音楽配信サイトを用いた違法配信コンテンツの推定

阿部　佑樹† 　仮屋　慶太†† 　糸川　剛† 　北須賀輝明† 　有次　正義†

† 熊本大学大学院自然科学研究科 〒 860-8555 熊本県熊本市黒髪 2-39-1
†† 熊本大学工学部情報電気電子工学科 〒 860-8555 熊本県熊本市黒髪 2-39-1
E-mail: †{abec,kariya}@dbms.cs.kumamoto-u.ac.jp, ††{itokawa,kitasuka,aritsugi}@cs.kumamoto-u.ac.jp

あらまし　現在，ウェブ上に著作者に承諾を得ず，違法に配信されている音楽や動画が多々ある．そのようなデータそのものを解析して違法か判定を行う手法があるが，ウェブ上のデータを解析するにはコストがかかる．本研究では，違法なデータを配信しているコンテンツを推定することを目的とする．合法的にデータを配信しているサイトのリストは著作権を管理している団体が管理していると考え，合法的にデータを配信しているコンテンツから特徴量を抽出し，違法な配信を行っているコンテンツの推定を行った．特徴量には名詞に基づくものとアンカーテキストに基づくものを用いた．実験により両特徴量に差はあるが，違法な配信を行っているコンテンツを推定することができることを確認した．

キーワード　違法配信サイト，著作権

Detecting Illegal Webpages Using Legal Online Music Sites

Yuki ABE†, Keita KARIYA††, Tsuyoshi ITOKAWA†, Teruaki KITASUKA†, and Masayoshi ARITSUGI†

† Graduate School of Science and Technology, Kumamoto University
2-39-1 Kurokami, Kumamoto, Kumamoto 860-8555, Japan
†† Dept. of Computer Science and Electrical Engineering, Kumamoto University, 2-39-1 Kurokami, Kumamoto, Kumamoto 860-8555, Japan
E-mail: †{abec,kariya}@dbms.cs.kumamoto-u.ac.jp, ††{itokawa,kitasuka,aritsugi}@cs.kumamoto-u.ac.jp

Abstract　There have been many music and movie data which do not have the consent of the copyright owners and are thus available illegally, in WWW. One way to find such data is to analyse data in terms of their contents, it tends to be costly, though. In this paper, we propose a method to infer whether a given webpage provides such data illegally. In WWW, there are many webpages providing data legally. Our method extracts features based on nouns and anchor texts from them and attempts to make use of the features in the inference. Our experimental results show that our method can detect some illegal contents.

Key words　Illegal music site, Copyright

1. はじめに

近年，インターネットのブロードバンド化や携帯電話の 3G 化により，音楽や映像などの大容量のデジタルデータを高速で送受信できるようになり，パーソナルコンピュータや iPod などの携帯メディアプレイヤー，携帯電話などで楽しむことが可能となった．そのようなデジタルデータを配信や販売するサービス，例えば iTunes Store や着うたなどの利用者とダウンロード数が年々増えている [1]．その約 9 割が携帯電話向けの着うたや着うたフルなどの配信によるものである．しかし，着うたや着うたフルなどを著作権所有者に承諾を得ずに配信している，いわゆる違法な音楽配信サイトが増えており，音楽ファイルをダウンロードしているユーザも多くなっている．また，日本レコード協会の調査によると 2007 年 10 月から 2008 年 9 月の間，合法的に配信しているサイトの音楽ファイルのダウンロード数が 3 億 2900 万回であるのに対し，違法なサイトからのダウンロード数が約 4 億回以上という推定がある [2].

2009 年 6 月に成立し，2010 年 1 月 1 日に施行された著作権法の一部を改正する法律では，著作権法第 30 条が改訂された．改訂以前は音楽や映像などのデジタルデータの著作権所有者に無断で配信することは違法であったが，そのようなデータのダウンロードは私的使用の範囲内で合法とされていた．改訂され

た著作権法では配信されているデータが違法だと知りつつダウンロードした場合，著作権所有者の権利を侵害したとし違法となる．また，違法配信を行っているサイトの運営者の収入となる広告にも問題があり，バナー広告やリンク広告などの多くがアダルト系や出会い系の広告である．そのため，青少年が詐欺や恐喝などの被害にあうケースがある．これらの問題に対し，クローラーでウェブ上を巡回し，ウェブページを収集し，解析を行うことで違法配信を行っているコンテンツを発見し，ISP などに連絡しコンテンツを削除することで，一般ユーザが未然に法に触れないようにすることや，そもそも法改訂以前から違法であったアップロードの行為を抑えることなどを目的とし，本研究ではクローラーが得ることができるウェブページの情報を用いて違法配信を行っているコンテンツの推定を行う．

P2P などファイル共有を用いた違法なファイル交換も行われているが，P2P で共有されているファイルは P2P ユーザでなければ削除することができず，削除したとしても，P2P の特性上，一旦共有されてしまったファイルの拡散を防ぐことが困難なため，今回はウェブページで配信されている音楽と映像を対象とし，そのようなデータを配信しているコンテンツの推定を行う．

データの不正複製を検出するには電子透かしを用いたり，デジタルデータ自体を解析する手法もあるが，解析時間が長いなどコストが高いため，違法配信を行っているコンテンツを推定することで解析対象を限定する手段として本手法を用いるといった利用を想定する．

本研究では，ウェブで公開されているコンテンツを対象として，そのコンテンツが違法かそうでないかの推定を行う．違法に配信を行っているコンテンツの推定に，著作権所有者に許諾を得てデータを配信しているサイトから抽出した特徴量を用いる．抽出した特徴量を用いて機械学習を行い，映像や音楽を配信しているコンテンツを推定する．この推定結果には違法と合法の両方のコンテンツが含まれることになるが，合法的に配信を行っているコンテンツは権利者団体などが所有しているリストを用いることで除外できると考える．

本稿の構成は以下の通りになっている．2 章で著作権侵害に関して対策を行うための関連技術について，3 章で提案する手法について述べる．4 章で実験手順を述べ，5 章で実験の結果と考察を述べ，6 章でまとめとする．

2. 関 連 技 術

本章では音楽や映像などのデジタルデータの不正複製や再配布の防止のためにデータに情報を埋め込む技術に関しての説明を行う．本研究で提案する推定手法を用いることで，埋め込まれた情報を解析する対象を限定することができ，本章で述べる技術にかかるコストを削減できると考えられる．

2.1 電子透かし

電子透かしとは，デジタルデータに情報を埋め込む技術のことである．電子透かしをデジタルデータの著作権保護のために用いることは以前から研究されており [3]，埋め込む情報にはコピー可能数や著作権所有者などの著作権関連の情報が多い．電子透かしを検出することができるソフトを使用し，著作権の情報を取り出すことで不正にコピーされたものであるかどうかを検査し，著作権を侵害しているか判断することができる．しかし，電子透かし自体が著作権侵害を防止できるわけではなく，また，加工することによって電子透かしの情報が失われることもある．デジタルデータとして配布されたものにしか付与できないため，音楽 CD から取り込まれたデータには付与されていない．

2.2 電 子 指 紋

音楽 CD などの事前に電子透かしを付与していないデジタルデータに対しては電子指紋が有効だと考えられる．電子指紋とは，個々のデータ特有の特徴を抽出しパターン化し，登録しておくことで，同一のデータかどうか検出できる技術である．電子指紋は符号化方式が変わっても検出することが可能であり，また，検出対象となるデータが改変されていても検出することが可能である．大西ら [4] はサーバとクライアントが一体となってデータに電子指紋を埋め込む手法を提案し，実際に放送システムを構築し，評価を行い実用性を示した．

上記二つの技術をデータに用いることで，配信されているデータが違法であるか確認することができる．しかし，大量に配信され，日々増加していくデータをクローラーで逐一解析するには解析コストがかかる．

3. 提 案 手 法

3.1 判別システム

本稿では，違法に配信を行っているコンテンツの推定に，著作権所有者に許諾を得てデータを配信しているサイトから抽出した特徴量を用いる．抽出した特徴量を用いて SVM を用いて機械学習を行い，映像や音楽を配信しているコンテンツを推定する．音楽配信を行っているコンテンツを推定するため，結果として，違法，合法かどうかに関わらず両方のコンテンツを推定してしまうが，合法的に配信を行っているコンテンツは権利者団体などが所有している URL や許可している認識番号などのリストがあり，そのリストを用いることで合法な配信を行っているサイトのコンテンツを除外できると考える．

判別システムは，学習データから判別器を生成する学習部と判別を行う判別部で構成される．図 1 は判別システムの概要である．学習部では，学習データから 3.2 節で説明する特徴量を抽出し，機械学習を行い判別器を生成する．判別部では，違法配信であるかどうかを推定したいページを入力とし，学習部で用いられたものと同様の手法を用いて特徴量を抽出し，その特徴量を用いて，学習部で判別器を生成する．生成された判別器を用いて違法配信コンテンツの推定を行う．

3.2 判別に用いる特徴量

本研究で提案する手法で用いる特徴量は学習データ d_i $(i = 1, 2, \ldots, m)$ における単語の出現頻度や共起頻度を用いる．本章では機械学習に用いる特徴量について述べる．

手法 1:名詞の出現頻度

MeCab [5] に学習データ d_i を読み込ませ，出力結果

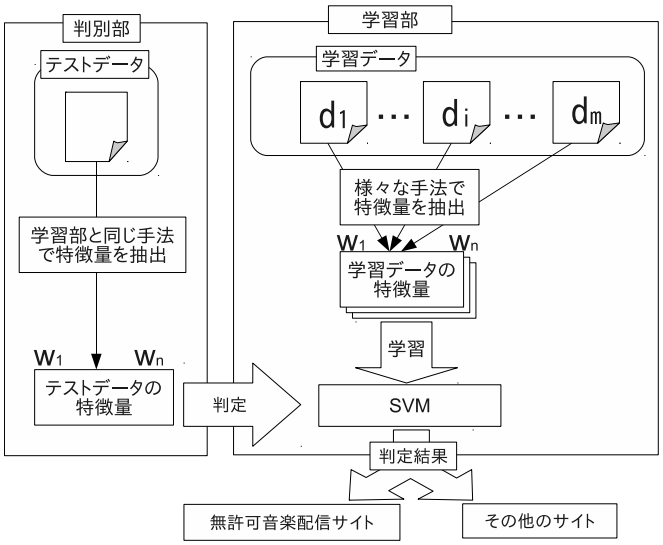

図 1　判別システム概要

から名詞を抽出する．抽出した名詞をページの先頭から順に $w_{d(i,1)}, w_{d(i,2)}, \ldots$ とし，この並びを $W_{d_i} = (w_{d(i,1)}, w_{d(i,2)}, \ldots)$ とする． W_{d_i} を作成するとともにすべてのページに出現した名詞を出現頻度順に並べた $W = (w_1, w_2, w_3, \ldots)$ を作成する．すなわち w_i は学習データすべての出現回数の和が i 番目に多い名詞である．後述する定義の説明のため，頻度順位が n までの頻出名詞群を $W_n = (w_1, w_2, \ldots, w_n)$ とする． W_n の各成分が，学習データのページそれぞれの名詞並びである W_{d_i} に出現した回数を特徴量とする．

ウェブページ内での名詞の出現頻度を調べることで，そのページの特徴をとらえることが多く，本研究では音楽配信サイトの名詞の出現頻度を用いることで，違法配信コンテンツを推定できるのではないかと考え，この手法を用いる．

手法 2:名詞の共起頻度

名詞の共起頻度を用いる判別手法では，次の二つの研究を参考にして考慮する．ユーザのコマンドの共起に注目した侵入検知の研究で，岡ら [7] は ECM(Eigen Co-occurence Matrix) を用いることを提案している．また，ECM を使った処理を効率的に行った研究 [8] がある．本稿では，これを，学習データから抽出した名詞に適用する．名詞間の共起の情報を行列として作成し，それを使った判別を考える．

具体的には，名詞の 2 項間の共起頻度を特徴量とした判別器を生成する．w_p，w_q は W_n の任意の異なる 2 成分とし，d_i について w_p における w_q の共起頻度を次のように調べる．学習データ d_i について，w_p における w_q の共起頻度を次のように定義する．学習データ d_i の名詞の並び W_{d_i} において w_p が出現する箇所を先頭に，ウィンドウ幅 l 個の名詞の並び $(w_{d(i,p')}, w_{d(i,p'+1)}, \ldots, w_{d(i,p'+l-1)})$ を取り出す．このとき $w_{d(i,p')} = w_p$ である．取り出した名詞の並び中に w_q が含まれる個数を求め，この個数を学習データ d_i についての w_p における w_q の共起頻度とする．ただし，学習データ d_i 中に w_p が複数回出現する場合は，出現のたびにウィンドウ幅個の名詞の並びを取り出し，複数の w_p のウィンドウが重なる場合も取り出す名詞の並びは重複を許し，重複箇所の w_q は複数回出現したとみなす．すべての学習データで頻出名詞群 W_n 内の単語同士の組合せで共起頻度を計算し，それをもとに共起行列 C_i を作成する．学習データ d_i についての共起行列 C_i は W_n をもとに作成するため，$n \times n$ の行列で，p 行 q 列の要素は，学習データ d_i での w_p における w_q の共起頻度とする．

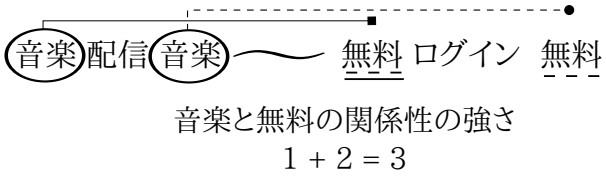

図 2　音楽と無料の共起表現

例として，ある学習データのページで名詞の並びを作成した結果，図 2 のような名詞の並びが得られたとする．名詞から延びている線はウィンドウ幅を示している．共起行列の項目となる頻出名詞群 W_n を “音楽”，“無料” とし，“音楽” における “無料” の共起頻度を求める．図 2 では 1 回目に出現した “音楽” から延びている実線内に “無料” が 1 回出現し，2 回目に出現した “音楽” から延びている点線内に “無料” が 2 回出現している．この例では 1 回目で数え上げられた 1 と 2 回目で数え上げられた 2 との合計，つまり $1+2=3$ が “音楽” における “無料” の共起頻度となる．用意した全ての学習データを共起行列の項目の名詞ごとに検査していき，共起頻度を算出して，得られた共起行列を特徴量とする．

手法 1 と同様に単語に着目したが，単語の出現している順序に特徴があるのではないかと考えてこの手法を用いる．

手法 3:アンカーテキストの出現頻度

学習データ d_i でリンクを貼られた文字，つまりアンカーテキストの出現頻度を特徴量として用いる．ページの先頭から順に $l_{d(i,1)}, l_{d(i,2)}, \ldots$ とし，この並びを $L_{d_i} = (l_{d(i,1)}, l_{d(i,2)}, \ldots)$ とする．L_{d_i} を作成するとともにすべてのページに出現したアンカーテキストを出現頻度順に並べた $L = (l_1, l_2, l_3, \ldots)$ を作成する．すなわち l_i は学習データすべての出現回数の和が i 番目に多いアンカーテキストである．頻度順位が j までの頻出アンカーテキスト群を $L_j = (l_1, l_2, \ldots, l_j)$ とする．L_j の成分が，学習データのページそれぞれのアンカーテキストの並びである L_{d_i} に出現した回数を特徴量とする．

手法 1 と同様であるが，アンカーテキストに着目することでコンテンツ作成者がどのような単語を用いて利用者の興味をひこうとしているか調べ，それを特徴語とし，その頻度に特徴があるのではないかと考えたためこの手法を用いる．

手法 4:アンカーテキストの共起頻度

アンカーテキストの共起頻度では，手法 2 で説明した名詞の共起頻度をアンカーテキストに用いて特徴量を得る．手法 2 ではページ内の単語を抽出し，単語集合 w_n から共起頻度行列を得たが，ページの先頭から順にアンカーテキストを得て，$l_{d(i,1)}, l_{d(i,2)}, \ldots$ とし，この並びを手法 2 と同様に共起頻度を計算し，共起頻度行列 C_i を得て，特徴量とする．

手法 2 と同様の理由からこの手法を用いる．

手法 5:名詞とアンカーテキストの出現頻度

学習データ d_i から名詞の出現頻度とアンカーテキストの出現頻度それぞれを併合し，出現頻度順に並べたものを特徴量とする．

手法 1 と手法 3 は個別に出現頻度を考えていたが，同一なものとして考えることで詳細に特徴をとらえられるのではないかと考えて，この手法を用いる．

手法 6:名詞の出現頻度とアンカーテキストの出現頻度

学習データ d_i から W_n と L_j を抽出し，組み合わせたものを特徴量とする．この時，特徴量は W_n から n，L_j から j 抽出するため特徴量の長さは $n+j$ となる．

手法 5 では二つの出現頻度を用いているが，どちらかの特徴が乏しくなるのではないかと考え，用いる特徴量の件数を固定した．

手法 7:名詞がアーティスト名の場合，「アーティスト名」に置換した名詞の出現頻度

学習データ d_i から名詞の出現頻度を抽出する際に，名詞がアーティスト名だった場合，その名詞を「アーティスト名」という単語に置換することによる，人名に重み付けを行った特徴量を用いる．これは，音楽配信サイトでは必ずといってよいほど，アーティスト名が記載されているため，そこに着目する．アーティスト名のリストを作成しておき，それを用いる．

手法 8:アンカーテキストの単語がアーティスト名の場合，「アーティスト名」に置換したアンカーテキストの出現頻度

学習データ d_i からアンカーテキストの出現頻度を抽出する際に，上記同様アンカーテキストがアーティスト名だった場合，そのアンカーテキストを「アーティスト名」という単語に置換し，特徴量を抽出する．

これは，音楽配信サイトで実際にデータをダウンロードする際にアンカーテキストにアーティスト名が記載されていることが多いため，それを考慮することで違法配信コンテンツを推定できるのではないかと考え，この手法を用いる．

手法 9:アーティスト名に重みを付けた名詞の出現頻度

学習データ d_i から名詞の出現頻度を抽出し，名詞がアーティスト名だった場合，そのアーティスト名の出現頻度を重み付けしたものを特徴量とする．重み付けをすることで，アーティスト名を重視できるのではないかと考えた．

手法 7 と同様の理由からこの手法を用いる．

手法 10:アーティスト名に重みを付けたアンカーテキストの出現頻度

学習データ d_i からアンカーテキストの出現頻度を抽出し，手法 9 同様アンカーテキストがアーティスト名だった場合，そのアーティスト名の出現頻度に重みを付けたものを特徴量とする．

手法 8 と同様の理由からこの手法を用いる．

4. 実 験 手 順

前章で述べた提案手法を用いて，学習データから生成した判別器がどの程度違法配信コンテンツであるかを推定できるか検証を行なう．

図 3　違法配信コンテンツの例 1

判別器を作成するための学習データとして社団法人日本音楽著作権協会 (JASRAC) [9] や社団法人日本レコード協会 (RIAJ) [10] などで承諾を得ていると確認できたサイト，つまり合法的に音楽配信を行っているサイトを用いた．また，今回提案した手法の評価のためのテストデータとして手動で違法配信コンテンツを収集した．違法配信しているコンテンツの定義として，著作権所有者に承諾を得ていない著作物を無断で公開し，ユーザが容易にその音楽ファイルなどのデータを取得できるウェブページとした．容易に取得できるということが問題であると考え，拡張子などを変更したり，データを分割することにより，偽装して配信を行っているコンテンツに関しては今回は対象外とした．

音楽配信をしているサイトが検索結果に出るキーワード，“音楽配信”，“無料” などのキーワードで検索を行い，検索結果のページを目視で確認し，そこで公開されている音楽が著作権所有者や著作権所有者から著作権の信託を受けた団体から承諾を得ているかを JASRAC が公開している J-WID [11] と呼ばれる作品データベース検索サービスを用いて確認した．承諾を得ていなければ，違法配信を行っているコンテンツとした．承諾を得ているかどうかの判断には，権利者団体が所有しているリストを用いることができないので，JASRAC の承諾マークや RIAJ の承諾マークであるエルマークがあるか否かなどを参考にした．違法配信と確認したコンテンツにあったリンク先も違法配信コンテンツではないかと考え，実際にアクセスし，違法配信だと確認できたらテストデータセットとして用いた．図 3，図 4 は違法配信を行っていると判断したの例である．図 3 は著作者に承諾を得ずにアップロードされたと考えられる動画を記事として投稿しているコンテンツである．図 4 も著作権所有者に承諾を得ずにアップロードされた音楽データを自由にダウンロードできるようになっているコンテンツである．

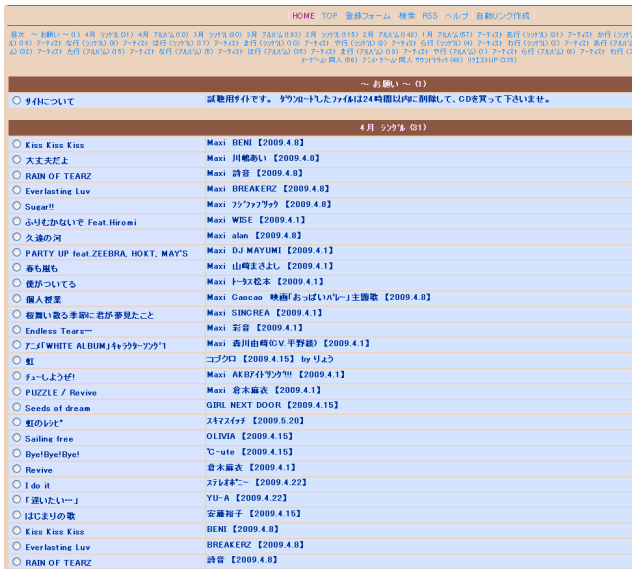

図 4　違法配信コンテンツの例 2

また，今回判別する音楽配信ではないサイトのコンテンツも収集し，それを判別できるかを確認する．こちらのコンテンツの具体例として，大学，企業などのページや本研究で提案した手法で音楽配信だと推定されそうなアーティストのブログやニュース記事なども収集し，どの特徴量がどのようなコンテンツに効果があるかを確認する．

本研究で用いたデータセットを表 1 に示す．

表 1　本研究で用いたデータセット

データ種別		件数
学習データ	許可	581
テストデータ	無許可	41
	その他	54

学習データから特徴量を抽出し，音楽配信コンテンツかどうか判別するため機械学習を用いて判別器を生成する．本研究では LIBSVM [6] を用いて One-Class の SVM で学習を行う．

本稿では 3.2 節で説明した，頻出単語 50 位までを特徴とし，共起頻度で用いるウィンドウ幅を 1000 とした．手法 9，手法 10 で用いる重みは 10 を用いた．

5. 実験結果

音楽配信サイトを学習データとし特徴量を抽出し，違法配信を行っているコンテンツの推定を行った結果を表 2，表 3 に示す．違法配信を行っているコンテンツを正しく判別できた件数を TP(True Positive)，その他のコンテンツを正しく判別できた件数を TN(True Negative) とする．

今回，実験で用いる手法は 3.2 節で説明したものを用いる．以下，ページに出現した名詞に基づいた特徴量とアンカーテキストにカテゴリを分け，カテゴリ毎に考察を行う．

5.1　名詞の抽出を行った手法

表 2 は名詞の抽出を行い，それを用いた手法とその結果である．名詞の抽出を行った手法は，手法 1，手法 2，手法 5，手法 6，手法 7，手法 9 である．

表 2　名詞の抽出を行った手法

手法	TP	TN
手法 1	4	48
手法 2	6	26
手法 5	5	47
手法 6	10	31
手法 7	4	46
手法 9	4	48

TP の件数を確認すると，件数は少ないが，同じような結果となった．判定されたコンテンツを確認したところ，どの手法でも同様のコンテンツを判定していた．つまり，判定精度はまだ低いが，我々の提案によって，適切に学習が行われ，期待通りの判別もできることが確認できた．以下，精度を向上するために，どのような判定が行われているか手法毎に確認を行う．

判定を行ったコンテンツを確認すると，アーティスト名を「アーティスト名」と置換し，名詞の出現頻度を特徴量として用いた手法 7 は，名詞の出現頻度を用いた手法 1 に包含されていた．これは，本実験では頻出語の上位 50 位を特徴量の項目として用いたが，手法 1 ではアーティスト名が項目として出現せず，手法 7 では手法 1 の 1 つの項目の代わりに「アーティスト名」となり，判定結果に差がでなかったと考えられる．

次に名詞の出現頻度を用いる手法 1 と名詞とアンカーテキストの出現頻度を用いる手法 5 だが，今回の実験で抽出された名詞の総数は 176,901 語，抽出されたアンカーテキストは 36,151 語と差があり，出現頻度上位 50 件内ではほとんどの項目が抽出された名詞であったため，差の無い結果になったと考えられる．

名詞の抽出を行った手法全体としては，更新頻度の低いウェブページや，更新を止めているページ，古い音楽がメインであるページなどは全く判別できていない．それは，学習させた多くのページが合法な音楽配信サイトの最新の楽曲やアーティストが書かれたページであったからだと考えられる．特に，アーティスト名に重みを付けた名詞の出現頻度を用いた手法 9 では他の手法に比べ，特徴量の項目にアーティスト名が多く，違法配信を行っているコンテンツでも，そのアーティストが掲載されていなければ違法配信と判定されず，その他のコンテンツであると誤判定されていた．その他のコンテンツを正しく判定するかに関しては，今回の実験では判別に One-Class の SVM を用い，違法配信と判定されなかったものは全てその他とするため，誤検出が少なかった．しかし，その他のページとして，アーティストのブログやアーティストの公式サイト，アーティストに関係するニュース記事などを含めていたが，そのようなページを違法配信コンテンツとして誤検出していた．

5.2　アンカーテキストの抽出を行った手法

次にアンカーテキストの抽出を行った手法である．表 3 はアンカーテキストの抽出を行い，それを用いた手法とその結果である．アンカーテキストの抽出を行った手法は，手法 3，手法 4，手法 5，手法 6，手法 8，手法 10 である．

TP となったコンテンツを確認したところ，名詞の抽出を行っ

表 3　アンカーテキストの抽出を行った手法

手法	TP	TN
手法 3	23	18
手法 4	40	3
手法 5	5	47
手法 6	10	31
手法 8	15	18
手法 10	31	7

た手法の結果と同じく，同様のコンテンツを判定していた．しかし，TP の件数には各手法にばらつきがあり，名詞を特徴語とした手法と違った特徴があると考えられる．以下，アンカーテキストの抽出を行った手法毎に確認した考察を述べる．

アンカーテキストの共起頻度を用いる手法 4 は，無許可配信を無許可配信と判定した数が 41 個中 40 個と全手法の中で一番高い．これは，共起行列を用いて判別器を生成したが，実際に作成された共起行列を確認したところ，名詞に比べ，アンカーテキストの共起は多くなかったため，共起行列の要素がほぼ 0 であった．SVM はそれを特徴としてとらえ，共起が起きなければ違法配信コンテンツであると判断したためだと考えられる．その他のコンテンツもほとんど無許可配信と判定している．

また，手法 10 もその他のページを多く誤判定している．これは学習データの中にアンカーテキストが少なく特徴がとらえられず，その他のページもアンカーテキストが少ないものが多かったため，その他のコンテンツを違法配信コンテンツだと誤判定が起こったと考えられる．

5.1 節で手法 7 が手法 1 に包含されていると述べたが，同様にアンカーテキストのアーティスト名を「アーティスト名」に置換した時の出現頻度を用いる手法 8 は，単にアンカーテキストの出現頻度を用いた手法 3 に包含されている．これも名詞の時と同様にアーティスト名という項目に変わったというだけで差がでなかったと考えられる．

アンカーテキストの抽出を行った手法全体としては，アンカーテキストが多いウェブページは無許可配信と判定され，少ないウェブページはその他のページと判定される傾向にある．それは学習データの中にアンカーテキストの多いものから少ないものまであり，特徴をとりづらいからだと考えられる．また，アンカーテキストの抽出を行った手法全体では，その他のページとラベル付けされたページの誤検出が多かった．

名詞を抽出した手法と同様に，アーティストのブログや公式サイト，ニュース記事などが違法配信だと誤判定された．さらに，企業のトップページや大学のページなども数ページ誤判定された．これは，学習データとして用いた合法な音楽配信サイトでよく出現する “ログイン” や “ヘルプ”，“サイトマップ” など一般的な単語がアンカーテキストとされていたためだと考えられる．

6. おわりに

本研究では，音楽配信サイトから特徴量を抽出し，著作権所有者に無断で音楽を配信しているいわゆる違法配信コンテンツの推定を行った．合法的に音楽配信を行っているサイトのリストなどを権利者団体は持っているであろうと考え，合法，違法問わず音楽配信を行っているコンテンツの推定を行い，合法的に音楽配信を行っているサイトをリストを用いることで除外し，違法配信コンテンツを推定できるという有用性を示したことになると考えられる．権利者団体がリストを作成，所有しているため，学習データとしての信頼性があると考えられる．また，学習器を生成・更新する際に，リストを用いることで学習データを容易に取得することが可能となる．配信されているデジタルデータが違法かどうか推定するにはデータ自体を解析する手法もあるが，そのデータが配信されているページを違法配信しているかどうか推定することで解析対象を限定できる．そのような前段階の処理手法として，本研究の手法が有効であると考えられる．

本研究では，音楽配信とそれ以外の二つに分類するため，誤判定が起こることは避けられない．そのため，今後は合法か違法か判断できないサイトをグレーゾーンのようなものとし，Multi-Class の SVM で学習を行うことで違法配信ページが発見できるようにしたい．また，今回用いた学習データを取得する期間が短く，最近の音楽に関するページを多く利用したため，名詞の抽出を行った手法は，更新頻度の低いウェブページや，更新を止めているページ，古い音楽がメインであるページなどは全く判別できていない．そこで，学習データの収集を長い期間で行うことで，より正しい判定ができるのではないかと考えられる．

文　献

[1] 社団法人 日本レコード協会，2008 年度 音楽メディアユーザー実態調査，http://www.riaj.or.jp/report/mediauser/pdf/softuser2008.pdf

[2] 違法な携帯電話向け音楽配信に関するユーザー利用実態調査 2008 年版，http://www.riaj.or.jp/report/mobile/2008.html

[3] 小川 宏，中村 高雄，高嶋洋一，“電子透かしを用いたデジタル動画像の著作権保護方式，” 情報処理学会全国大会講演論文集，Vol.55，No.3，pp.248–249, 1997.

[4] 大西 宏樹，上原 哲太郎，佐藤 敬，山岡 克式，“ 電子指紋により不正複を抑止するインターネット放送システム，” 情報処理学会研究報告，Vol.2006，No.26，pp.49–54, 2006.

[5] MeCab，http://mecab.sourceforge.net/

[6] LIBSVM，http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[7] 岡瑞起，小磯知之，加藤和彦，“Eigen Co-occurrence Matrix (ECM): 時系列データからの多層ネットワーク特徴抽出手法の提案，” 日本データベース学会 Letters， Vol.3， No.2，pp.9–12, 2004.

[8] Chen, L. and Aritsugi, M. “ An SVM-Based Masquerade Detection Method with Online Update Using Co-occurrence Matrix, ” Proc. Third International Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA), Lecture Notes in Computer Science, Vol.4064, Springer, pp.37–53, 2006.

[9] 社団法人 日本音楽著作権協会， http://www.jasrac.or.jp/

[10] 社団法人 日本レコード協会，http://www.riaj.or.jp/

[11] 作品データベース検索サービス ，http://www2.jasrac.or.jp/eJwid/