しく検索できたものは

ひとつもない．2 つのページから模倣したレポートに対しては 2 ページとも正しく検索できたのは 3%程度しかない．結果から見ると，模倣していないレポートが正しく検出できて，また模倣となるページ数が多いほど，検出が高いことが分かる．上記のすべての場合(模倣元ページ数 1,2,3)についての平均を取ると，模倣レポートを自動検出した総合性能としては 48%程度であった．

表 2 システムの総合正答率

模倣元ページ数	模倣レポート判定の正答率	
0（模倣していないレポート)	85%	
1	36%	
2	53%	50%（1 ページ正解)
		3%(2 ページ正解)
3	56%	46%(1 ページ正解)
		10 %(2 ページ正解)
		0% （3 ページ正解)

提案手法は期待通りに模倣レポートを判断することが分かった．しかし，キーワード適切さが大きく総合性能に影響している．複数のページから模倣した場合，模倣レポート判定の性能はキーワード抽出より低くなっている．その原因は被覆率の計算式が以下のことに対応できないことによるものである．

(1) 各 web ページから少しずつ文を取って繋ぎ合わせる．
(2) web ページにおける文数が非常に膨大である．

これらの場合に，模倣レポートと各 web ページの被覆率が低くなる．性能を向上させるため，適切な閾値を自動的に決定すること，被覆率の計算式の改良，キーワードを抽出する際，単語の出現頻度と複合語の長さを組み合わせて複合語を生成する手法の改良が必要であり，これらが今後の課題である．
また，3 通の模倣していないレポートを模倣レポートとして検出してしまった．これは，レポートに含まれる個々の文が元文書プール内の多数の web 文書に分散して類似しており，結果として被覆率が高くなってしまったものである．個々の文がいずれかの web ページ内の文と類似することはオリジナルなレポートでも避けられないため，これは明らかに誤検出である．模倣元文書の数を制限することにより，このような誤検出を減らすことが可能と考えられる．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）
〔学会発表〕（計 4 件）

[1] T.Yukawa, H.Takahashi, Y.Fukumura, M.Yamazaki, T.Miyazaki, S.Yano, A.Takeuchi, H.Miura, N.Hasegawa: Implementing e-Learning Technology for Project-Based Learning for the Development of Embedded Software, SITE2009, 2009 年 3 月 2 日-6 日, アメリカ・チャールストン.
[2] 鈴木結，S.Tansuriyavong，湯川高志，福村好美：e ラーニングにおける“つながり感”醸成ツールの効果検証，第 9 回CMS 研究会，2008 年 12 月 11 日-12 日，九州工業大学戸畑キャンパス・西日本総合展示場.
[3] T.Yukawa, K.Kawano, Y.Suzuki, S.Tansuriyavong, Y.Fukumura: Implementing a Sense of Connectedness in e-Learning, ED-MEDIA2008, 2008 年 6 月 30 日-7 月 4 日, オーストリア・ウィーン.
[4] 徐敏，湯川高志：改変に対して頑健な模倣文検出手法とその評価，平成 19 年度電子情報通信学会信越支部大会，2007 年 9 月 29 日，長野工業高等専門学校.

〔図書〕（計 1 件）

[1] T.Yukawa, Y.Fukumura: Chapter “Intelligent Interaction Support for e-Learning” in “E-Learning”, In-Tech, 2009.

６．研究組織
(1)研究代表者
湯川 高志（YUKAWA TAKASHI）
長岡技術科学大学・工学部・准教授
研究者番号：70345536

(2)研究分担者
なし

(3)連携研究者
なし