キーワード出現特性を用いたコンテンツ不正利用探索方式の検討

5D－2

大森 信行　　森 大二郎　　稲垣 博人
{ohmori, mori, inagaki}@aether.hil.ntt.co.jp
NTT サイバーソリューション研究所

1　はじめに

近年、コンピュータやインターネットの普及に伴ってネットワークに接続されたPCから世界中のコンテンツを利用することができるようになった。一方、ディジタルコンテンツが容易に入手・編集・配信できるため、著作物であるコンテンツの不正利用が増えてきた。そのため、ディジタル・コンテンツの著作権保護を目的として、コンテンツに電子透かしなどを利用し著作権情報を管理・監視する技術が研究されてきた。例えば、電子透かしを用いて、購入者情報などをコンテンツに埋め込んでおき、不正と思われるデータから付加されている情報を読み出して不正利用であるかどうかを判断することによって、不正利用を抑止するシステム [1] などが提案されている。しかし、コンテンツが不正に利用されているかどうかを判断するためには、不正と思われるコンテンツを見つけ、入手することが必要である。

ここで提案する探索方式は、インターネットのすべてのコンテンツを収集、チェックするというアプローチではなく、探索パターンに一致する特定のコンテンツを収集しチェックする手法である。コンテンツプロバイダやコンテンツ管理者が不正利用を探索し発見する技術を持つことで、コンテンツの不正利用者を突き止めることが可能となる。

2　キーワード出現特性を用いた不正利用探索

インターネット上のすべてのホームページを収集し、そこに掲載されているコンテンツもすべてを収集し、電子透かしのチェックを行えば不正利用コンテンツの収集漏れはなくなる。しかし、探索対象の収集や透かしのチェックに大きな計算時間が必要とされるため、インターネット上のすべてのコンテンツを収集し、透かしのチェックを短時間ではできない。そこで、探索対象を絞り、不正利用コンテンツの効率的な探索を可能とするために、探索対象に関連するキーワードに基づき、探索対象を収集する。

Detecting illicit copies of digital contents using characteristic of words appearance
Nobuyuki OHMORI, Daijiro MORI, and Hirohito INAGAKI
NTT Cyber Solutions labs.

2.1　探索範囲の決定

探索範囲を絞り込む方法としては、1.URLを指定する方法と、2.あるキーワードを含むページのURLを見つける方法が考えられる。

1.については、過去に不正利用が行われたページや画像などのコンテンツを多く掲載しているページなどのURLをユーザが指定し、そのURLの付近のコンテンツを調査する方法である。

2.については、ユーザがキーワードを指定し、そのキーワードを含むページを探索シードページとする。探索シードページに掲載されているコンテンツを収集し、電子透かしのチェックを行う。この場合は、不正利用を探索しようとするユーザは、対象となるコンテンツに関するキーワードを設定すればよく、URLまでを設定する必要はない。

探索シードページは以下の手順で決定する。探索シードページの決定を高速に行うため、あらかじめインターネット上のページを収集し、そこから単語を抽出して、単語インデックスと呼ばれる索引ファイルを作成しておく。ここで、日本語の文を単語に分割する処理が形態素解析処理である。形態素解析には、InfoBee[2] を用いる。単語インデックスは、単語とその出現ページのURLを組にした表である。ここから、ユーザが入力したキーワードとインデックス中の単語との照合を行い、その語を含むURLのリストを探索シードページとして得る。

得られた探索シードページは、サーチエンジンなどで一般的な *tf·idf* 法 [4] により得点がつけられている。得点によって探索シードページを順位付けすることができ、ユーザの入力したキーワードに関連するページほど、得点が大きくなる。ページの得点は、ユーザが指定したキーワードのページ全体での出現頻度と、各ページでの出現頻度から計算する。基本的にはユーザの指定したキーワードを多く含むものほど大きな値になる。ただし、どのようなページにも出現するようなキーワードについては、得点が低くなる。

2.2　コンテンツの収集

ユーザが入力したキーワードに応じて決定された探索シードページを基に、以下のような手順で電子透かしのチェック対象である検索対象コンテンツを収集する。

1. 探索シードページを収集する。
2. 収集したページ内のリンクを解析し、そのページからリンクされているページのURLと、そのページに掲載されている探索対象コンテンツのURLを得る。
3. 探索対象ページを収集し、2. の処理を繰り返す。
4. 2. の探索対象コンテンツを収集する。

2.3 電子透かしのチェック

電子透かしのチェックでは、コンテンツの種類、つまり画像、音声などに応じた電子透かしチェックモジュールを利用する。

3 不正利用探索システム

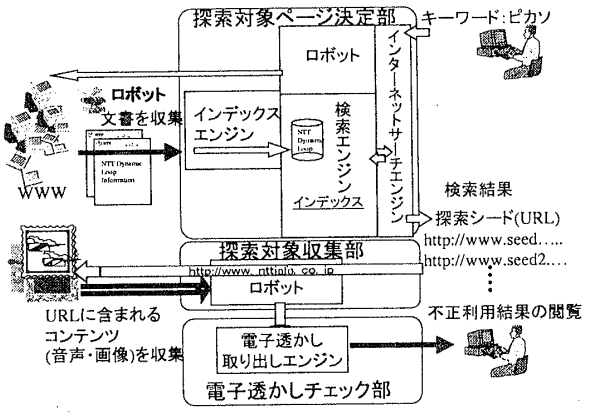

図 1: 不正利用探索システム

図1に試作したシステムの概要を示す。本システムは、3つのサブシステムから構成される。

探索ページ決定部 入力されたキーワードを含むホームページのURLをインデックスと照合して出力する。出力されたURLが探索シードページである。インデックスは、どの単語がどのURLに含まれているかが保存されている。あらかじめ収集されたホームページから単語を抽出し、インデックスを作成しておく。探索対象ページ決定部におけるインデックス作成とキーワードにより探索シードページを決定するシステムには森らのシステム [3] を用いた。

探索対象収集部 探索ページ決定部で出力されたURLのホームページに含まれる、コンテンツを収集する。

電子透かしチェック部 探索対象収集部で収集したコンテンツの透かしチェックを行う。不正なコンテンツから、透かし情報が発見された場合は不正利用と判定し、結果を出力する。ここでは、静止画像用のモジュールとして中村らの電子透かし技術 [5] を用いた。

4 評価実験

本システムにより、ユーザが指定したキーワードを入力し、日本国内のページに対して不正利用探索の実験を行った。今回の実験では、画像コンテンツ (gif, jpeg) を探索対象とした。

探索ページ決定部の入力は、あるコンテンツを示す3つ程度のキーワードである。このキーワードを含むURLを探索シードページとし、探索シードページとして約5000件のURLが得られた。探索対象収集部では、探索シードページの近傍、3ホップカウント以内のコンテンツを収集した。

収集後に画像コンテンツに対して、電子透かしを読み出し、不正利用のチェックを行った。ネットワークやサーバの設定によっても異なるが、キーワードの入力から、透かしの検出が終了するまでに半日以内で処理を完了することが可能であることが分かった。

5 今後の課題

電子透かしの不正利用探索方法として、探索パターンに一致する特定のコンテンツを収集し、チェックを行う方式を提案した。今回の評価実験では、あるURLの近傍のコンテンツの収集から、電子透かしのチェックまで半日以内に処理できる見通しを得た。

今後は、さらに効率的かつ効果的な不正利用探索技術を検討する。特に、検索対象ページの類似文書検索による方法などをインプリメントし評価を行う。

参考文献

[1] 大友他：著作権を考慮した画像流通システム,'97 信学春全大,A-7-9,1997

[2] 井上他:InfoBee テキスト情報検索技術, NTT R&D No.10,pp.1103-1108,(Vol.46) 1997

[3] 森他: 分散型大規模文書検索システムにおける適合度計算方法について, 情報処理学会ディジタル・ドキュメント研究会,DD-15-2,1998

[4] 大森他:$tf \cdot idf$ 法を用いた関連マニュアル群のハイパーテキスト化, 情報処理学会自然言語処理研究会,NL-121-16,1998

[5] 中村他: 静止画像に対する電子透かし技術, NTT R&D No.6 ,pp.711-714,(Vol.47) 1998