著作権侵害検出を目的とした類似文書発見手法

Similar document discovery technique for detection of copyright violation

高島 秀佳† 坂口 朋章† 長尾 壮史† 石田 崇‡ 平澤 茂一‡

Hideyoshi TAKASHIMA† Tomoaki SAKAGUTI† Masahumi NAGAO† 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 University

‡ School of Science and Engineering, Waseda University

要旨： 近年の情報技術の発達により，文書の剽窃が非常に容易なものとなり，Web 上での電子化された文書からのコピー&ペーストなどによる著作権の侵害が問題となっている．本研究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類似文書発見手法を提案する．その結果，Web 検索エンジンを活用することで，剽窃元の疑いのある被剽窃 Web ページを抽出し，それを盗用したと思われる不正（剽窃）ページと比較することにより，著作権侵害の検出を支援するシステムを開発することを目的とする．

Abstract: In recent years,it becomes very easy to plagiarize sentences by the development of the information technology and the violation of the copyright by copy and paste from sentences on the Web causes serious problems. We propose a new similar sentence discovery method by dividing sentences into a set of words, and by discovering the sequence of words that co-occurred between two documents. As a result,it makes possible to develop the system that supports the discovery of an illegal Web page by comparing them with the Web pages generated by a Web search engine.

1 はじめに

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 Web 上で情報を発信するようになった．しかし，それに伴って，新聞記事の無断転載，Web 上の文書をコピーして，それに多少の手直しを加えただけなどの著作権違反のページの存在が問題となっている．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エンジンサイトがある．調べたい事柄に関するキーワードを入力するだけで，そのキーワードに関する情報が書かれた Web ページを大量に探し出して表示してくれる．また，マウス等の装置を使って，Web ページに書かれている文書を簡単にコピーする事が可能であり，WORD 等の文書編集ソフト内に貼り付け，編集を行うことも極めて容易である．その結果，個人が簡単に他人の記事や文書を基にした剽窃ページを作成することができてしまう．

これらの剽窃文書をすべて人手で発見するのは大変困難な作業である．剽窃者が Web 上の検索エンジンを用いて剽窃元となるページを入手している場合には，まず剽窃元の Web ページ（被剽窃ページ）を探す必要がある．剽窃者は必ずしも自分のページの記事や内容に直接関連するキーワードを検索に用いているとは限らないため，検索の際には様々な検索キーワードを考えて試行錯誤しなければならない．

次に，検索の結果得られた Web ページが剽窃元であるか否かを判定しなければならない．剽窃者は自分が剽窃を行ったことが容易には分からないように，語尾の変更や文の順序の入れ替えなどの編集を行う．よって，剽窃元かどうかを判定するためには検索の結果得られた文書全てに目を通す必要がある．

最後に，得られた Web ページが剽窃元だと判定された場合には，剽窃者に対して，剽窃行為の証拠として適切に剽窃箇所を示す必要がある．このような剽窃を発見するための手法として学生レポートに関する研究がある [1][2]．また，最近 Web からの文書抽出に主眼をおいたクローリングに関する報告もある [3]．

本研究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類似文書発見手法を提案する．この手法を用いて，Web 検索エンジンの活用により，剽窃元の疑いのある Web ページ（被剽窃ページ）を発見し，それを剽窃ページと比較することにより，不正判定を支援する．また，その際に文書中の類似部分を抽出し，ユーザに提示することができるような，著作権侵害の可能性のある Web ページの発見を支援するシステムを開発することを目的とする．

なお，ここでは，著作権侵害検出を文書（テキスト）のみから成るページに限定する．

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 [7]．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ページである．

2.2 検出対象

検出対象をどのように設定するかによって，用いる手法も変わってくる．

本研究では1つのWebページに対して複数の著作権侵害候補を用意し，それぞれとの間で文書マッチングを行う．さらに，本研究ではあるページの名詞と動詞に着目したとき，単語が連続してT個以上他のページと一致するような場合，そのページには著作権を侵害している可能性があると判定することとする．

3 著作権侵害Webページ発見支援システムの設計

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．そのようなページを発見し検出するためには下記の3つのフェーズが必要になる．

3.1 Web検索フェーズ

チェック対象となるWebページが他のWebページ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，まずコピー元のWebページ（被剽窃ページ）を探す機能が必要となる．通常，剽窃者は検索エンジンを用いて被剽窃ページを得ると考えられる．そこで，本研究では効率よくコピー元のWebページを探すために，チェック対象のWebページから検索用のキーワードを生成し，Web検索エンジンを用いてコピー元のWebページの候補集合を収集する．

3.2 剽窃ページ判定フェーズ

被剽窃候補のWebページ集合を作成したら，次にチェック対象の被剽窃ページとその剽窃Webページとを比較し，剽窃している可能性を判定する機能が必要となる．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[1]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．

本研究では，Webページをコピーして一部を改変するようなケースを想定しているため，ある文書と表層的な表現が類似する文書が別の文書に含まれる場合に高く評価され，かつ文の入れ替えなどに対応可能な評価指標が必要となる．今回は，分子の部分解析に用いられるSmith-Waterman アルゴリズム[6]を改良し，文書を単語に分割することで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類似文書発見手法を使用する．

3.3 検査者提示フェーズ

剽窃している疑いが高いWebページ（剽窃ページ）とその剽窃元と思われるWebページ（被剽窃ページ）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（剽窃ページ）と被剽窃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，剽窃検査者に提示することによりこれを実現する．

4 システムの実装

本研究では，3節で述べた要件を満たすシステムを実装した．本節ではそのシステムの概要を，それぞれのフェーズごとに分けて説明する．

4.1 Web検索フェーズ

効率的に被剽窃Webページを検索するため，Webに出現する特徴的な単語をキーワードに用いることが望まれる．そこで，本研究では長さで単語をランク付けし上位3件を単語集合 $\{w_1, w_2, w_3\}$ として抽出した[2]．ただし，一般的に検索エンジンに単語を1つだけ与えた場合には，目的のWebページ以外のものが数多く検索されてしまうため，検索エンジンが持つAND・OR検索の機能を用いて複数の単語を用いた検索を行う．今回は，抽出した3つの単語を下記の4種類の論理演算式に当てはめたものを検索ワードとした．

1) $w_1 + w_2 + w_3$
2) $w_1 w_2$
3) $w_1 w_3$
4) $w_2 w_3$

これらは，実際にいくつかのケースで検索を試して，ある程度異なる結果が得られる式を選んだものである[1]．適切な検索ワードの生成方法については今後の課題である．

次に，上記の方法で作成された4つの検索ワードを用いて別々にWeb検索を行う．Web検索エンジンは，指定された検索ワードにマッチするWebページのURLを独自のアルゴリズムに従ってランキング表示する．この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5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，複数の被剽窃候補Webページが得られることになる．

[1] 効率的に検索をするためには，本質的には使用する検索エンジンの検索ランキングの仕組みに基づいて論理式を決める必要がある．しかし，ランキングの仕組みは非公開のものがほとんどであったため，このような手法をとった．

4.2 剽窃ページ判定フェーズ

4.2.1 剽窃発見のための Smith-Waterman アルゴリズム

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している．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 番目の単語と文書 Y の j 番目の単語が一致することを $X(i) = Y(j)$ 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 $S_{j,i}$ とする．

[Smith-Waterman アルゴリズム]

*Step*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．

*Step*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．

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

*Step*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．

*step*1 から *Step*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．

[例 4.1]

2 つの文字列

「XABCXDEXFGHXX」

および

「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の長さが $T = 10$ 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．予備実験によりスコアの最大値を 5 とした．これは欠落・挿入部分が長い場合，剽窃とは言えなくなることを考慮したものである．

4.2.2 入れ替えを検出するための改良手法

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．

本研究では，検出する連続単語系列の長さ (T) を小さくして，入れ替えが行われた場合に連続単語系列を結合する手法を提案する．結合して系列を長くすることにより今まで検出できなかった剽窃とみられる連続単語系列が検出できるようになることが期待される．提案により，以下のアルゴリズムが加えられる．

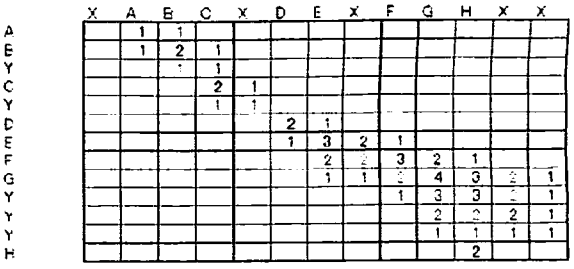

図 1: 剽窃発見のための Smith-Waterman アルゴリズム

[提案アルゴリズム]

*Step*4 連続単語系列の長さが $t(T > t)$ 以上の集合をすべて抽出する．

*Step*5 2 つの系列間にある単語数が L 個以下のとき，系列を結合する．

*Step*6 マージできる集合が無くなれば終了．

本研究では検出する連続単語系列の長さを $t = 3$ 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 $L = 5$ 以内の場合に結合するものとした．

4.3 検査者提示フェーズ

最終的に，剽窃している疑いが高い Web ページとその被剽窃と思われる Web ページを基に，剽窃であるか否かの判断は人手で行う．剽窃検査者の負担を軽減するため，本研究では剽窃検査者に対して 2 ページ間の連続一致単語系列が含まれる部分を 2 文書中から抽出し，提示する．これにより，文書全体に目を通す必要が無くなり，判断も容易となる．

5 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．本実験では擬似的に剽窃 Web ページの代わりに学生から実際に提出されたレポートを用いてシステムの性能を検証した．

5.1 実験データ

本実験では剽窃のしやすさや，編集の容易さの面で Web ページに良く似た特性を持つ学生レポートを用いて実験を行う．

課題：『アウトソーシングサービス事業について所感を述べよ』と『IT 社会の進展と個人情報保護について考察せよ』の 2 課題から学生が自由にどちらかを選択する

レポート数：２０件

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．

5.2 実験結果

全てのレポートにシステムを用いて調査を行ったところ，7件のレポートに剽窃行為の可能性があることが分かった．そこで，それぞれのレポートに対して著作権侵害の可能性のある部分の連続一致単語系列を表示して検査を行った．さらに，引用の有無をチェックし，最終的に6件のレポートが本システムにより剽窃であると判断された．実際に剽窃部分と判断された部分の例を図2に示す．

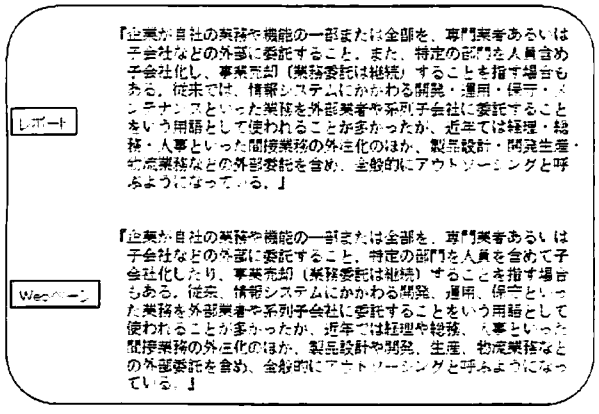

図 2: 検出された被剽窃 Web ページと剽窃候補ページの例

次に，各レポートの最長一致系列長と剽窃レポート件数，加えて本システムで検出できた件数を図3に示す．

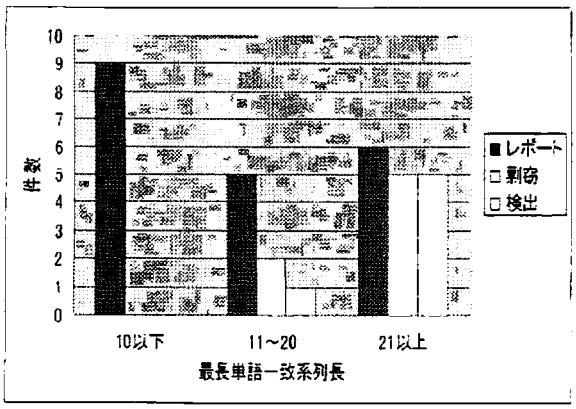

図 3: 最長一致系列長と検出結果

この結果より，本システムは2文書間で行われた剽窃行為に対して，類似部分を正しく発見することができる有効な手段であることが確認できる．

6 今後の課題

- Web ページへの適用

  今回はレポートの剽窃問題に適用しシステムの評価実験を行ったが，今後は実際の Web ページに対して適用し性能を検証する必要がある．

- 同義語・多義後の考慮

  本研究は文単位の入れ替えには対応できるが単語の換言には対応できない．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への改良が必要である．

- 検索キーワードの生成

  今回用いたキーワード生成法でも良い結果が得られることは示したが，キーワードの生成方法を改善する事でさらに良い結果が得られる可能性がある．

  今後はより Web における文書の特徴を考慮したキーワードの生成法を考える必要がある．

参考文献

[1] 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.

[2] 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.

[3] 田代崇，上田高徳，堀泰祐，平手勇宇，山名早人，“Ｗｅｂページを対象とした著作権違反自動検知システム” 信学技報，DE2006-54, 2006.

[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] T. F. smith，M. S. Waterman，“Identification of commonmolecular subsequences”，Jounal of Molecular Biology，147. pp. 195–197，1981 .

[7] 中山 信弘，マルチメディアと著作権，岩波書店,1996.