社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DBS－140 (II) (34)
2006／7／13

Webページを対象とした著作権違反自動検知システム

田代 崇[†] 上田高徳[†] 堀 泰祐[†]
平手勇宇[††] 山名早人[†,†††]

近年のWebページ総数の飛躍的な増加に伴い，歌詞や新聞記事の無断引用などの著作権侵害のWebページの数も増大している．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出するシステムを提案する．本システムではまず，検索ワードを，指定された文章を文節単位に区切り組み合わせることにより生成し、GoogleやYahoo!が提供している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．次に候補ページを類似度をもとにランキングを行ない，ユーザーに提示する．ランキングに用いた類似度は文節をもとにした最長共通部分列から求める．評価実験を行った結果，歌詞，新聞記事，ブログ等からなるWebページをシードとして，著作権侵害ページを検出することができた．

Copyright violation detection system for Web texts

TAKASHI TASHIRO,[†] TAKANORI UEDA,[†] TAISUKE HORI,[†]
YU HIRATE[††] and HAYATO YAMANA [†,†††]

Due to explosive increase of the number of web pages, the number of copyright violation web pages, such as lyrics or news citation pages without permission, has also been increased. To solve this problem, we propose a system for detecting copyright violation web pages. The proposed system consists of three steps. Firstly, the system generates search keywords on phrasal units, called "bunsetsu", which are included in the "seed page." Secondly, on search keywords generated by the first step, the system gathers candidate of 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 proposed system is able to extract copy violation web pages correctly.

1. はじめに

近年のブログ等の普及によりWebページの作成は容易になり，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．その結果、Webページの質は多様となった．その中でも，新聞記事の盗用や，歌詞の無断引用などの著作権違反のページの存在が問題となっている．このようなページを手作業で検出するためには，検索エンジンなどを用いて検索を行い，検索した結果のページの中から著作権侵害が行なわれているページを検出する必要がある．しかし，検索エンジンに入力するためのキーワードを，検索元となる文章から抽出し，得られたページを一つ一つ確認するのは，多大な労力を必要とする。Webページの総数が膨大になった現在では，このような作業は困難である．そこで本稿では，このような作業を自動化するシステムを考案した．

本稿では2節で抽出対象とする著作権侵害ページについて述べ、3節では関連研究を述べる。4節ではWebサービスを通してページを収集するための「検索ワード生成手法」と，ランキング付けのための「類似度解析」を述べる。5節ではシステムの実装について述べ、6節で評価実験結果をのべ、7節でまとめを述べる．

† 早稲田大学理工学部
Science and Engineering, Waseda University
†† 早稲田大学大学院理工学研究科
Graduate School of Science and Engineering, Waseda University
††† 国立情報学研究所
National Institute of Informatics

2. 著作権侵害ページ

本節では，検索元となる文章であるシードパラグラフと，検出対象のページについて述べる．

2.1 シードパラグラフ

提案システムは，著作物となる文章を入力とし，著

作権侵害の疑いのあるWebページを抽出する．

Web上の著作権侵害ページは，論文などの専門的な文章の盗用から，新聞記事の盗用などまで，さまざまなものが考えられる．したがって，シードパラグラフは，専門性が高いものだけでなく，歌詞や新聞記事などの専門性の低い，一般性の高いものも想定する必要がある．

2.2 検出対象のページ

文章がそのままコピーされる場合以外にも、Webページの作成者によって改変がなされている場合があるが，これらは主に以下の2つに分けられると考えられる．

(1) 著作物である文章のアイディアや内容を盗用し，新たに文章を書く．

(2) 著作物である文章を表面的に変化させる．

(1)のような，深層的な内容の盗用は，それが盗用であるか判断することは難しい．そこで，本稿が対象とする改変された著作権侵害のページとは，(2)のような表面的に変化させたものとする．

また，文章がコピーされたページが，正規の引用か著作権の侵害にあたる盗用なのか判断することは，人間の目でも難しく，計算機で行なうことは困難である．そこで，本稿では，そのような正規の引用か著作権の侵害にあたる盗用かは判断はしない．

3. 関連研究

2.2で述べた，表面的に変化させた文章の類似度の既存研究として，与えられた2つの文章の類似性を定量的に示すための類似度に関する研究[1),3)]がある．しかし，[1),3)]では類似性を判定する文章は全て手元にあるのに対し，本稿で提案するシステムでは著作権違反文章はWeb上に存在する点で異なる．

また，[2)]ではWeb検索エンジンを使って学生レポートの剽窃を判定するシステムを提案している．しかし[2)]で提案している手法と提案手法は以下の2つの点において異なる．

(1) [2)]では，コピーされた文章は全て手元にあり，著作権がある元の文章がWeb上にある．これに対し，提案システムでは著作権があるもとの文章が手元にあり，コピーされた文章はWeb上にある．

(2) [2)]では，対象を学生レポートとしているため，検出対象とする文章は，特徴語などを多く含む専門的な文章である．これに対し提案システムでは，歌詞，新聞記事記事といった一般的な単語を多く含む一般的な文章も検出される必要がある．

したがって，提案システムは，[1),2),3)]と比較して以下の特徴がある．

- 著作権違反検出対象文章はWeb上に存在する．
- 一般的な単語で構成されている文章も，検出対象文章とする．

4. 類似ページの収集とランキング付け

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) 指定されたシードパラグラフを基にして検索ワードを生成する．

(2) 商用サーチエンジンが提供するWebサービスを用いて著作権違反ページの候補を取得する．

(3) 取得したページの集合から，本稿で提案する類似度を基にして結果の順位を並び替える．

本節では，まず4.1で基本事項を述べる．つぎに，4.2で検索ワードの生成手法を述べ，4.3に類似度の計算手法について述べる．

4.1 基本事項

4.1.1 検索エンジン Web サービス

本システムでは，バックエンドの検索エンジンとしてYahoo![6)]およびYahoo! JAPAN[7)]，そしてGoogle[8)]の3社が提供する3種類のサービスを利用した．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．Googleが提供する検索サービスは、SOAPによりアクセスする．利用にあたっては，ライセンスキーが必要であり、1ライセンスあたり1日1000回まで検索が可能である．Yahoo! 及びYahoo! JAPANが提供するWeb検索サービスはRESTと呼ばれる方法で利用でき、24時間の間に1つのIPアドレスあたり5,000回までの検索が可能である。1度の検索でYahoo!では最大100件、Yahoo! JAPANでは最大50件までの検索結果を取得することができる。2社が提供するこのサービスはGoogleに比べ検索結果を返すまでの時間が短い．特に、Yahoo! JAPANのWeb検索サービスは日本向けのサービスとなっているため，日本語で検索した場合のヒット件数がYahoo!、Googleに比べて多い。3社のサービスの特徴を表1に示す．

4.1.2 文要素

一般的に文章を解析するためには，次のような，ある特徴をもった要素の列として表す必要がある．

- 形態素
- 名詞，動詞などの自立語

☆ URLの長さの制限(8000文字程度)までは，可能であった．

表 1 検索エンジンの機能

	Yahoo!	Yahoo! JP	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
検索回数	5 千回/24h	5 千回/24h	千回/1day
検索結果取得数	100	50	10
検索速度	高速	高速	低速
日本語サイトの数	少ない	多い	少ない
最大検索ワード	非公開 ☆	非公開 ☆	32

- 文節

このような要素を，本稿では一般に「文要素」と呼ぶことにする．すべての文章は，ある文要素の列として表すことができる．なお，本システムの形態素解析器として，茶筌[5] を用いた．

4.2 検索ワードの生成

4.1.1 で述べた Web サービスを用いて著作権侵害の候補ページを集める場合，以下のような問題がある．

- 検索結果が多い場合，現実的に検索結果の上位ページしか取得できない．
- 2.2 のように，改変された文章の存在．
- 検索エンジンの，検索ワード数制限．

以上の問題に解決するためには，検索ワードを生成する時に以下の点を考慮しなければならない．

- 検索結果を絞り込み，上位に候補ページを集めること．
- 文章の改変を許容できること．

したがって，検索ワードを作る際には，文章の改変された部分を含まないような，出来るだけ長い（条件を絞り込める）検索ワードを作成することが望ましい．しかし，文章のどこが改変されるかは，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．

そこで，文章を文要素単位に分割し，下のようなアルゴリズムにより，検索ワードを生成する．

検索ワードの生成アルゴリズム

(1) シードパラグラフを n 個の文要素列 $L_{in} = a_0, a_1, ...a_{n-1}$ 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．

このように文章を小さな単位に分割し，一つずつずらしながら網羅的に検索ワードを作成することによって被検索文章の中で改変をしていない部分に対してヒットさせることができる．その結果，改変が存在する文章も，部分的に改変されていない部分が存在すれば，検出できる．また，シードの一部のみしかコピーしていないページに対しても，この検索手法は有効である．具体的に，この検索手法では，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．

4.3 類似度解析

4.2 で述べた，網羅的に検索ワードを生成する手法は，改変されたページが検出される分，全く関係のないページが結果として含まれる可能性がある．そこで，シードパラグラフと被検索ページの類似性をもとにランキング付けを行なう．以下では本システムで用いた類似度の計算手法について述べる．

4.3.1 表面的な改変

本システムが主に対象としている歌詞や新聞記事などの転載時の改変は，以下に示す改変にとどまるものと考えられる．

- シードの一部分のみのコピー
- 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
- 漢字からカタカナや，英単語からカタカナ表記など，同音であるが「別表記表現」への置換
- 替え歌などの主語や動詞の，「全く別の意味」への置換

このような改変が行なわれた時，当然，シードの一部分のみをコピーしたものよりは，シードの全体をコピーしたもののほうが類似度が高くなることが望ましい．しかし，コメントのルビなど，補足的な意味を表すものの挿入については，類似度が変化することは望ましくない．

このような特性を数値的に表すものとして，シードパラグラフの文要素列と被検索対象の文要素列との最長共通部分列 (Longuest Common Subsequence 以下 LCS) の長さを用いた．

4.3.2 最長共通部分列 (LCS)

LCS は，要素列の類似性を数値化したものであり，以下のように定義される．

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

$$0 \leq i_0 < i_1 < ... < i_{L-1} \leq N-1$$
$$0 \leq j_0 < j_1 < ... < j_{L-1} \leq M-1$$

を選んだ時，$a_{i_0}, a_{i_1}, ..., a_{i_{L-1}}$（または $b_{i_0}, b_{i_1}, ..., b_{i_{L-1}}$）を共通部分列といい，その中で最長ものを最長共通部分列 (LCS) という．具体的な例は以下のようになる．

配列 1: 2,0,0,6,7,1,2
配列 2: 0,2,1,0,2,7,1
LCS : 2,0,7,1

このLCSを，シードパラグラフと被検索対象の文要素列に対して適用する．このとき、LCSはシードパラグラフと被検索対象の文要素列の両方に含まれる文要素列である．したがって，被検索対象の文要素列に，シードと関係のない要素（ルビなど）が挿入されても、LCSの長さに変化が起こることはない．また，被検索文章の文要素列の中にシードに含まれる文要素が存在しなければ（つまり部分的に引用している場合には)、LCSの長さは短くなる．

4.3.3 類似度の定義

4.3.2で述べたLCSの長さに対し正規化を行なうことで，次のように類似度を定義する．

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 L_{web} のとき，シードパラグラフから見た，被検索文章の類似度 $Sim(L_{in}, L_{web})$ は，

$$Sim(L_{in}, L_{web}) = \log_2(\frac{LCS(L_{in}, L_{web})}{L_{in}} + 1)$$

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．上式によって計算される類似度は，被検索文章内に，どの程度シードパラグラフと同じ文章が存在するかを表す値になっている．

5. システム実装

本節では，4節で提案した，検索ワードの生成と類似度計算手法をもとにした，システムの実装方法について述べる．

5.1 文要素の選択

実装するにあたって，検索ワードの生成や，類似度解析で用いる文要素を選択する必要がある．文要素の候補となるのは，4.1.2にあげた以下のようなものが考えられる．

- 形態素
- 名詞，動詞などの自立語
- 文節

提案システムが主に対象とするシードページは，歌詞や新聞記事などの専門性が低い文章である．専門性が低い文章では，名詞や動詞などの自立語は一般的なものが多く，特徴語にはなりにくい．したがって，自立語を用いた場合，以下のような問題があると考えられる．

- 検索ワードで用いると，著作権違反ページ候補を絞りこむのは難しい．
- 類似度解析において，一般性の高い単語は，文章がコピーされた部分以外にも一致してしまい，適切な類似度が計算できなくなる．

これに対し，文節を用いると，自立語だけでなく付属語が追加されることで，次のような利点がある．

- 検索ワードで用いると，付属語が追加されただけ，文章を絞り込むことができる．
- 類似度解析でも，シードパラグラフとの一致している部分のみに一致しやすく，適切な類似度の計算ができる．

したがって，検索ワードと，類似度解析では，文要素として文節をもちいる．

5.2 全体構成

システムの全体構成を図1に示す．　以下，図1

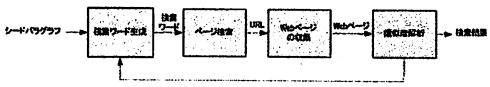

図1　全体構成

における，各処理について述べる．

5.2.1 検索ワード生成とページ検索

4.2で述べたアルゴリズムを用いて検索ワードを生成する．この際検索ワードは，組み合わせる文節をスペースで区切るだけのAnd検索であり，フレーズ検索は行なわない．このとき，文節を結合する数(4.2で定義したkの値)は，経験的に2とした．次に、Google,Yahoo!の提供するWebサービスを用いて，検索ワードに関連したURLを取得する．また，検索エンジンが提供するWebサービスでは，検索結果を最大N件しか取得することが出来ない(GoogleではN=10，Yahoo!ではN=100、Yahoo! JAPANではN=50)．そこで次のような手法をとった．

(1) R ← 1

(2) 検索ワードを元に，検索結果R番目からR+N番目のページを取得する．

(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(システム固定)を超えた場合，$R \leftarrow R + N$ とし，(2)へ戻る．

なお，この時用いた閾値は，本システムでは経験的に0.2として実装した．

5.2.2 ページ収集

URLを元に，実際のWebページを取得する．

5.2.3 類似度解析

得られたWebページを，4.3の手法により，類似度を計算しランキング付けを行なう．

6. 評価実験

評価実験では実際に違反ページの抽出を行い，シス

表 2　一致度の定義

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，シードの転載と分かるもの
0	全く関係のないもの

テムの評価をおこなった．なお，類似度によるランキングの評価尺度として、DCG を正規化したものを用いた．

6.1 実験内容

歌詞や新聞記事，公的機関の Web ページなどをシードパラグラフとして，本システムで違反ページの抽出をし，抽出結果の評価を行なった．本システムでは，バックエンドとして用いる検索エンジンに抽出結果が依存する．そこで，システムの性能を最大限発揮させて評価するために，バックエンドとして Google、Yahoo!、Yahoo! Japan の 3 つの検索エンジンを用いた場合の抽出結果の Web ページを URL を元にマージしたものを評価に用いることにした．そして，マージした抽出結果のうち，類似度上位 20 件の Web ページを Web ブラウザで閲覧し，シードパラグラフとの一致度と，著作権違反の確認を人手で行なった．

ここで用いる一致度とは，シードパラグラフと抽出結果の類似の程度を示す 4 段階の指標である．抽出結果がシードの 8 割以上一致している場合を一致度 3 とし，シードの 3 割から 7 割一致しているものを一致度 2 とした．そして，一致度 3 と 2 に当てはまらないが，シードパラグラフの転載と分かるものを一致度 1、全く関係のないものを一致度 0 とした．また，替え歌や新聞記事のパロディなど，シードをもとにしているが，違う意味になっているようなものは，一致度 2 として分類した．

一致度の定義をまとめたものを表 2 に示す．

6.2 使用したシードパラグラフ

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 件である．

6.2.1 歌　　詞

歌詞のシードパラグラフ 20 件を選ぶにあたり，有名な歌からマイナーな歌まで含まれるように次の手順を踏んだ．

(1) 歌詞検索サイト[9] から無作為に 2000 曲選んだ．

(2) (1) で選んだ 2000 曲それぞれについて、Yahoo! Japan の検索エンジンを用いて「アーティスト名 and 曲名」をキーワードに検索した．

(3) 検索結果の件数が多い曲を有名な歌，少ないものをマイナーな歌として、2000 曲の中から (2) で取得した検索結果の件数が偏らないように 20 件を人手で選択した．

選択の参考にした検索結果の件数は，表 3 に「有名度」として示してある．

6.2.2 新聞記事

新聞記事のシードパラグラフとして最新の記事を用いた場合，著作権違反を犯しているページが検索エンジンにインデックス化されておらず，違反ページを抽出できないという問題がある．このため，調査を行なった日から 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合計 15 本の新聞記事を選択した．

6.2.3 そ の 他

歌詞・新聞記事以外のシードパラグラフとして、Wikipedia[13]・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．

6.3 類似度の評価尺度

類似度によるランキングの評価尺度として、DCG を正規化したもの用いた．

DCG(Discounted Coumulative Gain)[4] は Järvelin らによって考案された検索システムの評価尺度であり，以下のように定義される．

$g(i)$ が i 番目にランクされた文書の得点のとき，

$$dcg(i) = \begin{cases} g(1) & if\ i = 1 \\ dcg(i-1) + g(i)/\log_c(i) & otherwize \end{cases}$$

これは，得点が高い，つまり適合した文書，がランキング上位に多くあれば高い数値を示すことになる．本実験では，以下のようにパラメータを指定した．

- 得点 $g(i)$：ページの一致度
- $c = 2$

また、DCG は値の範囲が定まっていないため，次のように正規化を行なった DCG_n を評価として用いた．本システムの検索結果ランキング上位 N 件に対して，本稿の類似度でランキングした場合の DCG を $DCG_s(N)$，一致度でランキングした場合の DCG を $DCG_c(N)$ で表す時，

$$DCG_n(N) = \begin{cases} 1 & if\ DCG_c(N) = 0 \\ DCG_s(N)/DCG_c(N) & otherwize \end{cases}$$

これは，本稿の類似度が，どの程度人手による評価である一致度に近いかを表した数値になっている．

6.4 実験結果

歌詞をシードパラグラフとした時の上位 20 件の検出結果を，表 3 に，新聞記事をシードパラグラフとした時の上位 20 件の検出結果を，表 4 に，Wikipedia[13]・

有名人のブログ・公的機関のWebページをシードパラグラフとした時の上位20件の検出結果を表5に示す.

表3 歌詞をシードとした時の上位20件の検出結果

No	有名度	一致度別ページ数				違反数	DCG_n
		3	2	1	0		
1	942000	19	1	0	0	20	1.000
2	350000	20	0	0	0	17	1.000
3	217000	19	1	0	0	20	0.996
4	107000	20	0	0	0	19	1.000
5	104000	20	0	0	0	19	1.000
6	55500	20	0	0	0	20	1.000
7	33200	17	0	3	0	20	1.000
8	30400	19	0	0	1	18	1.000
9	25800	11	8	1	0	19	0.999
10	20900	20	0	0	0	14	1.000
11	20400	20	0	0	0	20	1.000
12	18700	16	4	0	0	19	0.997
13	13100	20	0	0	0	15	1.000
14	9150	14	6	0	0	20	1.000
15	6300	1	4	0	15	4	1.000
16	1140	0	0	1	19	0	0.263
17	728	2	8	1	9	10	0.905
18	612	1	4	1	14	6	0.980
19	550	2	1	3	14	4	0.964
20	16	0	0	1	19	0	0.356

表4 新聞記事をシードとした時の上位20件の検出結果

No	記事発行日	一致度別ページ数				違反数	DCG_n
		3	2	1	0		
21	95/1/17	0	0	0	20	0	1.000
22	05/3/25	5	2	9	4	6	0.996
23	05/5/17	5	13	2	0	4	0.999
24	05/6/11	10	0	10	0	2	1.000
25	05/6/28	13	3	1	3	15	1.000
26	05/12/9	19	1	0	0	19	0.999
27	06/1/7	12	8	0	0	14	0.998
28	06/1/9	17	3	0	0	6	1.000
29	06/1/15	8	7	4	1	10	1.000
30	06/1/22	20	0	0	0	11	1.000
31	06/1/23	9	11	0	0	4	1.000
32	06/1/28	10	2	8	0	4	1.000
33	06/1/29	3	3	14	0	4	1.000
34	06/2/3	14	4	2	0	1	1.000
35	06/2/6	10	5	5	0	2	1.000

表5 歌詞, 新聞記事以外のWebコンテンツをシードとした時の上位20件の検出結果

No	種別	一致度別ページ数				違反数	DCG_n
		3	2	1	0		
36	Wikipedia	20	0	0	0	5	1.000
37	Wikipedia	20	0	0	0	14	1.000
38	Wikipedia	18	2	0	0	2	1.000
39	Wikipedia	10	9	0	1	4	1.000
40	Wikipedia	20	0	0	0	16	1.000
41	ブログ	20	0	0	0	9	1.000
42	ブログ	2	4	14	0	0	1.000
43	ブログ	0	0	0	0	5	0.996
44	ブログ	4	7	4	5	5	0.980
45	ブログ	14	6	0	0	7	1.000
46	ブログ	14	5	0	1	7	1.000
47	公的機関	4	16	0	0	8	1.000
48	公的機関	10	10	0	0	5	1.000
49	公的機関	3	2	1	14	1	0.985
50	公的機関	10	9	1	0	17	1.000

6.5 抽出結果

6.5.1 歌詞

歌詞をシードにした抽出結果400件のうち, 一致度2以上のページは77%, 著作権違反と判定されたページは71%であった. また, 抽出実験3,10,12,14,17では, 上位20件の中に, 替え歌を掲載しているページが発見された. また、9では, 歌詞のフレーズごとに男女のパートを挿入しているもの、11では, 歌詞のフレーズ間にページ作成者のコメントが挿入されているページが発見された.

6.5.2 新聞記事

表4では, 一致度2以上のページが抽出結果300件の90.6%を占めている. なお, 違反ページが抽出されなかったNo.21では, インターネットが普及していなかった頃の新聞記事である。. また, 新聞記事においても、26,30などで, である体からですます体へ変えてある, 改変されたページが検出できた.

6.5.3 その他のコンテンツ

表5では, 一致度2以上のページが抽出結果300件のうち86.3%を占めている. だが, 著作権違反と判定されたページは41%にとどまる. これは、Wikipediaが引用を許可しており, 正しく引用すれば著作権違反とならないからである. Wikipedia・有名人のブログ・公的機関のWebページをシードパラグラフとした抽出実験41〜50の結果では, 一致度2以上のページは抽出結果200件のうち80%である. 検出例としては、37,44では, 有名人のブログをパロディにしているページもあった.

6.6 考　　察

本実験で実際に確認した全1000ページのうち，著作権侵害のページの割合は49.1%であった．正規の引用であるページや検索元のページまで含めると、86.0%となる．比較対象とするシステムが無いので，客観的な判断は難しいが，実用可能な性能であるといえる．

また，歌詞，新聞記事，その他のコンテンツともに，替え歌や文体を変えたものなど，改変されたページが検出でき，本稿の検索ワード生成手法が有効であることが確かめられた．類似度解析では、No.16,20などでは、ランキング下位になってしまった一致度1のページがあるが，これらのページは歌詞のワンフレーズのみの転載である。著作権侵害の疑いの強い、一致度3や2のページについては、No.17,18,19,23,44,49など，上位20件のページがさまざまな一致度が混在しているような検索結果に対しても，DCG_n が9割以上の値を示している．これは，本稿が提案する類似度と，人手による判断がおおよそ一致しているものだと考えられ，類似度の有効性が確認できる．

7. 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．本システムはシードとなる文章を指定することで，著作権侵害の疑いのあるページを収集し，類似度をもとにランキング付けをおこなう．ページを収集する段階では，検索ワードを生成し，そのワードをGoogleやYahoo!が提供しているWebサービスに問い合わせることで著作権侵害の候補ページを取得する．検索ワードは，改変に対応するため，シードパラグラフを文節単位に分割し，組み合わせることで作成する．また，類似度は，シードパラグラフの文節列と，被検索対象の文節列のLCSによるものである

本稿ではまた，提案システムの評価を行なった．その結果，本システムによる著作権侵害ページの検出が実用的な性能であることが確認された．また，検出結果には，替え歌や文体を変えたものも含まれ，提案された検索ワード生成手法が，改変に有効であったことも示された．

謝　　辞

本研究の一部は経済産業省「情報大航海」プロジェクトの先導研究として実施した．

参 考 文 献

1) 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).
2) 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).
3) 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).
4) Kalervo Järvelin & Jaana Kekäläinen, "IR evaluation methods for retrieving highly relevant documents, "Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.41-48(2000).
5) 形態素解析システム茶筌, http://chasen.naist.jp/hiki/ChaSen/
6) Yahoo!,http://www.yahoo.com/
7) Yahoo! JAPAN,http://www.yahoo.co.jp/
8) Google,http://www.google.com
9) 歌ネット,http://www.uta-net.com/
10) Yahoo! NEWS, http://headlines.yahoo.co.jp/accr?ty=t&c=all
11) 聞蔵 DNA for Libraries, http://database.asahi.com/library/
12) Impress Watch, http://www.watch.impress.co.jp/
13) Wikipedia,http://ja.wikipedia.org/wiki/