レポート・論文での剽窃の言語的特徴と対策について

On linguistic features and measure of plagiarism reports

酒井　純
SAKAI Jun

1．はじめに

本研究は、レポートや論文における剽窃をチェックするソフトウェアをウェブサービスの形で開発するとともに、このソフトウェアを用いた剽窃抑止のための学習教材の開発を目的としている。このために、まずレポート・論文での剽窃された文章の特徴をまとめるとともに、これに基づいた剽窃チェックのソフトウェアを試作し、ウェブサービスとして教員や学生が利用可能な形での公開を目標とする。

2．剽窃の特徴

2.1．剽窃の定義

剽窃とは通常「他人の詩歌・文章などの文句または説をぬすみ取って自分のものとして発表すること」（広辞苑第5版）とされているが、本研究では次のように定義している。

- 剽窃とは、インターネット上にある文章を、コピー＆ペーストすることで、レポートや論文に盗用すること。また他の学生のレポートや論文をコピーすること。

本来の剽窃の意味は、インターネットに限らず書籍や雑誌などからの盗用、またレポート・論文に限るものではない。これに対して本研究での剽窃では、インターネット上からのコピー＆ペーストと、学生間のコピーに限定して考えるものである。

ただし、正しい引用の方法を理解しておらず、学生本人は引用しているつもりが剽窃になっていると考えられるものもある。このような「引用崩れ」の剽窃については、教員など人手による最終的な判断にゆだねるしかないと考えられる。

2.2．コピー＆ペーストによる剽窃の背景

学生によるレポート・論文での剽窃は、昔から変わらぬ古典的問題である。コンピューターやインターネットの普及以前より、レポートの書き写しや、書籍の書き写し、ほかの学生による代筆などが行われ、課題評価時に問題となっていた。しかし、コンピューターとインターネットの普及とこれに伴う検索サイト[1)]の発達、またWikipediaのようなインターネット上の辞典・辞書が一般化するとともに、学生による剽窃の質が変化してきたと言える。それは、コピー＆ペーストという方法により、他の人の文章を簡単に写すことが可能になり、その文章の内容を確認することなく、自分のレポートとして提出することが可能になってしまったことである。

インターネットの普及は、複数の人間の知識を集約するとされる集合知をもたらし、誰もが簡単に、専門的な情報にアクセスできるようになった。また、電子情報の特徴は複製可能性にあり、全く同じ形でのコピーを作ることが可能であることから、誰もが剽窃という正しくない形での情報の再利用をすることが可能になってしまった。

一方で、検索サイトの発達も、剽窃を助長しているとも言える。たとえば、「地球温暖化と環境問題について」というレポート課題が出たときに、課題文をそのまま検索サイトに入力し、検索ボタンをクリックするだけで、適切なウェブページがいくつも提示される。これは、図書館や書店に行って本を探したり、論文を探したりしていた時代に比べて非常に簡単に情報を得ることが可能になったことを示している。その上、その内容が簡単にコピー＆ペーストできるとなると、剽窃をするなという方が無理があると言えるかもしれない。

神戸親和女子大学　発達教育学部　ジュニアスポーツ学科　准教授

しかしながら、法律、研究、教育などの面から剽窃は許されざる行為であり、積極的に防ぐ必要がある。

2.3. 剽窃チェックの現状

一方、教員による剽窃のチェックは、現在までほぼ手作業にゆだねられている。現在すでに剽窃チェックプログラムもパッケージソフトとして販売されているが[2]、その価格の問題もあり一般に普及するには至っていない。

手作業で剽窃をチェックするには、剽窃が疑われるレポート・論文の1文を検索サイトで検索することで可能であり、実際にこの方法でいくつもの剽窃が発見できている。ただし、これらを元に成績評価をつけるには、証拠となる剽窃元のページを印刷しておいたり、剽窃箇所を明示できるようにしたりと非常に手間のかかるものである。これらの作業をすべてのレポートで行うことはほぼ不可能である。

この点で言えば、コンピューター・インターネットなどICTの普及は、レポート・論文の剽窃に関して学生には多大なメリット（？）を与えているが、教員にとっては現在までのところメリットは大きくないといえる。

2.4. 剽窃の特徴

2.4.1. 剽窃の元となる文章

コピー＆ペーストの元となる文章としては、通常のウェブページやブログ以外に、Wikipediaやナレッジコミュニティー、レポート共有サイトがあげられる。

まずWikipediaであるが、いわゆるインターネット上のフリーの百科事典である。集合知の象徴とされ、様々な項目についての網羅的な記述されている。その記述内容については、インターネットが利用可能なものであれば誰でも編集可能であり、このため項目の内容については必ずしも正しいとは限らない。また、議論のあるものや話題の項目については、編集合戦と呼ばれる状態になると、1日も何回も内容が変わってしまうものもあり、レポートや論文の参考文献には適さないとも言われている。

ただし、内容的には学部レベルのレポートでは、ほぼそのまま回答となりそうな記述がされている項目も多くある。このため、Wikipediaの項目を読んでしまうと、それ以上の内容について書くことが難しくなってしまう可能性も否定できない（学生が自分なりに試行錯誤して考える段階を飛ばして、いきなりまとまったものを見せられてしまい、やる気をなくす可能性は十分ある）。

次にナレッジコミュニティーについてであるが、これは一般には質問サイトやQ&Aサイトと呼ばれるウェブサイトである[3]。お互いに答えられる質問に答えることで、疑問や質問を解決するためのサイトで、これも集合知の一つの形であると言える。このようなサイトを利用して、直接、レポートの課題が質問としてあげられ、それに対する答えが利用されることも考えられ、また回答の文章がそのまま剽窃される場合もある。

一方、レポート共有サイトについては、まだ日本国内ではあまり表立って利用されていないが、学生同士でレポートを交換し、共有することを目的にしているウェブサイトである[4]。これらのサイトでは、レポート・論文の作成の参考などとして、他の人の文章を利用できるようにしている。しかしながらこれは表向きの説明であり、実際には剽窃の温床になっているのではないかと考えられる[5]。また、海外のレポート共有サイトにも日本語のレポートが共有されているようである。

ただし、レポート共有サイトから文章を入手しようとすると、ウェブサイトへの登録に手間がかかるだけでなく、文章のダウンロードにポイント[6]やお金がかかることがある。このことから、レポートや論文で「手抜き」をしようとしている学生が、そこまで積極的に行うとは考えにくい状況ではある。

2.4.2. 剽窃の単位・手法

剽窃が行われる文章の単位については、レポート・論文全体、段落ごと、文ごと、または文全体でのあらすじが考えられる。

まずレポートや論文全体を剽窃、つまりは丸写しして盗用する場合であるが、インターネットからの剽窃ではあまりこの例はない。授業等で課せられる課題では、800字以上や2000字程度、また卒業論文などでは16000字程度や20000字以上など、文字数が指定されていることが多い。このため、この条件にあった文章をインターネットから検索することは難しく、そのまま全体をコピー＆ペーストできないと考えられる[7]。その一方で、学生間でレポートをコピーするときには、文字数の制限が逆に限定的に働き、レポート全体を丸写しすることが多いようである。

次に、段落ごとの剽窃についてであるが、レポート・論文での剽窃は多くの場合この単位で行われることが多い。そして、剽窃元となる文章が複数にわたることも珍しくない。次の表は、筆者が担当した授業のレポートで見つかった剽窃での、剽窃元となっ

たウェブサイトの数と、そのまま剽窃された段落の数、剽窃の上で改編された段落の数である。

表1　剽窃元サイトの数と、剽窃された段落の数

レポート名[8)]	元となったウェブサイト数	そのまま剽窃された段落数	剽窃の上で改編された段落数
剽窃レポートM	10	4	7
剽窃レポートN	2	0	5
剽窃レポートY	4	5	5

これら3つのレポートすべてで、段落ごとでの剽窃がされている。実際には、そのまま段落をコピーしたり、一部の接続詞が省略されたり、段落を要約したりと様々な手法が使われている。

次にあげるのは、段落冒頭で節が補われた例、逆に段落冒頭の節が削除された例、また段落の内容を要約している例である。

剽窃例1　段落冒頭に節が補われた例

剽窃元1	メールの受信者と差出人の環境が異なるとき、……
剽窃 レポートM	このように、メールの受信者と差出人の環境が異なるとき、……
	下線部が追加された部分

剽窃例2　段落冒頭の節が削除された例

剽窃元2	米国では、たいていの企業に電話を掛けると……（中略）…… この10年でコンピューターの音声はかなりわかりやすく……
剽窃 レポートY	たいていの企業に電話を掛けると……（中略）…… コンピューターの音声はかなりわかりやすく……
	下線部が削除された部分

例1および例2は、段落冒頭に節が追加されたり削除されたりしている例である。剽窃では、学生自身が書いた文章と、他から剽窃した文章の整合性を取るために、段落冒頭部分での節の改編がされやすい。実際に、上記であげた3つのレポートの中でもこのような段落冒頭のみが変更された例が26段落中で4回見られている。

剽窃例3　段落の一部を要約している例

剽窃元3	情報の拡散とは、情報が（しばしば無秩序に）拡散してしまう現象を指す。特に近年では情報技術の発達により、情報の複製が容易となった事から、コンピュータネットワークを通じて拡散現象が発生しやすい。この現象は、簡単にいえば情報の複製等が広まる現象である。これは、末端に行くほどに情報が細分化されてその価値を失う散逸とは違い、元の情報やその複製情報が流布される。本来流通すべきでない方法で流通する可能性があるとして問題視される場合がある。 これの最も顕著な例が、近年度々発生している個人情報の流出である。個人を特定しうる個人情報を複製・流布（大抵は個人情報リストとして販売される）することで、本来は迷惑メールや電話セールスなどの望まれない営業行為に利用されないことを前提に集められているこれら情報が、そのような営業行為に用いられてしまうため、情報収集を行った側の信用問題にも発展する。
剽窃 レポートN	近年情報技術の発達により、コンピュータネットワークを通じて拡散現象が発生しやすくなった。この現象は、情報の複製等が広まる現象である。最近では多く言われている個人情報流出である。個人を特定できる情報を複製・流布している。迷惑メールや電話セールス等の望まれない営業行為に用いられてしまうため、情報収集を行った側の信用問題にも発展する。
	下線部が共通する部分

剽窃例3では、長い段落を390文字から165文字に短くまとめ、途中つなぎの部分で言葉を少し補っている。ほとんどそのままの文章を「継ぎ接ぎ」しただけにしては、よくまとまっているとも言える。

次に見る例は、いくつもの文章から段落を剽窃し、一つにまとめたものである。

剽窃例4　複数のウェブページの文章を1つの段落にまとめている例

剽窃元4	文字だけでは伝えきれない微妙なニュアンスや、いろんな感情をマークで表現するのが顔文字です。メールやチャットが普及するにつれて種類も増え、動く顔文字などのバリエーションも出てきました。
剽窃元5	顔文字には大きく分けて、日本式と欧米式がある。(^_^) が日本式で、:-) が欧米式、それぞれ起源が異なっている。欧米式のは、顔が横倒しになっているのが特徴だ。
剽窃元6	米国には半角文字しかないので、目が半角のコロン（:）で鼻が半角のハイフン（-）の場合が多く、どれを見ても似たような顔になる
剽窃元7	パソコン通信やインターネットでは文字が主な伝達手段であるため、感情を表現するために文章に混ぜて用いることがある。文章そのものだけでは誤解を与えると思われる時に、語調を和らげることができるという利点があるが、あまり親しくない人に対して使うと馴れ馴れしい印象を与えることがあるので注意が必要である。一般に、仕事で相手先とメールを交換する場合など、

	改まった場では用いるべきでないとされている。スマイリー（smilly）などとも言う。欧米では :-) のように横に倒した顔文字が使われることが多い。
剽窃 レポートＭ	また機種依存文字は顔文字を使う際にも注意が必要です。顔文字とは、文字だけでは伝えきれない微妙なニュアンスや、いろんな感情をマークで表現するのが顔文字です。メールやチャットが普及するにつれて種類も増え、動く顔文字などのバリエーションも出てきました。顔文字には大きく分けて、日本式と欧米式があります。(^_^) が日本式で、:-) が欧米式で、それぞれ起源が異なっています。欧米式の顔文字は、顔が横倒しになっており、米国には半角文字しかないので、目が半角のコロン（:）で鼻が半角のハイフン（-）の場合が多く、どれを見ても似たような顔になるのが特徴です。パソコン通信やインターネットでは文字が主な伝達手段であるため、感情を表現するために文章に混ぜて用いることがあります。顔文字には、文章そのものだけでは誤解を与えると思われる時に、語調を和らげることができるという利点があります。一般に、仕事で相手先とメールを交換する場合など、改まった場では用いるべきでないとされています。スマイリー(smilly) やエモティコン（emoticon）などとも言います。欧米では :-) のように横に倒した顔文字が使われることが多いです。
	共通する部分に下線または二重下線

この例４では、複数の文章から段落を持ってきているため、元々の文体が「である体」と「ですます体」の二つになっており、剽窃をした学生はすべてを「ですます体」に統一し直している。また、一見意味が通っているようにも見えるが、日本式と欧米式で顔文字の向きが違うという内容の重複が見られ、不自然な点も多い。

また、内容を完全には確認・理解せずに文章を無理矢理継ぎ接ぎしたために、意味が通じなくなったものもよく見かける。

剽窃例５　無理な編集により意味不明になっている例

剽窃元３	まだ活版印刷以前の時代に於いて情報は、その情報量と複製コストの兼ね合いから保護されていた。情報の複製を作る事は写本を意味していた事もあり、一定の情報量のある情報は、その量的な問題から保護されていたのである。これは同時に、本来広めるべき情報までもが広めにくいという事でもあったため、印刷技術の発達を促し、やがてそれは20世紀中頃よりコピー機に代表される一対一で複製を製作する機械の普及を発生させた。 ファクシミリの登場 　コピー機と同時期に利用されるように
	なってきたファクシミリ（FAX）という通信と複製を同時に行う装置の発達は、情報が拡散しやすい形での社会基盤となった。……
剽窃 レポートＮ	まだ活版印刷以前の時代は、その情報量と複製コストの兼ね合いから写本することによって保護されていた。コピー機やファクシミリという通信と複製を同時に行う装置の発達は、情報が拡散しやすい形になった。……
	下線部が剽窃された部分

この例では、「写本」の意味を理解せずに文をつなげてしまったために、１文目が「写本することによって保護されていた」という意味のわからない文章になっている。また２文目も「……、情報が拡散しやすい形になった」と「社会基盤」という意味的に重要な単語を省略してしまったために、意味不明の文章になってしまっている。

このように、改編が行われるものも含め、段落ごとに剽窃されることが多い一方で、剽窃された段落の特徴として、漢字の出現頻度が高い可能性がある。今回もちいた剽窃レポートで、学生本人の書いた文章が入っているのが剽窃レポートＮのみのため、データとしては有効性は低いが、次の表に漢字使用頻度の違いを例示する。

表２　剽窃レポートＮでの漢字使用頻度の違い

	本人記述部分		剽窃部分	
	文字数	文字数／総数	文字数	文字数／総数
漢字	413	**32%**	260	**37%**
ひらがな	676	52%	322	46%
カタカナ	151	12%	92	13%
英数字	16	1%	0	0%
記号	32	2%	32	5%
総数	1288		706	

一方、これまでの例でも見られるように、実際の剽窃手法としては、文頭または段落頭での節の挿入・変更・削除、文末における終助詞の変更による文体の変更、剽窃部分の長さの調節のための文章の省略や要約などがある。

以上のことから、学生のレポート・論文での剽窃の特徴を次のようにまとめることができる。

①剽窃が行われるのは、段落単位が多い。

②コピー＆ペースト後に編集されるときは、段落頭の節が削除・挿入されるか、または段落全体にわたって要約が行われる。

③剽窃された段落は、漢字の使用頻度が高い可能

性がある

この特徴を参考に、後述する剽窃チェックプログラムの試作を行っている。

2.4.3. 留学生による剽窃の特徴

留学生についても、剽窃に用いられる手法にかわりはないが、日本人学生の剽窃に比べると段落を改編せずに剽窃される場合が多い。次の表は、留学生が書いた剽窃を含むレポートでの、段落改変率についてである。

表3　留学生による剽窃レポートでの段落内改編率

	C1	C2	G	H	R	Y	計
改編あり	2	5	11	0	3	2	23
改編なし	9	5	4	11	16	10	55

個人差はあるものの、段落内のことばの入れ替えや変更は日本人の学生ほどは行われない[9)]。また、剽窃を含むレポートは全体に長くなる傾向が見られる。これは段落を要約することが日本語運用能力の面から難しく、剽窃した段落を無秩序につなぐためであると考えられる。

また留学生による剽窃については、別の問題として日本国外のレポート共有サイトからの剽窃もあると考えられるが、これについては今のところ有効な解決策は見いだせていない。

2.5. 剽窃に関する学生の意識

現在、本学の1年次情報リテラシー科目の授業では、「剽窃はなぜダメなのか？」というレポートが課せられている。このレポートの結果や、学生へ剽窃チェックプログラムの試作版を見せたときの反応から、学生の剽窃に対する意識は、次のようにまとめることができるであろう。

①罪悪感がない

②引用と剽窃を明確に区分できていない

③簡単にばれないと考えている

④教員が簡単にチェックできないと考えている

まず、①についてであるが、コピー＆ペーストして剽窃でレポートを作成することに、入学してすぐの時点ではあまり罪悪感を持っていないようである。これは高校までの授業において「レポート・論文の書き方」といった授業を受けておらず、そのために「剽窃はだめ」という指導も受けていないためだと考えられる。また、高校までには自分の意見を表現するといった科目や課題がないため、他の人の意見と自分の意見を区別して、論を組み立てるという訓練がされておらず、これも原因の一つになっていると考えられる。

実際に学生との話の中で、海外の大学での剽窃への対応が厳しいという話をすると「ありえない！」という感想が多数を占めた。それだけ剽窃が気軽なものであり、罪と考えていないことがわかる。物心ついたときからコンピューターに囲まれて育った、デジタルネイティブ世代にとって、情報はコピー＆ペーストできるのがあたりまであり、分からないものは調べればよいものである。この点を踏まえた指導が、今後必要であろう。

またこのような理由から、引用と剽窃の違いを明確には理解していないようである。レポートを作成するときに、どうすると剽窃になるのか、引用にするにはどうしたらよいかをよくわかっていない。本学では初年次教育の基礎演習という授業で引用の仕方について学ぶが、それ以前からレポート課題がどんどん出される。このため、正しい引用の仕方が身につく前に、剽窃が増えて行ってしまうとも考えられる。また、初年次教育の授業以外で引用の仕方をはっきりと学ぶ機会は少なく、学生によっては卒業論文を書く時になって、正しい引用の仕方を初めて知る学生が実際にいる状況である。

次に③の剽窃が簡単にばれないと考えていることについてである。教員にとって、レポート・論文の文章を一読すればどの部分が剽窃なのかを見抜くことは容易である。これは専門用語・言葉遣い・言い回し・慣用句・文脈構成など、日本語の様々な要素を総合的に判断していると考えられる。しかし、前述のような漢字の使用頻度のような簡単な指標でチェックができるようになると、学生にも剽窃は簡単にばれると説明がしやすくなるかもしれない。

最後に教員が剽窃を簡単にはチェックできない、と考えていることについてである。実際に学生の目の前で剽窃チェックの仕方を、手作業とプログラムのデモ版を用いた方法両方を見せたことがある。このときに学生からは「こんなに簡単にチェックできるとは思わなかった」などの意見が多く聞かれ、中には「ずるい」とまで発言する学生もいた。このことから、剽窃チェックプログラムを学生に触れさせることで、剽窃に対する抑止力になると考えられる。

3．剽窃チェックプログラム

3.1. 剽窃に関する問題点と対策

剽窃を無くすために、障害となっている問題点をまとめると、次の2点に集約される。

①教員が剽窃チェックをする作業量が多いこと
②剽窃に対する学生の理解度、意識が低いこと

単純にレポート・論文の剽窃をチェックするには、剽窃チェックプログラムを作成することで、教員側の作業量を減らすことはできる。これは現在も、剽窃チェックプログラムを購入・導入することでも可能である。しかし実際に剽窃チェックの作業量を減らそうとする場合には、まずコンピューター上で扱えるデータでの課題ファイルの提出手順の確立が必要である。また、複数ファイルの剽窃チェック方法の確立、また結果の出力方法（画面上に表示するのか、ファイルに書き出すのか、など）の確立など、一貫したチェックシステムの構築が必要になってくると考えられる。この点では、上田和志、他2010のプラグイン化の手法が参考になると考えられる。本研究では、剽窃をチェックする段階にとどまっているが、将来的にはファイルの提出時点でのチェックなど、一貫したシステムを構築していきたいと考えている。

次に剽窃への学生の理解度、意識が低いことについてであるが、まず通常の授業の中で剽窃の問題をしっかりと教育する必要がある。また本学の情報リテラシー科目でも、情報モラルとしてeラーニングの1ユニットを割り当てているが、現在の状況を見る限りまだまだ不足していると考えられる。そのため、剽窃チェックプログラムをウェブサービスとして実装し、eラーニングの教材に組み込む。学生には、実際に例文をチェックさせることで、これほど簡単に剽窃をチェックできるのだと言うことを示すことで、これが抑止力となり、剽窃を積極的に防ぐことが可能になると考えられる。

3.2. よく用いられる検索サイト

学生による剽窃の起点となるのが検索サイトである。学生がよく使う検索サイトと同じものを使用することで、その効率を上げることが可能になる。このためにも剽窃チェックプログラムを作成するのに、これを考慮する必要がある。

2011年4月時点での、日本国内の検索サイトシェアは、次の通りである。

表4 国内検索サイトシェア[10)]

順位	サイト	シェア
1	Yahoo Japan	55.9%
2	Google Japan	36.4%
3	Google USA	3.3%
4	BIGLOBE Japan	1.7%
5	Nifty Japan	1.1%

現在まで、日本国内での検索サイトシェアトップは、Yahoo!が保持してきており、次いでGoogleの順番になっており、近年この順位に変化はない。このため、これまで剽窃をチェックするときには、まずYahoo!で検索した後にGoogleで検索をしていた。ただし、この状況は、最近になって大きく変化した。

2010年の12月から、Yahoo! JapanでのGoogleの検索結果と同じものとなった。これは、Yahoo! Japanが独自の検索エンジン[11)]を使用していたのを、Googleの検索エンジンを用いるようになったためである[12)]。この結果、剽窃をチェックするにはYahoo!またはGoogleでの検索結果を検証するだけで90%の結果を網羅することが可能となり、剽窃チェックに用いるには十分な情報量と考えられる。

ただし、海外からの留学生の場合には異なる可能性がある。全世界でのシェアと異なり、日本での検索サイトのシェアはYahoo!が1位なのと同様に、韓国・中国の検索サイトシェアも世界、また日本とも異なっているためである。たとえば、2009年の韓国での検索サイトシェアはNAVERが約62%[13)]であり、また2010年の中国の検索サイトシェアは百度が72%であり、留学生の剽窃を防ぐためにはこれらの検索サイトの利用も視野に入れる必要がある。ただし、今回の研究ではこれは取り入れず、今後の課題としている。

3.2.1. Web API

通常、剽窃のチェックを手作業で行う場合にはウェブブラウザで検索サイトを開いて検索を行うが、今回開発するようなプログラム、ソフトウェアの場合には、Web APIといわれるツールを用いて検索サイトでの検索を行う。Web APIとは、各検索サイトが公開しているプログラム用の検索エンジンといえるであろう。ただし、これらのAPIには使用条件には課せられており、たとえばYahoo!の検索APIであれば1つのIDにつき、1日上限50,000件となっている。このため、剽窃チェックプログラムを作成する場合には、Web APIを利用する回数制限を意識しながら行う必要がある。

このため今回のプログラムでも、なるべくWeb APIを利用する回数を減らすことを意識し、開発を行っている。

3.3. 剽窃チェックプログラム

3.3.1. 形式

剽窃チェックプログラムとしては、これまでにク

ライアントソフト、サーバソフト[2)]、課題管理システムへのプラグイン（上田和志、他2010）のような形で提供されてきた。特にクライアントソフト、サーバソフトについては、前述のようにパッケージとして市販されるところまできている。

これに対して、本研究ではウェブブラウザ上で剽窃チェックのできる、ウェブサービス[14)]の形でプログラムの開発を行っている。

ウェブサービスとして開発するメリットとしては、次の4つが考えられる。

メリット1：インストール環境の違いなどを意識する必要がない
メリット2：eラーニング教材との親和性が高い
メリット3：複数開講科目で、学生同士のレポートのコピーをチェック可能
メリット4：キャッシュの利用により、Web APIを通じたクエリの上限数の制限緩和

まずメリット1についてであるが、ウェブサービスとして開発を行うことで、ソフトウェアを各パソコンにインストールする必要がなくなる。また、ブラウザが動作するコンピューターであれば、基本的にOSなどに限定されず、利用することが可能である。また、サーバー上のプログラムを変更するだけで、バージョンアップも随時行うことが可能であり、機能の追加も後から行うことができる。特に今回は学生にも利用させることを考えており、予算・手間もかからないという意味で理想的である。ただし、サーバマシンの負荷と、Web APIの利用制限数の問題から、まずは学内に限定して公開の予定である。

次にメリット2についてであるが、現在本学で利用しているeラーニングのシステムでは、教材をhtmlのファイル形式作成しており、ウェブサービスとして剽窃チェックプログラムを作成することで、教材の中から剽窃チェックプログラムをシームレスに利用可能になる。今回の研究では、この教材への組み込みが重要な部分の一つであり、この点ウェブサービスであれば簡単に利用が可能である。

メリット3として、ウェブサービスとして展開することにより、1つのサーバーにデータを集約することとなり、レポート・論文のデータと、検索した結果のデータのデータベースを共通化することが可能である。これにより、チェックしたレポートや論文、およびそのときに収集したデータを再利用できる。このことにより、複数開講科目で共通のレポートをチェックしたときなどに、学生間のレポートのコピーをチェックすることが可能となる。

また技術的な問題であるが、メリットの4として、データベースを共通利用することにより、Web APIの利用回数を減らすこともできるのではないかと考えられる。これは、剽窃チェックの際に検索した結果をデータとして蓄積することにより、次に同じキーワードで検索されたときにはこれらの結果をキャッシュとして再利用することが可能となる。これによりWeb APIの利用回数を減らすことができると考えられる。

一方、デメリットとしては、クライアントにインストールする形のソフトに比べて、ドラッグ＆ドロップの利用や複数ファイルの指定のしやすさなど、ユーザーインターフェースが少し使いにくくなる可能性がある。しかしこれは、AJAXや、これから普及すると考えられるHTML5などの利用により、今後解決できるのではないかと考えている。

3.3.2. アルゴリズム

剽窃をチェックする方法については、いくつかの方法がすでに提案されている。

①検索サイトを利用して全文検索[15)]
②キーワードで検索、その結果をキャッシュして本文と比較。
③レポートの一部を抜き出して検索、その結果をキャッシュとして本文と比較。

本研究では②と③の方法を組み合わせている。

まず、入力されたキーワード、または本文の冒頭の題目をWeb APIで検索する。これにより、検索結果としていくつかのウェブページのURLが得られるので、これらのページの情報をすべて収集してキャッシュとする。これとレポート・論文の本文の比較をする。

このときに剽窃が見つからなかった段落や部分については、段落ごとに漢字率を計算して、漢字比率の高い段落から1文を抜き出し、Web APIで検索する。これで見つからなければ、剽窃はないと判断する。これを文章の中の段落の数に応じて、必要な段落分繰り返す。

この結果を、左側に元の文章、右側に見つかったウェブページの形で表示する。特に剽窃が判断された部分についてはオレンジに色をつけ、また剽窃が見つかった中でも2文以上連続している部分には赤色に色をつけて、目立つように表示する。

これらの検索結果、レポート・論文ファイルをキーワードごとに保存して、同じキーワードで検索が行

われたときにキャッシュとして再利用する。このことにより、Web APIの使用頻度を下げる。

3.3.3. 完成サイト

現在、作成したウェブサービスを学内限定で以下のアドレスで公開している。

http://visnu.kobe-shinwa.ac.jp/dps/

3.4. 教材への応用

剽窃チェックプログラムを元にした、eラーニング教材を現在作成中で、授業の中に組み込む予定である。内容的には、これまでの情報モラル、剽窃に関するユニットに、剽窃チェックを行うページを追加する予定である。これにより学生が実際に自分で剽窃チェックを行うことで、剽窃に対する抑止力となることを想定している。また、この教材に剽窃に関するアンケートを組み込み、学生の意識調査と効果測定を同時に行う予定である。

4. おわりに

今回の研究で、レポート・論文での剽窃の特徴をとらえることができたと考えられる。実際の剽窃では、単に段落をそのままコピー&ペーストして剽窃するものだけでなく、様々な手法が用いられていることが分かった。特に、段落頭を改変し、段落中の文章はそのままであることが多く、また文体を統一させるために「である体」と「ですます体」が統一されることが分かった。

この特徴を元に、剽窃チェックのプログラムを試作しているが、公開段階に至るにとどまっており、今後この剽窃チェック効率を計測するとともに、その抑止効果についても評価測定が必要であると考えている。

また剽窃チェックプログラムの今後の展開として、いくつかの改善を予定している。まずウェブサービスを発展させた形でミドルウェア化できればと考えている。これにより、剽窃チェックプログラムの本体とインターフェースを分離することが可能になる。そして、ウェブブラウザからの利用に限らず、プラグインを開発することでeラーニングのシステムに組み込み、レポート提出時にチェックをかけられるようにできればと考えている。

5. 参考文献等

5.1. 注

1）検索エンジンとはYahoo!やGoogleなどに代表される検索サービスのこと。検索サイトとも。
2）「コピペルナー」http://www.ank.co.jp/works/products/copypelna/
3）Yahoo!知恵袋やOKWaveなど。2011年度京大入試ではカンニングで利用されて有名になった。
4）レポート共有サイトとしては、Report Report（http://www.reportreport.jp/）やHappy Campus（http://www.happycampus.co.jp/）などがある。
5）もちろん悪いことばかりではなく、就職活動での合格者のエントリーシートの書き方を学べるなど有効な使われ方もしている。
6）自分で書いたレポートを登録したり、アンケートに答えたりすると付与されるポイント。
7）もちろん例外はあり、ある課題において「練馬区環境作文コンクール中学生部門（http://www.city.nerima.tokyo.jp/manabu/kankyogakushu/sakubun/index.html）」のページからの丸写しの剽窃が確認されている。
8）ここでの剽窃レポートとは、筆者の担当していた授業で実際に提出されたものである。
9）これはもちろん個人差があり、剽窃レポートGでは、多くの段落に改編がされている。ただし、改編内容を見ると、日本人に比べて文単位の削除が多いように感じた
10）GA-Proから引用。http://www.ga-pro.com/
11）検索エンジンとは、ウェブ上の情報を検索するためのプログラムを指す（狭義）。ただし、Yahoo!やGoogleなど検索サイト全体のこと（広義）を指すこともあるので注意が必要。本論では検索エンジンと検索サイトを区別して使用している。
12）SEMリサーチ「ヤフー、検索エンジンGoogleへの移行完了」http://www.sem-r.com/google-2010/20101130144509.html
13）http://www.comscore.com/Press_Events/Press_Releases/2009/6/NHN_Corporation_is_Most_Visited_Web_Property_in_South_Korea
14）ウェブサービスとは、「WWW関連の技術を使い、ソフトウェアの機能をネットワークを通じて利用できるようにしたもの。」（IT用語辞典、2011/3/31確認）
15）実際には、Web APIの利用上限回数にすぐに到達してしまうため、現実的ではない。

5.2. 文献

太田貴久、増山 繁「学生レポート採点支援のためのレポート類似部分発見手法」、電子情報通信学会技術研究報告. NLC、言語理解とコミュニケーション105（594）、37-42、2006-01-26
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-I、情報・システム、I-情報処理 J86-D-I(9)、702-705、2003-09-01
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「Webサイトからの剽窃レポート発見支援システム」、電子情報通信学会論文誌. D、情報・システム J90-D(11)、2989-2999、2007-11-01
上田和志、富永浩之「類似性に基づくレポート剽窃の検出ツールのプラグイン化」、情報処理学会研究報告 Vol.2010-CE103 No4、2010/3/6

5.2. 参考サイト

アンク社「コピペルナー」
http://www.ank.co.jp/works/products/copypelna/Client/

5.3. 剽窃元サイト

剽窃元1：パソコン教室ネバーネバーランド「機種依存文字について」（2011/03/31確認）

http://www.nnland.co.jp/web/kishuizon.html
剽窃元 2 ：日経サイエンス「話し上手なコンピューター」
（2011/03/31確認）
http://www.nikkei-science.com/page/magazine/0509/speak.html
剽窃元 3 ：Wikipedia「情報の拡散」2010年10月26日版
（2011/03/31確認）
http://ja.wikipedia.org/wiki/% E6% 83% 85% E5% A0% B1% E3% 81% AE% E6% 8B% A1% E6% 95% A3
剽窃元 4 ：いろいろやりたい初心者のページ「顔文字。」
（2011/03/31確認）
http://rose.zero.ad.jp/~zac06238/PC/kaomozi.htm
剽窃元 5 ：マンガの中の聴覚障害者「顔文字の起源 (^_^)」
（2011/03/31確認）
http://www001.upp.so-net.ne.jp/wakan/Others/FaceMark.html
剽窃元 6 ：パソコントラブルQ＆A「顔文字について (^-^)」
（2011/03/31確認）
http://www.724685.com/weekly/type/table07.htm
剽窃元 7 ：IT 用語辞典 e-Words「顔文字【emoticon】」
（2011/03/31確認）
http://e-words.jp/w/E9A194E69687E5AD97.html