Web 上の文章を対象とした著作権違反自動検知システム

Copyright Violation Detection System For Web Texts

田代 崇♣ 上田 高徳♥ 堀 泰祐♥
平手 勇宇♦ 山名 早人♠

Takashi TASHIRO Takanori UEDA
Taisuke HORI Yu HIRATE
Hayato YAMANA

近年，Web ページは飛躍的に増加している．その一方で，著作権侵害の Web ページの数も増加しており，問題となっている．そこで本稿では，著作権違反の疑いのあるページを自動検知するシステムを提案する．本システムはシードとなる文章を入力とし，その文章を無断掲載しているページを検出することを目的とする．システムではまず，指定されたシード文章を文節単位に区切り，それを組み合わせることにより検索ワードを作成し，Google,Yahoo!等が提供する検索エンジン Web サービスを用いて著作権違反の候補ページを収集する．次に得られた候補ページとシード文章との間で文節の最大共通部分列を元に類似度を算出し，候補ページをランキングする．評価実験の結果，歌詞，新聞記事，ブログ等をシード文章とした場合は，検出結果上位 20 件に対し，49.1% の精度で違反ページを検出することができた．

Due to explosive increase of the number of web pages, the number of copyright violation web pages has also been increased. To solve this problem, we propose a copyright violation detection system which aims to detect and rank candidate web pages violating copyright that resemble to user's inputted seed page. The system consists of three steps. Firstly, the system generates search keywords based on phrasal units, called "bunsetsu" included in the "seed page." Secondly, on the search keywords, the system gathers candidate web pages violating copyright by using Google or Yahoo! web service. Finally, the system re-ranks the candidate web pages with similarity to the seed page. Here, we adopted "Longest Common Subsequence" of phrasal units, as a similarity measurement. Our evaluation confirmed that the proposed system is able to extract copyright violation web pages with 49.1% accuracy among top 20 ranked web pages.

♣ 学生会員 早稲田大学理工学部ＣＳ学科
ttashiro@yama.info.waseda.ac.jp
♥ 早稲田大学理工学部ＣＳ工学科
ueda@yama.info.waseda.ac.jp,
mail_to_hori@toki.waseda.jp
♦ 学生会員 早稲田大学理工学研究科
hirate@yama.info.waseda.ac.jp
♠ 正会員 早稲田大学理工学部,国立情報学研究所
yamana@yama.info.waseda.ac.jp

1. はじめに

近年，ブログやWiki等，Webページを容易に作成できる環境が普及している．その結果，以前にも増して多くのユーザーがWeb上で情報を発信できるようになった．その一方で,マナーのないユーザーも増え，新聞記事の盗用など，著作権違反のページの存在が問題となっている．Webページが飛躍的に増加している現在では，このような違反ページを人手で監視することは不可能である．そこで本稿では，著作権侵害の疑いのあるページを自動的に抽出するシステムを提案する．違反ページを容易に抽出することができるようになれば，侵害を企てる個人・会社に対する抑止力となることができると考える．

本稿では2節で提案システムの概要について述べ，3節では関連研究を述べる．4節ではシステムの具体的手法について述べる．5節で評価実験結果を示し，6節でまとめを述べる．

2.提案システムの概要

提案システムは，文章を入力とし，その文章と類似したページを全世界の Web ページから抽出すことで，著作権侵害の疑いのある Web ページを抽出するシステムである．本稿では以下，本システムが入力とする文章を，シードパラグラフと定義する．また，全世界の Web ページに対してアクセスするための手段として，検索エンジン Web サービスを用いる．

2.1 シードパラグラフ

本システムが入力とする文章は，歌詞や新聞記事のような，1000 字程度までの文章である[1]．これは，１つのWebページに掲載される程度の量であり，この程度の量の文章が検出できれば，十分に実用可能であると考えられる．

2.2 検出対象の Web ページ

本システムが真に目的とする検出対象は，著作物である文章を無断で掲載しているページである．しかし，著作権の侵害にあたる盗用と，正規の引用を判断することは，人間の目でも難しい．そこで提案システムでは，正規の引用であるか著作権の侵害にあたる盗用であるかは判断せず，「類似した文章を掲載したページ」を検出することを目的とする．

「類似した文章」とは，以下の 2 つに分類できる．

(1) 深層的に似ている文章
(例 1)違う記者が書いた，同じ事件の新聞記事
(例 2)人のアイディアを盗用した論文
(2) 表面的に似ている文章
(例 3)軽微な語尾変化を行った文章
(例 4)英語⇔カタカナなど，表記が変わった文章

(例 2) で上げた例は著作権侵害であるが，このように内容のみを盗用したものは，その文章が盗用なのかそうでないのか判断することは難しい．一方で，(例 3)や(例 4)は，人間が見れば一目で著作権侵害であるとわかることから，計算機でも判別可能であり，また違反である可能性も高い．そこで，本システムが対象とする著作権侵害ページとは，(2) で上げた例のような「表面的に類似している文章を掲載した Web ページ」とし，(1)のような深層的に似ている文章は対象としない．

2.3 検索エンジン Web サービス

本システムで用いたバックエンドの検索エンジンは，

[1] 実用的な処理時間を考慮した文字数であり，1000 文字以上の文章に対しても提案システムは動作する．

Yahoo![1] および Yahoo! JAPAN[2], そして Google [3] の 3 社が提供する 3 種類のサービスである．各社とも Web サービスによる検索サービスを提供している．3 社のサービスの特徴を表 1 に示す．

表 1 検索エンジン Web サービスの仕様
Table1 Web Service Specification of Search Engines

	Yahoo!	Yahoo!JAPAN	Google
アクセス方法	REST	REST	SOAP
アクセス制限方法	IP アドレス	IP アドレス	ライセンスキー
検索制限回数	50000 回/24h	50000 回/24h	1000 回/1day
1 検索当たりの最大検索結果数	100	50	10
1 クエリ内の最大検索語数	非公開[2]	非公開[2]	32

3. 関連研究

文章間の類似性については，これまでに多数の研究が行われている． 八太らは[4]において深層的な類似性を定量化した検索システムを提案している．また，深谷ら[5]や村田ら[6]は，表面的な類似性を定量化している．このように，文書間の類似性を定量化する研究は行われているが，Web 上で公開されているテキストを対象とした場合，こうした類似性判定を直接適用することができない．その理由は，比較対象となるテキストが Web 上に存在し，手元に存在しないためである．

Web 上に存在する多数のテキストを対象に類似度判定を行うためには，自前で Web 上のテキストを収集するか，商用検索エンジンを利用して比較対象候補となる Web ページを収集しなければならない．前者のアプローチは全世界の Web ページ数が 400 億ページと推測[7]されており現実的ではない．このため，本論文では後者のアプローチをとる．

後者のアプローチを用いた手法としては，宮川ら[8]の学生レポートを対象とした手法が提案されている．[8]では，レポートに特徴的な語を抽出し，検索エンジンのクエリとして用いることにより，比較対象となる Web ページを収集している．このように比較対象とする文章から専門性の高い特徴語を抽出できる文章に限定した場合，[8]は効果的な手法である．

しかし，本論文で対象とする文章は，新聞記事，歌詞，百科事典など一般的な文章であるため，[8]のような特徴語を抽出することが困難であり，そのまま[8]の手法を著作権違反自動検知システムに採用することができない．

また，検索エンジンを用いて類似度判定対象となる Web ページを収集するためには，以下のような検索エンジンが持つ制約を考慮した手法が必要不可欠となる．

(1)クエリと正確に一致した文章しか検出されない[3]
(2)クエリ内の検索ワード数の制限がある

以上，本論文で提案する手法の新規性は，類似度判定の対象となる Web ページを現存の商用検索エンジンが持つ制約を満たしつつ収集する点にある．

4. 提案システム

提案システムは，以下に示す 3 つのステップで著作権違反ページを検出する．

(1) シードパラグラフを基にした検索ワード生成．
(2) 商用検索エンジンが提供する Web サービスを用いた著作権違反ページの候補取得．
(3)本稿で提案する類似度を基にした(2)の候補ページの並び替え．

なお，(1)，(3)では，シードパラグラフや得られたページを文節の列として表現する．そして，文章間の一致する文節に着目することで違反ページの検出を行なう．名詞や動詞などの自立語の集合として文章を表現することも考えられるが，文章間の自立語が一致しただけでは，文章が表面的に似ているのか，深層的に似ているのか区別することはできない．また，文節を用いることで文章を構成している単語が一般的であっても文章を絞り込むことが可能である．したがって本システムでは，文節列として文章を表す．また文節列を作成する際，形態素解析器として茶筅[9]を用いた．

4.1 検索ワードの生成

文節列を元に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，以下の点を考慮する必要がある．

(1) 検索結果を絞り込むこと．
(2) 改変されたページも抽出すること．

(1)の観点からは，文節 1 つではなく，複数の文節を結合することで，なるべく長い検索ワードを生成する方がよい．一方，(2)の観点からは，長い検索ワードは改変された部分を含む可能性が高くなり，改変されたページを抽出できなくなってしまうため，短い検索ワードの方が適していることになる．

(1),(2)両方を考慮することを考えると，検索ワードは，文章の改変された部分を含まないような，できるだけ長いもので，かつ，検索結果が絞り込めるものがよい．しかし，文章のどの部分が改変されるかは，シードによって異なり，一概に特定することは不可能である．

以上を踏まえ，著作権違反候補ページを一回の検索で抽出するのではなく，複数回の検索結果の和集合として抽出することを考える．シードパラグラフの文節列から文節単位の N グラムを生成し，それらを各回の検索ワードとする．これにより，連続する N 文節を用いた検索が可能となり，(1)の絞り込みを実現しつつ(2)の改変されたページを含むことができると考える．

検索ワード生成アルゴリズム

(1)シードパラグラフを n 個文節列 $L_{in} = a_0, a_1, ..., a_{n-1}$ に分割する．
(2) $i \leftarrow 0$ とする．
(3) a_i から連続する 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，$i \leftarrow i+1$ として，(3) を繰り返す．

この手法では，被検索文章内の連続する k 個の文要素が，シードパラグラフのものと同じであれば検出ができる．

4.2 検索ワードからの Web ページの取得

生成された検索ワードを検索エンジンへのクエリとし Web ページを取得する．この際，表 1 で示すように，検索エンジンには一度に取得できるページ数に制限があり十分な数の Web ページを取得することができない．また，検索結果数が膨大である時，上位何件を取得するかも重要な問題である．そこで，以下に示すように 4.3 節で定義する類似度がある一定値以下になるまで上位から取得することにした．なお，以

[2] 8000 文字程度までは実験的に確認

[3] Yahoo! JAPANでは，カタカナ表記のゆれなど，多少の違いには対応している．

下のアルゴリズム中，N は１回のクエリで取得するページ数を表す．

ページ取得アルゴリズム

(1) R ← 1
(2) 生成された検索ワードのランキング上位 R 番目から R+N 番目のページを検索エンジンにより取得する．
(3) もし，取得されたページが N 件未満であれば終了
(4) N 件以上取得でき，かつ，N 件の平均類似度がある閾値 Th 超えた場合，R ← R+N とし，(2) へ

このページ取得アルゴリズムは4.1で生成されるすべての検索ワードに対して実行される．多くの類似ページを取得できる検索ワードに対し､多くのWebページを取得することで，効率的に類似性の高い Web ページを取得することができる．

4.3 類似度解析

ブログや掲示板など，一般的に Web ページは１つの文章にのみで構成されることは少なく，複数のテーマの文章や単語が１つのページに存在する．したがって類似度は，シードパラグラフ全体に対し Web 上の文章の一部に対する類似性を定量化したものが望ましい．また，改変されたとしても，単語や文の出現順序は変わらず，以下に示すような改変にとどまるものと考えられる．

・ 英語，カタカナ，ひらがな，などの表記の違い
・ コメントや，ルビの挿入など，文章の内容を変えない，補足的な意味を表す文や単語の挿入

以上のことをもとに，本システムでは，類似度を以下のように定義する．

シードパラグラフの文要素列 L_{in}，被検索文章の文要素列を L_{web} とした時，シードパラグラフから見た，被検索文章の類似度 $Sim(L_{in}, L_{web})$ を以下の式で定義する．

$$Sim(L_{in}, L_{web}) = \log_2 \left\{ \frac{|Lcs(L_{in}, L_{web})|}{|L_{in}|} + 1 \right\}$$

シードパラグラフと被検索文章が完全に一致した場合には１になり，全く類似性が無い場合には 0 となる．

$Lcs(L_{in}, L_{web})$ は列 L_{in}，L_{web} の最長共通部分列(以下 LCS)を表す．具体的な LCS の定義は本節最後に示す．

LCS はシードパラグラフと被検索対象の文節列の両方に含まれる文節列である．したがって，被検索対象の文節列の前後や内部に，コメントやルビなど、シードと全く関係のない文節列が挿入されても，LCS の長さに変化はない．また，部分引用したときのように、被検索文章中にシードに含まれる文節が少なければ，LCS の長さは短くなる．

LCS(最長共通部分列)の定義

LCS とは，配列間の類似した部分列で，定義は以下である．．

配列 $a_0, a_1, ..., a_{n-1}, b_0, b_1, ..., b_{m-1}$ について，

$$a_{i_0} = b_{j_0}, a_{i_1} = b_{j_1}, ..., a_{i_{L-1}} = b_{j_{L-1}}$$

を満たすように，配列のインデックスの列

$$0 \le i_0 < i_1 < ... < i_{L-1} \le n-1$$
$$0 \le j_0 < j_1 < ... < j_{L-1} \le m-1$$

を選んだ時，$a_{i_0}, a_{i_1}, ..., a_{i_{L-1}}$（または $b_{j_0}, b_{j_1}, ..., b_{j_{L-1}}$）を共通部分列といい，その中で最長ものを LCS という．

5. 評価実験

5.1 実験内容

実験は，以下の手順により行った．

(1) ３つの検索エンジンで，侵害ページの候補を収集．
(2) ３つの検索エンジンの検索結果をマージ
(3)提案手法の類似度によりランキングした，候補ページの上位 20 件に対してその精度を人手により確認

(3)の確認作業では，以下の項目について調査した．

・表 2 に基づき，実際に似ているかを人手により判断(正規の引用かは問わない)
・ 類似しているページの中で，正規の引用でない，著作権侵害ページにあたるかどうかを判断

実験に使ったシステムのパラメータは以下の通りである．

・検索ワードの文節数： k = 2
・一度に取得するページ数: N = 10
・次の N 件を取得する閾値: Th = 0.2

提案手法のランキングに対する評価として，DCG[10]を用いた．DCG はランキング対象の文書集合がどの程度適合度順に並んでいるかどうかを判断することを目的としたランキング評価手法である．本実験ではランキングの上位 20 件について表 2 の一致度をもとに評価した．評価にあたっては，ランキング上位の 20 件が表２の一致度順に完全に並んでいる場合を１として正規化した．これは，提案した類似度がどの程度人手によるランキングに近いかを表している．

表 2 一致度の定義
Table2 Definition of Relevance Measurements

一致度	説明
3	シードと 8 割以上が一致しているもの
2	シードと 3 割から 7 割一致しているもの
1	一致度 3 と 4 以外で，シードパラグラフの転載と分かるもの
0	全く関係のないもの

5.2 シードパラグラフ

評価実験で用いたシードパラグラフは，歌詞 20 件，新聞記事 15 件，歌詞・新聞記事以外の Web コンテンツ 15 件の合計 50 件である．

歌詞は，[11]より無作為に 2000 曲選出し，Yahoo! Japan の検索エンジンを用いて「アーティスト名 and 曲名」の検索結果数を歌詞の有名度合いとし，有名な曲からマイナーな曲まで選出されるよう，人手で選択した．

新聞記事のシードパラグラフは，調査を行なった 2006 年 2 月 7 日から 1 ヶ月前の新聞記事を中心に，10 年前の古い新聞記事も含めて[12],[13],[14]から計 15 本の新聞記事を選択した．最新の記事は検索エンジンによってインデックス化されてないため，シードとして選ばなかった．

また，歌詞・新聞記事以外のシードパラグラフとして，Wikipedia[15]・有名人のブログ・公的機関の Web ページの３分野から合計 15 件を選択した．

5.3 実験結果

抽出結果 50 事例の上位 20 件，計 1000 ページの確認結果を表 3 に表し， 1000 ページの一致度の内訳を図 1 に示す．

表 3 では，検出されたページのうち，49.1%が違反ページであることがわかる．中でも歌詞においては検出結果上位 20 件において，7 割以上の精度で違反ページを検出することができた．新聞記事やその他の事例では，違反ページは 3 割と歌詞と比べて少ないが，これは新聞記事や Wikipedia などは，正式に引用すれば違反とならないからである．正規の引用ペ

ージ（一致度 1 以上のページ）も正解とすれば，歌詞，新聞記事，その他の事例において，上位 20 件の精度は 70%以上であり，実用可能な精度といえる．

また，歌詞の事例において抽出された違反ページの中には，替え歌をしているもの，男女パートや歌詞に対するコメントなど歌詞に関係の無い単語，文の挿入があるものなどが検出された．また，新聞記事やその他の事例においては，文体を変えて自分が書いたかのように掲載しているもの，有名人ブログをパロディにしているもの．などが検出された．これは，提案手法が文章の改変にロバストであることを示している．

また，表 3 の正規化した DCG を見ると，歌詞，新聞記事，その他の事例ともに 0.9 以上の値を示している．これは DCG の観点から，人手によるランキングと提案手法によるランキングが 9 割以上同じことを示している．

表 3 50 事例における上位 20 件の確認結果
Table3 Results of Top 20 Ranked Pages Based on the 50 Seed Paragraphs

	歌詞	新聞記事	その他	すべて
事例数	20	15	15	50
確認した全ページ数	400	300	300	1000
一致度 3	261	155	169	585
一致度 2	37	62	70	169
一致度 1	11	55	20	86
一致度 0	91	28	21	140
違反数	284	102	105	491
違反数/全ページ数	0.710	0.340	0.350	0.491
一致度 1 以上/全ページ数	0.745	0.723	0.797	0.754
正規化した DCG	0.923	0.999	0.997	0.973

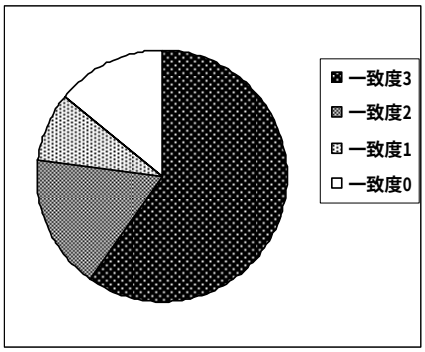

図 1 確認した 1000 ページの一致度の内訳
Fig.1 Relevance Measurement Distribution of the 1000 Pages which were Manually Checked

6. おわりに

本稿では，著作権侵害の Web ページを自動検知するシステムを考案し，そのシステムの評価を行なった．その結果，上位 20 件に対して，49.1%の精度で著作権侵害ページを検出することができた．また，ランキングの評価では，人手よるランキングの DCG に対する，提案手法によるランキングの DCG の割合は，0.975 であった．これらから，本システムが容易に著作権侵害ページを検出できることを示した．

検出されたページのほとんどは，一般ユーザーが気軽に Web 上に掲載してしまった文章であり，その個人に対して著作権侵害を問うことは難しい．しかし，このようなページが容易に検出できれば，違反が多いホストの管理者や作成者個人に対する警告も容易に行なうことができる．その結果，著作権侵害の抑制に役立てることができると考える．

[謝辞]

本研究の一部は経済産業省「ＩＴによる「情報大航海時代」の情報利用を考える研究会」の先導研究として実施した．

[文献]

[1] Yahoo!, http://www.yahoo.com/
[2] Yahoo! JAPAN, http://www.yahoo.co.jp/
[3] Google, http://www.google.com
[4] 八太絵美, 福本徹, 横山節雄, 赤堀侃司: “文書間の類似度に基づく論文検索システムの開発と評価”, 日本教育工学会研究報告集, pp.91-96(2002).
[5] 深谷亮, 山村毅, 工藤博章, 松本哲也, 竹内義則, 大西昇: “頻度統計と概念辞書を用いた文章の類似性の定量化情報処理学会研究報告”, Vol.153, No.4, pp.73-79(2003).
[6] 村田哲也, 黒岩大輔, 高橋勇, 白井治彦, 小高知宏, 小倉和久: “学生レポートの n-gram による類似度評価の検討”, 情報科学技術フォーラム, pp.101-102(2002).
[7] 加藤真, 山名早人: “Fact of the Web – 30 億ページの Web 解析”, 第 17 回電子情報通信学会データ工学ワークショップ, ３B-i6 (2006)
[8] 宮川勝利, 高橋勇, 小高知宏, 白井治彦, 黒岩丈介, 小倉久和: “Web ページの剽窃により作成された学生レポートの検出手法の提案”, 教育システム情報学会研究報告, Vol.20, No.4, pp.33-40(2005).
[9] 形態素解析システム茶筌, http://chasen.naist.jp/hiki/ChaSen/
[10] Kalervo J¨arvelin and Jaana Kek¨al¨ainen: ”IR evaluation methods for retrieving highly relevant documents, ” Proc. of the 23rd Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp.41-48(2000).
[11] 歌ネット, http://www.uta-net.com/
[12]Yahoo! NEWS, http://headlines.yahoo.co.jp/accr?ty=t&c=all
[13] 聞蔵 DNA for Libraries, http://database.asahi.com/library/
[14] Impress Watch, http://www.watch.impress.co.jp/
[15] Wikipedia, http://ja.wikipedia.org/wiki/

田代 崇 Takashi TASHIRO
早稲田大学理工学部ＣＳ学科在学中．日本データベース学会会員．

上田 高徳 Takanori UEDA
早稲田大学理工学部ＣＳ学科在学中．

堀 泰祐 Taisuke HORI
早稲田大学理工学部ＣＳ学科在学中．

平手 勇宇 Yu HIRATE
2005 早稲田大学大学院理工学研究科修士課程修了．同大学同研究科博士課程在学中. 2006 年より同大学メディアネットワークセンター助手．ACM，情報処理学会，日本データベース学会各学生会員．

山名 早人 Hayato YAMANA
1993 早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．1989-1993 同大学情報科学研究教育センター助手．1993-2000 電子技術総合研究所．2000 早稲田大学理工学部助教授. 2005 同大学理工学術院教授, 現在に至る. IEEE, ACM，IEICE，IPSJ，日本データベース学会各会員