

複数の Web ページから剽窃したレポートの発見支援システムの実装

上野 修 司[†] 高橋 勇^{††} 黒岩 丈 介^{††}
白井 治 彦[†] 小高 知 宏^{††} 小倉 久 和^{††}

本研究では、我々は複数の Web ページから剽窃した学生レポートを発見するためのシステムを提案する。このシステムを実現するために、我々は検索エンジンを用いて学生レポートの剽窃元になった Web ページの検出と剽窃の可能性を評価する際の 2 つの問題を解決した。また、これらの新しい手法を用いたシステムを実装し、システムの有効性を検討するため 2 つの実験を行った。その結果、複数の Web ページから剽窃したレポートの発見支援が可能であることが示された。

Implementation of a support system to find out of the report plagiarized from several Web pages

SHUJI UENO,[†] ISAMU TAKAHASHI,^{††} JOUSUKE KUROIWA,^{††}
HARUHIKO SHIRAI,[†] TOMOHIRO ODAKA^{††} and HISAKAZU OGURA^{††}

In the present investigation, we propose a system in supporting faculties to find out the learner's report plagiarized from several WEB pages. In order to realize the system, we have solved mainly two problems: (i) How to find out original WEB pages from which learners plagiarized by means of a certain WEB search engine?, and (ii) How to evaluate possibility of plagiarism? We implemented the new two algorithms in the system, and performed two experiments to show the their effectiveness. From the results, we have almost succeeded to find out the reports plagiarized from several WEB pages except for a few reports.

1. はじめに

近年、情報技術が発達し、パソコンが広く普及したことによって、個人でパソコンを所有することも一般的になった。そのため、学習者がパソコンを介して、Web 上から容易に様々な情報を手に入れるようになった。それに伴ない、学習者同士の剽窃によりレポートを作成するだけでなく、Web ページのコピー&ペーストによってレポート課題を作成する学習者が増えている。このような剽窃行為は、学習者の学習機会を奪うだけでなく、正しい成績評価の妨げになる。さらに、著作権の侵害にあたるという問題もある。また、海外でもこのような剽窃行為が問題となっている。¹⁾

このような剽窃行為により、作成されたレポートを確認する際、教師は剽窃の可能性のあるレポートについて、それらの文章がどこから剽窃されているか Web 検索をかけ、多くの Web ページの中から剽窃元となった Web ページを見つけ出さなければならない。剽窃行

為が発覚しないよう文章改編をしたレポートを提出された場合、剽窃元を見つけ出すことがより困難になり、確認作業にも時間がかかる。そのような剽窃レポートをチェックする作業は教師にとって大きな負担となる。

このような剽窃を発見するために、様々な研究が行われている^{2)~7)}。例えば、概念辞書を用いて文章の深層情報を利用して解析する剽窃発見手法⁶⁾や英国の Turnitin UK というオンラインの剽窃探知ソフトウェア⁷⁾などである。しかし、これらは実運用においては学生レポート同士の比較や、論文などデータベースに蓄積されたデータを主に対象としており、頻繁に更新される Web ページのような対象を剽窃する行為には十分に対応できていない。

我々はこれまでに Web ページから剽窃された可能性のあるレポートを選出し、剽窃元の候補となる Web ページを自動で検出する剽窃チェック支援システム⁸⁾を提案した。この既存のシステムは、レポートから検索ワードを抽出し、検索エンジンを用いて収集した Web ページとレポートの文章の表層情報を解析し、その類似性を評価することで、剽窃の可能性が高いレポートを検出する。

本研究では、既存のこれらのシステムでは検出することができなかった複数の Web ページを検出する手

[†] 福井大学

Faculty of Engineering, University of Fukui

^{††} 福井大学大学院工学研究科

Graduate School of Engineering, University of Fukui

法と、それらの Web ページとレポートとの間で剽窃の可能性を評価する手法を提案し、その手法を用いたシステムを実装する。また、実際に大学生から提出されるレポートを対象とした実験を行い、システムの有効性の検証を行う。

本稿の 2 章では我々がこれまでに提案したシステムの手法を紹介し、複数の Web ページからの剽窃を検出する際の問題とそれを解決する手法を提案する。3 章では提案した手法を用いたシステムの実装について説明し、4 章では本システムを利用したシミュレーション実験、実際の学生レポートを対象とした実験を紹介する。5 章は 2 つの実験についての考察、6 章は本稿のまとめである。

2. 複数からの剽窃を検出する手法の提案

ここでは、複数 Web ページからの剽窃を検出する手法を提案する。本手法では、基本的には我々がこれまでに提案した手法^{*)}を利用する。まず、その手法の概要を示す。次に、一般的に剽窃を検出する手法を用いて複数の Web ページから剽窃したレポートを検出する際の問題点を挙げ、最後に、それを解決する手法を提案する。

2.1 剽窃検査システムの検査手法

このシステムでは、まず、レポートとの比較を行う剽窃元候補となる Web ページを収集するために、Web 検索に用いる検索ワードを抽出する。レポートの文の中から長い単語上位 3 つを、そのレポートの特徴的な単語として抽出する。これらを組み合わせて検索ワードとして検索エンジンにより Web 検索をし、その結果上位にランクされた Web ページを取得する。

次に、剽窃されたレポートか評価するため、文章の解析を行う。本研究では Web ページをコピー&ペーストして作成されたレポートの発見を支援することを目的としている。よって、解析には表層情報を利用した解析法である n-gram 解析を用いる。ここでは、日本語の解析に最適とされている 3-gram を用いている⁵⁾。まず、レポートと取得した Web ページについて、それぞれ n-gram 解析を行う。解析した結果を用い、我々が提案した評価式⁶⁾により、レポートと各 Web ページとの類似性を評価し、類似度を算出する。この類似度は、レポートの文字列が Web ページに表層的にどの程度含まれているかの概算を表している。この類似度が高い Web ページから順に表示することにより、剽窃元となった可能性の高い Web ページをチェックしやすくする。

この列挙された Web ページの中の 1 つを選択すると、その Web ページとレポート中の文字列が表層的に一定数以上一致している部分を剽窃された可能性が高い部分としてマークして表示する。この一定数は、

長すぎると一部改編した部分を含んだ文字列にマークがつかなくなる。一方、短いと剽窃されていない部分にもマークがついてしまうという問題が生じる。マークをつける基準として、今回は経験的に 5 文字以上とした。この表示機能により、剽窃された部分を視認でき、剽窃レポートを発見することが容易になる。

2.2 複数からの剽窃発見における問題

Web からの剽窃を発見するには、剽窃元の可能性の高い Web ページ群を収集し、レポートと各 Web ページとの間でその可能性の評価をする必要がある。複数の Web ページからの剽窃に対応するためには、Web ページ収集のための検索ワードの設定と、剽窃の可能性の評価方法について検討する必要がある。

検索ワードの決定方法については、剽窃チェックを行う対象となる Web ページを取得する際、レポートの中から検索に用いる単語群を抽出し、それらを用いて Web 検索する方法が考えられる。これは剽窃の元になった Web ページには、レポートに使用されている単語が多く含まれているという特徴を利用している。

しかし、複数の Web ページから剽窃して作成されたレポートの場合、剽窃の元になった Web ページが複数存在する。そのため、抽出した単語群が必ずしも全ての剽窃元 Web ページに含まれているとは限らない。よって、検索に用いるために単語を組み合わせる際に、どの単語を用いるかを検討しなければならない。仮にレポート固有と思われる単語 3 つの OR 検索をしたとしても、全ての単語を含む Web ページが検出されてしまい、剽窃元 Web ページを取得できないことがある。

また、評価においては、一般的な文章間類似性評価手法を用いて、レポートと Web ページ 1 つの間の比較して、類似度を算出する方法が考えられる。しかし、その方法では多くの Web ページから剽窃し、それらを組み合わせて作成されたレポートの場合、各 Web ページからの剽窃部分がレポート全体に対して少ないため、類似度は低い値となる。そのため、剽窃せずオリジナルで書かれたレポートと複数の Web ページから剽窃されたレポートの類似度が同程度となり、剽窃レポートを区別することが困難になる。

実際に学習者から提出される剽窃レポートは 1 つの Web ページからのみの剽窃レポートは少なく、多くの Web ページから剽窃されたレポートが多い。このようなレポートは、現段階では適切には検出できない可能性が高い。

2.3 複数からの剽窃に対応する手法

2.2 節で挙げた複数の Web ページからの剽窃発見においての問題点を解決し、複数 Web ページからの剽窃であっても適切に発見できる手法を提案する。我々が提案した剽窃検査システムでは、1 つの剽窃元 Web ページを発見することは可能である。そこで、このシステムで 1 度剽窃検査した結果、剽窃元とされた Web

* 電子情報通信学会投稿中

ページから剽窃したと思われる部分を削除した文章に対して再検査することで、この問題を解決できると考える。

削除を行う際、コピーして一部改編した文には、オリジナルに作成した文に比べ、Web ページと一致している文字列が多く含まれている点に着目する。そこで、レポートを文に分割し、分割された文ごとにそれぞれ評価を行う。

具体的には、分割された文の中の Web ページと一致する文字列の割合が一定値以上の文を剽窃された文とみなし、レポートの文章の中から削除する。この操作により新しく作成したレポートの文章に対して、再度検索ワードを抽出し、Web 検索を行う。これにより検索結果に基づいて、1 度目は異なる単語群が得られ、異なる検索ワードでの検索が行われるため、前章の検索の部分の問題が解決できると考える。

評価は、検索の結果得られた Web ページと削除後のレポートの間で行う。剽窃レポートでは、削除により全体における一致する部分の割合が増えるため、削除前に類似度が低かったレポートのうち、剽窃レポートの類似度のみを高くできる。これにより、2.2 節で述べた評価における問題も解決し、剽窃元候補の Web ページを検出することができ、複数の Web ページから剽窃したレポートを検出することができる。

3. システムの実装

3.1 提案手法を用いたシステムの実装

提案した手法を用いて剽窃チェック支援を行うシステムを実装した。図 1 はそのインターフェースである。システムは Web ブラウザ上で動作するアプリケーションとして実装した。これはネットワーク環境があればどこからでも利用できるようにするためである。開発言語には、スクリプト言語である PHP を使い、Web サーバは PHP との相性がよく、手に入れやすい Apache を使用した。

図 2 にシステムの構成を示す。教師は、学習者からテキストファイルとして電子的に提出されるレポートをディレクトリ内に保存し、教師がインターフェース上で、その中からレポートを指定し、処理を開始する。

剽窃検査システムはレポートを取り込み、そのレポートから検索ワードを抽出し、Web 検索を行う。検索の結果 Web ページを収集し、それらの Web ページとレポートに対して類似度評価をし、類似度を算出する。類似度が高い順に Web ページをソートし、類似度ランキングを作成する。そして、この類似度ランキングと Web ページとレポートが出力される。

この剽窃検査システムの出力したデータを元に、剽窃されたと思われる部分の削除を行う。この削除機能では、次のような処理を行う。

1. 最長マッチング処理



図 1 システムのインターフェース
Fig. 1 Interface of the system

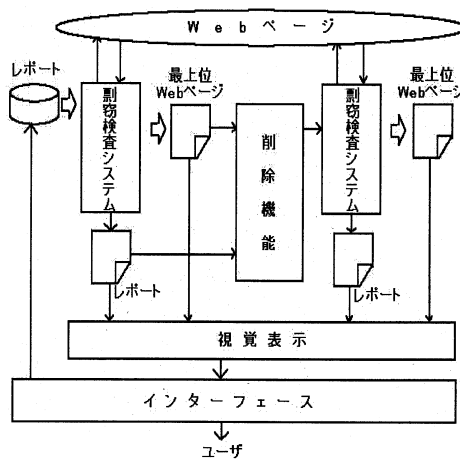


図 2 システムの構成
Fig. 2 Constitution of the system

類似度ランキングで最上位の Web ページとレポートの間で最長マッチング処理を行う。まず、レポートの文章にある文字列をもとに Web ページの文章の照合を行う。この処理では、5 文字以上の一致をマッチしたとみなす。マッチした文字列には、後で利用するためにマークをつける。

2. レポートの文章を文単位に分割

レポートの中の文が必ず読点によって区切られていると仮定し、読点ごとに文字列を切り出すことによってレポートを文に分割する。

3. Web ページと一致した文字列の割合の算出

ひとつの文の全文字数に対し、上述の 1 でマークした文字が含まれる割合を算出する。この処理をレポート中のすべての文に対して行う。



図 3 複数のページからの剽窃チェック部のインターフェース
Fig.3 Interface of the plagiarism check part from several Web pages

4. Web ページと一致する文字列の多い文を削除
閾値 n を設定し、割合が閾値を超える文をすべて削除する。そして、削除後のレポートデータを出力する。

このデータを用いて再度、剽窃検査システムにかけ、検査を行う。

この結果、1 度目の剽窃検査システムにかけた結果、最も高い類似度を示した Web ページと元のレポートとの間で一致した部分を視覚表示する。さらに、2 度目の結果、最も高い類似度を示した Web ページと削除後のレポートとの視覚表示を行う。出力結果の 1 例を、図 3 に示す。また、ここではそれぞれの類似度と Web ページアドレスも表示する。これらを一括して表示し、比較することで、複数の Web ページからの剽窃をチェックすることができる。

3.2 削除の基準となる閾値設定

前節で提案した手法では、文の中の Web ページと一致する文字列の割合が閾値 n 以上の場合、その文を剽窃された文とみなし、レポートの文章の中から削除する。実際にシステムを実装する際、その削除する基準となる閾値 n を設定する必要があるためアンケート調査を行った。

この調査では、本学科の大学生の「計算機システム」という授業で課された「DVD は発展途上の光ディスクである。最新の動向を調査せよ」というテーマで提出された 1000 文字程度のレポート 43 件の内、教師により剽窃と判断されたレポート 2 件を対象として行った。

レポートを文ごとに分割し、文中に一致する文字列が一定の割合以上含まれる文にマークを付けたレポートを用意する。今回はその割合を 0.6, 0.5, 0.4, 0.3 の 4 種類とし、それぞれの割合で文にマークしたレポートを用いた。これらのレポートと剽窃元 Web ページを示し、どのマークのつけ方が剽窃した部分を最も適

切に示しているかアンケート調査した。このアンケート調査では 5 人を対象とした。

表 1 に調査結果を示す。今回はこの結果最も多かった 0.4 を閾値 n の値とした。

4. 実験

4.1 シミュレーション実験

提案した手法により、複数 Web ページからの剽窃レポートを検出することができるか調べるため、擬似的にレポートを作成し、シミュレーション実験を行った。擬似レポートは剽窃レポート 10 件、剽窃レポートとの比較のために剽窃せず、オリジナルに書かれたレポート (非剽窃レポート) を 5 件作成した。擬似的な剽窃レポートの作成方法を次に示す。まず、レポートのテーマを設定する。そのテーマに沿って Web 検索を行い、その結果の上位 100 件の中から無作為に 2 つの Web ページを選択する。それら 2 つの Web ページから合計 1000 文字以上の最小となるようにランダムに文章をコピーする。このような操作をして作成したレポートに、文末の変換、文の順序の変更、文の削除の 3 種類の文章改編を行い、疑似剽窃レポートとする。

また、非剽窃レポートの作成方法は、擬似的剽窃レポートの作成時と同様のテーマに沿って、Web 検索をし、その中の Web ページを 1 度読み、内容を理解した上で、オリジナルの文章で 1000 文字以上のレポートを作成し、これを疑似非剽窃レポートとする。

これらの作成された疑似レポートに対して、1 度目の類似度評価値と、2 度目の類似度評価値の比較、検討を行った。

その結果、疑似剽窃レポートは 10 件中 9 件において 1 度目の評価で剽窃元を検出することができた。また、1 度目の評価で検出できなかった 1 件においても、2 度目の評価により剽窃元を検出した。

この実験の評価結果を表 2 に示す。表中の平均は、各レポートにおいて類似度の最も高い Web ページとの類似度の平均、標準偏差はその類似度の標準偏差である。表 2 より、剽窃レポートの 2 度目の類似度の平均値が、1 度目の類似度の平均と比べ、高い値となった。これにより、複数の Web ページから剽窃されたレポートを検出できる可能性が高くなったと考えられる。

4.2 学生レポートを対象とした実験

実際に提出された学生レポートにおいて、複数の Web ページから剽窃されたレポートを検出できるか調べるため、実際に提出された学生レポートを対象と

表 1 閾値設定調査結果
Table 1 Result of research for set the threshold

削除する割合	0.3	0.4	0.5	0.6
1 件目	0	4	1	0
2 件目	0	3	2	0

して実験を行った。実験の対象とする学生レポートは、本学科の学生の授業で課された「DVDは発展途上の光ディスクである。最新の動向を調査せよ」というテーマで提出された1000文字程度のレポート43件である。

まず、1度目の剽窃検査を行い、その最上位にランクされたWebページとレポートの間で類似度を算出した。その後、本手法で2度目の検査をした結果得られた最上位Webページと、削除後のレポートとの間で類似度を求めた。また、システムの評価が実際に人の評価とどのくらい近い評価を行うことができるのか、目視により主観的に剽窃か否か判断し、システムの評価との差異を調べた。

ここでは、レポートの半分以上が主観的に剽窃されていると調査者がみなした文で構成されているものを剽窃レポートとする。文が剽窃された文か否かの判断基準として次の3つを設定する。

1. Webページの文をそのままコピー&ペーストされたのみの文である。
2. Webページの文をコピーして改編された文である。ここで、文章の改編は、文末の変換、文の削除、文の前後入れ替え、文の分割・接合とする。
3. 文の構成が剽窃元だと思われるWebページの文と変わらず、同じ意味の違う単語に置き換えたのみの文である。

以上の判断基準によりレポートが剽窃かどうかを判断し、その評価結果とシステムの評価実験の結果について比較、検討を行った。

図4に結果を示す。このグラフは横軸を1度目の類似度結果を昇順にソートしたレポートの番号、縦軸を類似度としている。1度目の類似度を白、2度目の類似度を黒で示した。また、目視による主観的判断で剽窃とされたレポートは背景を灰色にしている。

グラフの右側のレポート群は、1度目の類似度が高いレポートであり、剽窃の可能性が高いと判断できる。グラフの中央付近の1度目での評価結果だけを見ると類似度が比較的高くなく剽窃の判断が困難なレポートの中に、2度目の評価結果で高い値を示しているものがある。これらは剽窃の可能性が高いレポートであるとシステムの評価結果から判断される。これらのレポートは目視においても、剽窃と判断されていることがわかる。目視において剽窃とされた13件のレポートのうち11件はシステムの評価においても高い値を示していた。しかし、目視による評価で剽窃と判断され

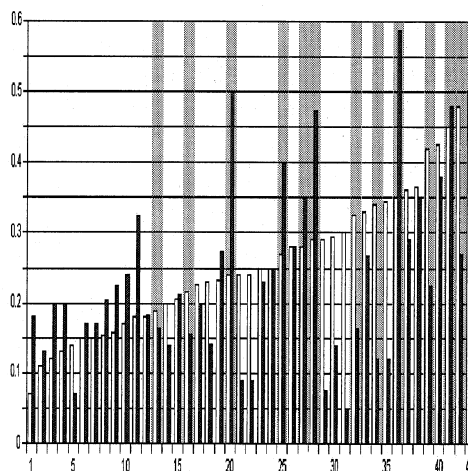


図4 目視による評価結果

Fig. 4 Result of evaluation by Viewing

ているレポートの中で、類似度の低いレポート、システムによる評価で類似度が高いレポートの中で、目視の評価では剽窃でないとして判断されているレポートが存在した。

5. 考 察

シミュレーション実験と実際に提出された学生レポートに対して行った実験結果から、本手法を用いたシステムが、複数のWebページから剽窃されたレポートの検出に有効であることを示した。また、目視による主観的な評価とシステムによる評価の比較により、実際に人がレポートを見て剽窃レポートか判断した場合の評価に近い評価をシステムが行えることが示された。

しかし、システムの評価と人の評価で違う結果を示すレポートが存在した。目視の結果、剽窃レポートであるにも関わらず、システムの評価する類似度が低いレポートと、システムの評価では高い類似度を示したにも関わらず、人の評価では、剽窃でないとして判断されたレポートである。

前者のレポートを、システムが正しく検出することができない原因について検討した。このシステムにより自動で収集されたWebページを調べた結果、目視によって剽窃元であると判断されたWebページが収集されたWebページの中に存在しなかった。よって、検索ワードの設定にどのような問題があるのか検討した。

検索が適切にされなかった原因は、次の3点であった。第1の原因は検索ワードとして抽出した単語群が調査者がみつけた剽窃元Webページ群に分散して存在しているケースであった。第2の原因は単語の表記を改編したもので、Webページ中で漢数字で表記して

表2 予備実験結果

Table 2 Result of a preliminary experiment

	剽窃レポート		非剽窃レポート	
	平均	標準偏差	平均	標準偏差
1 度目の類似度	0.538	0.138	0.182	0.045
2 度目の類似度	0.864	0.145	0.172	0.032

いたものをアラビア数字に置き換えたものや、外来語・カタカナ語などの発音による微小な改編であった。第3の原因は、誤字・脱字がある単語が検索ワードとして抽出されたものだった。本文のほとんどが Web ページの剽窃で作成されているレポートであってもこれらの単語が、抽出された検索ワードに含まれると剽窃元が適切に検出されない可能性があることが確認された。

この問題を解決するために、検索ワードの設定の方法を検討する必要がある。

前述の第2の原因については漢数字とアラビア数字の対応と異表記を含む外来語の辞書を作成して、用いることで解決できると考える。

第3の原因については、抽出した文字列に誤字、脱字などが含まれる場合は、単語辞書を用いて、辞書にある単語との類似性から正しい表記を抽出するという方法が考えられる。また、3つの単語を検索ワードとしてレポートから抽出しているがこの数を増やし、それらを組み合わせて検索をすることで収集する Web ページの件数を増やし、誤字などの影響を軽減するという対策も考えられる。

しかし、第1の問題のように検索ワードがそれぞれの剽窃された部分から抽出された場合には、この手法では適切に検出することができない。よって、検索ワードをレポートから抽出する以外の方法と組み合わせる設定方法についても考える必要がある。学習者が Web ページを剽窃する際、まず、レポート課題のテーマを検索ワードとして Web 検索をかける予想されるので、課題レポートのテーマの中の単語をそのまま使用するという手法が考えられる。また、学習者が剽窃に利用する Web ページには、用語説明のページや、そのページにある単語を利用し、さらに検索してみつけたページを用いるなど傾向がある。そのため、一般的な検索エンジンだけでなく、用語解説ページの検索機能を利用する方法や、検索されたページ中の特徴的な単語を抽出して再度 Web 検索を行うなど、検索ワードの設定を含めた検索の手法についてさらに検討する必要がある。

また、後者のレポートについては、その Web ページがレポートに比べ文字数が非常に多く、単語が多く一致しているとみなされたためであった。この一致した単語とは、DVD に関連した単語で DVD についてレポートを書く場合、よく使用される単語（「記憶メディア」や「ポリカーボネート」など）であった。また、「～している。」や「～を行った。」など語尾の検出も多かった。また、このような理由から1度目の評価で高い値を示したため、レポートから多くの文が削除され、削除後の文字数が他のレポートに比べ、少なくなっていた。これにより、2度目の類似度も高くなっていた。

この問題は、一致している文字列をマークする際、ソーラス辞書や頻出単語を記録した辞書などを用いて、それらを対象から外して評価することにより解決できると考える。

6. ま と め

本稿では、複数の Web ページから剽窃したレポートを発見する手法を提案し、その手法を用いた剽窃チェック支援システムを構築し、その評価を行った。さらに、システムが適切に剽窃レポートを検出できるのかを調べるため、シミュレーション実験をし、実際に提出される学生レポートを対象とした実験を行った。

それらの実験により、実装したシステムは、複数の Web ページから剽窃されたレポートを検出することが可能であることがわかった。しかし、このシステムにおいても人の判断する剽窃レポートをすべて検出することはできなかつた。

今後は、考察において述べた検索ワードの問題を解決する必要がある。

参 考 文 献

- 1) 浅見文絵：“大学における剽窃行為とその対策-英国・JISCPAS を中心に-”カレントウェアネス、No.285, CA1567,2005
- 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：言語処理学会第10回年次大会発表論文集 pp.729-732,2004
- 3) 太田貴久, 増山繁：模倣レポート判定支援システムの開発：言語処理学会第11回年次大会発表論文集 pp.293-296,2005
- 4) 村田哲也, 小高知宏, 小倉久和：n-gram 解析による学習者レポートの類似性比較：平成13年度電気関係学会北陸支部連合大会 F-116, pp.456,2001
- 5) 小高知宏, 高建武, 白井治彦, 黒岩文介, 村田哲也, 諏訪いずみ, 高橋勇, 小倉久和：n-gram を用いた学生レポート評価手法の提案：電子情報通信学科論誌 D-I Vol.J86-D-I No.9 pp.702-705,2003
- 6) 深谷亮, 山村毅, 工藤博章, 松本哲也, 竹内義則, 大西昇, 単語の頻出統計を用いた文章の類似性の定量化-部分的類似性の考慮-, 電子情報通信学会論文誌, D-II, Vol.J87-D-II, NO.2 pp.37-42,2004
- 7) JISC:Solutions for a new era in education, (http://submit.ac.uk/static_jisc/ac_k/index.html)
- 8) 宮川勝年, 高橋勇, 黒岩文介, 白井治彦, 小高知宏, 小倉久和：レポート剽窃チェックのための Web 検索システム：電気関係学会北陸支部連合大会 E-15,2005