著作権侵害検出のための剽窃Webページ発見システム
A Plagiarism Web Page Detection System for the Copyright Infringement Detection

坂口 朋章† 雲居 玄道‡ 石田 崇‡ 平澤 茂一‡

Tomoaki SAKAGUTI† Gendo KUMOI‡ Takashi ISHIDA‡ Shigeiti HIRASAWA‡

† 早稲田大学 大学院理工学研究科
‡ 早稲田大学 理工学部

† Graduate School of Science and Engineering, Waseda Univ.
‡ School of Science and Engineering, Waseda Univ.

要旨：
近年の情報技術の発達により，blogなどを通じて多くのユーザーが情報を発信している．それに伴って，他人の文書を剽窃している文書の存在が問題となっている．しかし，Webページの増加により剽窃文書を人手で調査するのは困難となっている．本研究では，対象とする文書から自動的に剽窃の疑いのある文書を発見するシステムを提案する．このシステムは，検索エンジンで類似文書を抽出し，対象文書との比較により著作権侵害検出を支援するものである．実験・評価には，剽窃が新聞記事や実際のWebページを剽窃した文書などを用いる．

Abstract:
It becomes very easy to plagiarize sentences by the development of the information technology in recent years, and the violation of the copyright with copy and paste from sentences that to be made to the electron on Web etc. The proposed method is similar sentence discovery technique by dividing sentences into the word in this research, and discovering the continuous words that co-occurred between two documents.Moreover, it aims to develop the system that supports the discovery of an illegal Web page by discovering the Web page with the doubt of the plagiarism origin by using the Web search engine, and comparing it with former page.

1 はじめに

近年の情報技術の発達により，wiki, blogなどを通じて多くのユーザーがWeb上で情報を発信できる環境が整ってきている．また，検索エンジンサイトにより，キーワードを入力するだけで，関連した情報が書かれたWebページを大量に探し出すことが可能になった．その結果，新聞記事の無断転載，Web上の著作物文書のコピーなど著作権違反のページが増加していて問題となっている．

これらをすべて人手で発見するのは非常に困難である．まず，剽窃元文書を探すには類似文書を探さなければならない．

高橋らの剽窃文書発見手法[3]では長さの長い単語を組み合わせて検索を行っている．しかし，長さの長い単語には外来語が多く，一般的な文書には適用が難しい．田代らの研究[6]では，隣接する文節を組み合わせて検索を行っているが，検索回数が多くなってしまうという問題点がある．

[7]では文書を単語に分割し，2文書間に共起する連続単語系列を発見することによる，剽窃発見システムを提案した．しかしWeb検索を行う際のキーワードが適切でなく，あまり類似文書を得られなかった．

そこで，本研究ではWeb検索を行う際のキーワード選択を改良したシステムを提案する．システムを試作し，学生による剽窃レポートと著作権侵害のあった新聞記事に適用し，有効であることを示す．

2 Webページの著作権侵害

2.1 著作権侵害

著作権とは，著作物の創作者である著作者に保障される権利の総称であり，知的財産権の一種である．現行の著作権法では，いくつかの条件を満たせば権利者の許諾を得ることなく文書をコピーして掲載することができる．以下にその条件を示す．

1) その部分を引用する必然性がある．

2) 引用であることが明記されている．

3) 著作物全体の中で自分の書いた部分が「主」，引用部分が「従」である

以上の条件を満たしていれば，正当な引用となる．しかし，他人の文書の単なる丸写しや，「てにをは」などを少し変えただけの文書を掲載するのは無断転載あるいは剽窃となり，著作権侵害に当たる．本研究において検出対象となるのはこのような文書である．

3 著作権侵害 Web ページ発見支援システム

本研究では Web ページをコピーすることにより作成された著作権侵害ページを探すことを目的とし，そのようなシステムを考える．今回の研究では，太田らの研究 [2] と同様に下記の 3 つのフェーズを用いる．

3.1 Web 検索フェーズ

対象文書が他の文書をコピーして作られたものであるか否かを調べるためには，まず類似文書を検索する必要がある．検索エンジンに問い合わせるための検索ワード生成にあたっては，文書の改変を考慮する必要がある．

検索ワードは，文章の改変された部分を含まないような，検索結果が絞り込めるものがよい．そこで，複数回の検索結果の和集合を抽出することを考える．

田代らの研究 [6] では連続する k 個の要素を検索ワードとしているが，長い文書では検索回数が非常に多くなり非効率的であると考えられる．そこで本研究では，検索ワード作成のために日本語構文解析システム (KNP[8] など) を用いて，使用する要素を決定することにする．構文解析の例を図 1 に示す．

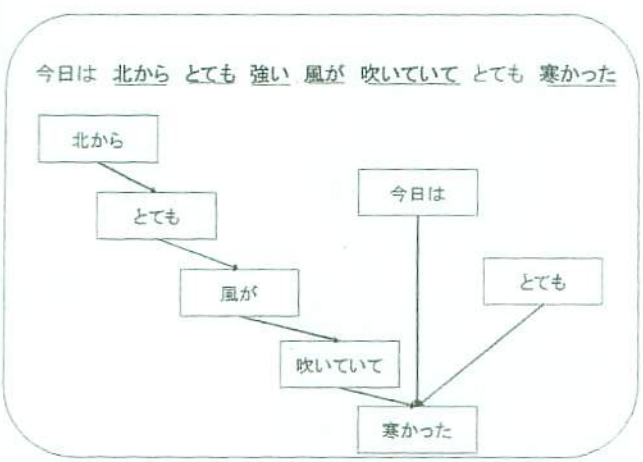

図 1: 構文解析の結果の例

独立した文節は改変される可能性が高いが，長いパスの全てが改変されてしまうことは少ない．そこで，最も長いパスに含まれる文節を利用して検索する手法を提案する．この手法によって，図 1 の例では下線のついた文節が抽出される．

検索ワード生成アルゴリズム

(1) 元文書を文節列に分割する．

(2) 係り受け解析を行い最も長いパスに含まれる文節を抽出する．

(3) 連続する k 個の要素を and で結合し検索ワードを生成する．

(4) $n-k+1$ 個の検索ワードが作成されるまで，(3) を繰り返す．

検索結果の上位にランキングされている Web ページ程，剽窃者が参考にしている可能性が高いと考えられるので，それぞれの検索ワードに対し上位 N 件の Web ページの URL を剽窃元候補集合として収集する．これにより，1 つの Web ページに対して，複数の剽窃元候補 Web ページが得られることになる．

3.2 剽窃ページ判定フェーズ [7]

類似文書集合を作成した後，次に対象文書と比較し，剽窃している可能性を判定する機能が必要となってくる．これは文書間の類似性を評価する問題とみなすことができる．文書間の類似度評価の手法としては，例えば文中の名詞と動詞を用いて文間の類似度を計算する手法 [2] や，n-gram 解析により文字列の出現頻度分布を用いる手法 [4] などの様々な手法が提案されている．

3.2.1 剽窃発見のための Smith-Waterman アルゴリズム [5]

本研究では，Robert W. Irving[5] によって提案されたアルゴリズムを利用する．このアルゴリズムは，2 文書間に一致する単語の情報を用いて剽窃とみられる連続単語系列を発見するものである．

文書 X の i 番目の単語 $X(i)$ と文書 Y の j 番目の単語 $Y(j)$ が一致することを $X(i)=Y(j)$ と表す．文書 X の i 番目の単語と文書 Y の j 番目の単語の組におけるスコアを $S_{i,j}$ とする．

Smith-Waterman アルゴリズム

(1) 文書 X と文書 Y の中から一致する単語の組を見つけ，連続単語系列の始点とする．ここでは $X(i)=Y(j)$ の場合を考える．初期スコアを $S_{i,j}=1$ とする．

(2) 一致した単語以降のスコアを以下のように求める．スコアが 0 になる単語の組以降のスコアは求めない．

$$S_{m,n}=\begin{cases} S_{m-1,n-1}+1, & \text{if} \quad X(m)=Y(n) \\ \max(0, S_{m-1,n}, S_{m,n-1}, S_{m-1,n-1})-1, & \text{otherwise} \end{cases} \tag{1}$$

(3) スコアを求めた範囲の一致した単語の中で，始点から最も遠い組を終点とする．

(1) から (3) で剽窃とみられる連続単語系列が 1 組得られる．これを繰り返すことで 2 文書間の全ての剽窃とみられる連続単語系列が得られる．

2 つの文字列「XABCXDEXFGHXX」と「ABYCYDEFGYYYH」が与えられたとき，図 1 のようにスコアが計算され「ABCXDEXFGH」と「ABYCYDEFGYYYH」が得られる．スコアが 3 のときは互いの文書の 4 語先までの中から一致を調べる．3 語の挿入・欠落を許容するということになる．

本研究では，得られた連続単語系列内での最大スコアが 10 以上のものを剽窃とみられる連続単語系列とみなす．今回対象とする単語は，名詞と動詞のみとする．

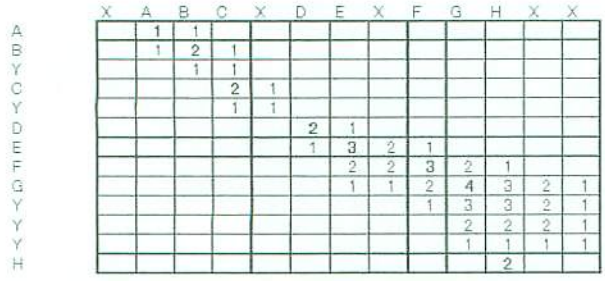

図 2: 剽窃発見のための Smith-Waterman アルゴリズム

Robert W. Irving の提案したアルゴリズムは単語の欠落・挿入には対応できる．しかし，日本語の剽窃に見られる文節単位や文単位の入れ替えは考慮されていない．よって，入れ替えにより連続単語系列の長さが短くなり検出ができなくなるという問題点がある．そこで，本論文では我々が以前に提案した手法 [7] を用いる．

本研究では検出する連続単語系列の単位を 3 単語とした．そして，「AXB」と「BYA」となった場合「X」と「Y」の部分がともに 5 語以内の場合に結合するものとした．

3.3 検査者提示フェーズ

剽窃している疑いが高い対象文書とその剽窃元と思われる文書が見つかったとしても，最終的に剽窃か否かを判断するのは人手にまかされることになる．そのため，剽窃か否かをできるだけ容易に判断できるようにするために，それらを効果的に表示する機能が必要となる．本研究では，チェック対象ページと剽窃元候補のページの連続一致単語系列が含まれる部分を文書中から抽出し剽窃検査者に表示することによりこれを実現する．

表 1: 検索手法による結果の例

	検索回数	剽窃判定数
田代らの手法 [6]	207	48
提案手法	124	133

4 評価実験

本研究で述べたシステムを実装し，評価実験を行った．本実験では，検索フェーズには Web ページの剽窃の含まれる学生レポートを用いる．判定フェーズには著作権侵害のあった新聞記事を用いる．形態素解析には茶筌 [9] を用いた．パラメータは $N = 20, k = 2$ とした．

本システムでは，Yahoo Japan [1] の提供するサービスを利用している．なお，検索の制限回数は 50000 回/24h，1 検索あたりの最大検索結果数は 50 である．

4.1 学生レポートを用いた評価実験

本実験では Web ページの剽窃の含まれると思われる学生レポートから剽窃元の Web ページを検索する．

課題：情報化社会とアウトソーシングについて

レポート数：20 件

科目名：情報化社会概論

レポートは電子メールで提出され，形態素解析には茶筌を用いた．

4.1.1 検索フェーズ

検索フェーズでは全ての文節を用いる手法と本手法の比較を行った．その結果を表 1 に示す．

我々の提案により検索回数，計算量などを削減しながらも多くの剽窃候補が得られた．これは提案手法により不要な文節が取り除かれて，剽窃の検索に適した文節がキーワードとして選択されたためである．

1 つのレポートから多くの剽窃が発見された理由は，有名な Web ページを他の Web ページが剽窃しているためである．剽窃文書から他の剽窃文書を発見できていることから検索フェーズとして良い性能を示しているといえる．

4.1.2 剽窃判定フェーズ

提案手法では従来手法より多くの文書が剽窃元文書と判定された．他の Web ページを剽窃した Web ページを多く検出しているので，剽窃判定数がかなり多くなった．

実際に剽窃があったかはわからないが，我々が人手で確認した結果，全ての剽窃元文書が妥当であると判断された．これは，学生レポートを作る際に，文書の改変をあまり行わなかったため検出が容易であったためだと思われる．

4.2 新聞記事を用いた追加実験

学生レポートは実際に剽窃をしたのかが確認できないため，剽窃の判明している新聞記事で剽窃判定フェーズの追加実験を行う．

実験データには著作権侵害のあった山梨日日新聞の新聞記事を用いる．社説の盗用があったと発表されている文書は対照表として公開されており，その中の文書を対象として実験を行った．

これらの新聞記事には改変が多く，これらの判定ができれば一般的な文書十分な結果が期待できる．

今回対象とした剽窃箇所は 51 カ所，そのうち剽窃と判断されたのは 44 カ所であった．

実際に剽窃部分と判断された部分を図 2 に，対照表の中で検出できなかった部分を図 3 に示す．

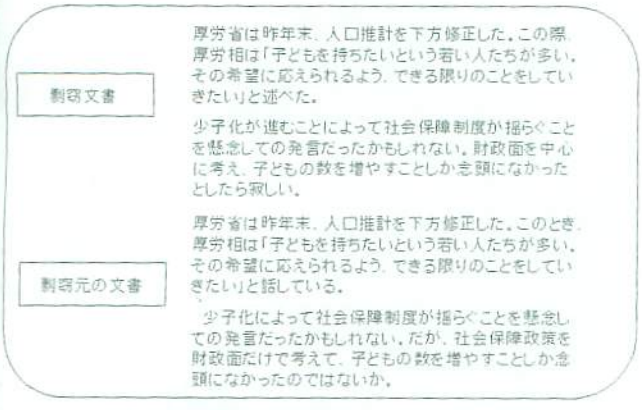

図 3: 剽窃と判断された文書の例

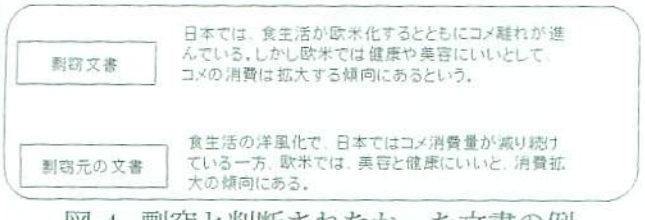

図 4: 剽窃と判断されなかった文書の例

剽窃と判断されなかった部分は，言い換えや入れ替えが多く今回のシステムでは検出することができなかった．

また，対照表には無いが検出された文章があった．事実を述べている文章が多く新聞社は剽窃と判断しなかったもの思われる．人手で調査しても見解が分かれるところであり，システムの動作としては問題ないと考えている．

5 おわりに

本研究では，本研究では Web 検索を行う際のキーワード選択を改良したシステムを提案した．その結果，本システムが効率的に剽窃を検出できることを示した．特に検索フェーズでは計算量の削減しながら多くの剽窃候補ページを収集することに成功した．

これにより，違反が多いホストの管理者や作成者個人に警告を行うことができる．また，著作権侵害の抑制にも役立つと思われる．

今後の課題としては，シソーラスなどを用いることにより，同義語や多義後を考慮したシステムの改良などがあげられる．

本研究は，(財) 電気通信普及財団による研究助成を受けて実施しました．

参考文献

[1] Yahoo! JAPAN,http://www.yahoo.co.jp/

[2] 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”, 信学技報, NLC2005-112, pp.37–42, 2006.

[3] 高橋勇, 宮川勝年, 小高知宏, 白井治彦, 黒岩丈介, 小倉久和, “WEB からの剽窃レポート検出手法の実装と評価”, 人工知能学会研究会資料, SIG-ALST-A503-01, 2000.

[4] 深谷亮, 山村毅, 竹内義則, 松本哲也, 工藤博章, 大西昇, “単語の頻度統計を用いた文章の類似性の定量化”, 電子情報通信学会論文誌, J87-D-II ,02, pp.661–672, 2004.

[5] R.W. Irving, “Plagiarism and collusion detection using the smith-waterman algorithm”, Technical Report 164, Dept of Computing Science, University of Glasgow, 2004.

[6] 田代崇, 上田高徳, 堀泰祐, 平手勇宇, 山名早人, “Web 上の文章を対象とした著作権違反自動検知システム”, 日本データベース学会 Letters Vol.5, No.2, 2006.

[7] 高島秀佳, 坂口朋章, 長尾壮史, 石田崇, 平澤茂一, “著作権侵害検出を目的とした類似文書発見手法”, 経営情報学会, 2006 年度秋季全国研究発表大会予稿集, pp.58-61, 2006.

[8] 日本語構文解析システム KNP, http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[9] 形態素解析システム茶筌, http://chasen.naist.jp/hiki/ChaSen/