WWW サイト内の不正コンテンツ検出支援システムの構築

日田仁† 泉裕†† 齋藤彰一‡ 上原哲太郎‡ 國枝義敏‡

†：和歌山大学大学院システム工学研究科 ††：和歌山大学システム情報学センター
‡：和歌山大学システム工学部

内容梗概

通信回線のブロードバンド化により、WWW による情報発信が今まで以上に盛んになってきている。しかし一方で、著作権を侵害するものや公序良俗に反する不正コンテンツの WWW を用いた配布という危険性が、より一層高まっている。そのような WWW による不正コンテンツの配布を抑止するため、WWW サイト内の不正コンテンツ検出を支援するシステムを構築した。不正コンテンツはヘッダ改竄等の様々な手段によって偽装された上で公開されるため、直接の検出は困難である。そこで、当該サイト内で用いられている用語を用いて不正コンテンツ配布に使われている可能性の高い WWW ページを選び出し、作業効率を上げるツールを作成した。

The implementation of a support system for detecting illegal contents in WWW sites

Hitoshi HIDA† Yutaka IZUMI††
Shoichi SAITO‡ Tetsutaro UEHARA‡ Yoshitoshi KUNIEDA‡

†Graduate School of Systems Engineering, Wakayama University
††Center for Information Science, Wakayama University
‡Faculty of Systems Engineering, Wakayama University

Abstract

With the recent growth of the broadband Internet, WWW has become quite common as a medium for information distribution by individuals. On the other hand, it might be used for distribution of illegal contents. To care with the anxiety, the authors have developed a supporting system to detect inadequate contents in WWW sites.

It is very difficult to detect illegal contents in WWW sites straightforwardly, since they are mostly camouflaged with malicious modification of the file header. Therefore, this paper describes the implementation of an efficient tool, which detects the WWW pages with suspicion of distributing illegal contents, via checking the "Underground" words.

1. はじめに

近年、インターネット利用者人口の爆発的な増加や、インターネット回線の急速なブロードバンド化により、その利用方法が多様化してきている。現在主流である利用方法は電子メールと WWW であるが、それらに加え、動画像や高音質の音楽等のマルチメディアコンテンツを、インターネットを使って配信するという利用方法も普及してきている。

しかし、このようなインターネットの広帯域化により容易になったファイル交換の手軽さが、悪用される事例も増えてきている。具体的には、著作者に無断でコピーされたソフトウェアなどを配布している WWW サイト（いわゆる Warez サイトなど）の存在などが挙げられる。近年、このようなサイトの抑止のため、インターネットサービスプロバイダ(ISP)に不正コンテンツの配布に対する責任を科する動きがあり[1]、WWW サイト管理者にとっては自サイト内の不正コンテンツの検出は大きな課題となっている。しかし、そのような不正コンテンツのファイルは、多くの場合偽装が施されており、ファイルを直接調べることによる不正チェックが困難である。

そこで本研究では、WWW サーバ管理者の不正コンテンツ検出作業を支援するツールの構築を目的とし、WWW サイトにおける不正コンテンツ検出の方法の考察と、それに基づくツールの実装を行った。

2. 背景と目的

インターネットのブロードバンド化により、インターネット上で交換可能なファイルのサイズの制限はごく緩やかになった。それに伴って、インターネットを用いたソフトウェアの不正コピーや、音楽、映像の不正コピーが非常に容易となり、大きな問題として取

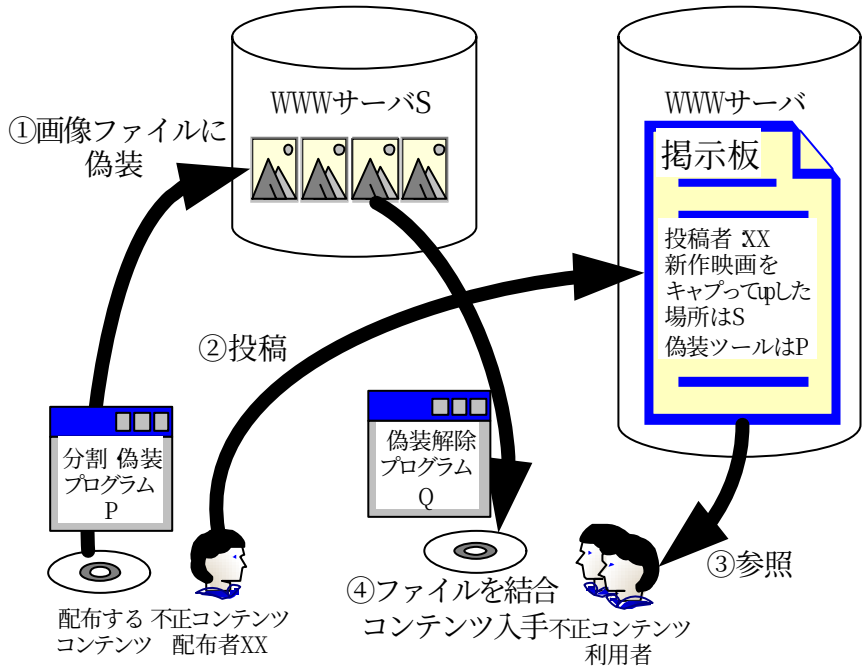

図１ WWW と掲示板を使った不正コンテンツ配布

り上げられている。

インターネット上での不正コンテンツの流通に使われる手法にはさまざまなものがあるが、特に多いのは Peer-to-peer ファイル交換システム(NapsterやGnutellaおよびそれらの互換システム)を用いた手法と、WWW を用いた手法である。

WWW を用いた不正コンテンツの典型的な配布手順を図 1 に示す。不法コンテンツの配布者は、配布したいコンテンツファイルを、ツールなどを用いて分割し、画像などほかのファイルとして偽装して WWW サーバ上に置く。そしてその置き場所と偽装手段を、電子掲示板などの上で公開する。多くの場合この電子掲示板のサーバはコンテンツの置き場所とは全く無関係の場所である。不正コンテンツの入手者は、掲示板の内容からコンテンツの置き場所と偽装解除方法を知り、そのコンテンツを得る。また、この掲示板を用いて、欲しいコンテンツの配布をリクエストしたりすることも多い。このような不正なコンテンツの配布に使われる WWW サーバやページなどを称して、UG 系(Under Ground)サイトと呼ぶ。

UG 系サイトがサーバに存在した場合、そのサーバ管理者には著作権侵害に対する管理者責任を問われる可能性がある。特に大学のように、多数の構成員を持ち、管理者の監視が難しい WWW ページが多数存在しうるような環境ではその危険は大きい。しかし大学側にそのようなトラブルに対する危機管理が行き届いているとは言いがたく、多額の損害賠償が請求されるような事態への対処についてはほとんど考えられていないのが現状である。

以上のような背景から本研究では、特に大学の管理下にある WWW サーバ、特に情報処理センターなどのように完全にサーバの自身を管理できるようなサイトにおいて、そのサーバ内の UG 系サイトの検出を支援するシステムの構築を目的とする。

3. システムのモデル

一般に UG 系サイトでは、配布しようとする不正コンテンツを含むファイルに対して偽装を施す。偽装の手段としては、

- 拡張子の変更
- ファイルヘッダの改竄
- 特殊な形式での圧縮および分割
- ファイルフォーマットの拡張仕様を用いた埋め込み
- (パスワードつきの)暗号化

などが挙げられる。またこれらの偽装手段を組み合わせ、ファイルに対する偽装解除を困難にしている場合もある。さらに、このような偽装工作を簡略化するために作られたツール類も存在している。

したがって、不正コンテンツを検出する際に、公開されているファイルを直接調べても、有効な情報が得られない場合がある。特に暗号化が施されたファイルに対しては、偽装前のファイルを推測することはほぼ不可能である。

そこで本研究において、不正コンテンツの検出支援の方法として以下のようなものを検討し、組み合わせて利用することにした。

- WWW サイト内の用語チェック
- 公開ファイルへのアクセス頻度調査
- 既知の偽装パターンを利用したファイル形式のチェック

露無　（ROM の当て字）
割れ物　（不法コピーしたソフトウェア）
塩　（フリーWWW サーバ geocities）
でんこ　（ファイル偽装ツールの愛称）

図２ UG 語の例

以下、それぞれの手法について述べる。

4. 1 WWW サイト内の用語チェック

UG 系サイトで偽装されたコンテンツファイルを公開する場合の典型的な手法は、そのコンテンツファイルの場所（URL など）と、ファイルの偽装形式や、偽装解除の方法として偽装に使ったツールの名前などを掲載したページを使うものである。

このような UG 系サイトでは、UG 系独特の隠語が頻繁に用いられる。海外に存在するフリーホームページのサーバの名前や、偽装に用いられるツールの名前、あるいはインターネット用語を漢字を使って当て字で表現した隠語などがこれに当たる（図２）。

本研究では、これらの用語を UG 語とし、WWW サイト内で使用されている UG 語を自動的に検出することによって、不正コンテンツを配布している WWW サイトの検出支援を行うツールを構築する。

4. 2 アクセス頻度調査

UG 系サイトにおいて公開される不正コンテンツのファイルには、短期間にアクセスが集中する傾向がある。そこで、ある一定期間におけるファイルへのアクセス頻度を集計し、そのアクセス頻度の高いファイルをサーバ管理者に知らせることにより、不正コンテンツファイルのチェックを支援するツールも構築する。

4. 3 偽装ファイルのパターンチェック

偽装ファイルの直接の検出のために以下の手段を用いる。まず、WWW サイト内で公開されている全ファイルのファイル名やファイルサイズを手がかりに不正コンテンツの検出を行うツールを構築する。不正コンテンツの偽装ツールによっては、ファイルのサイズにある法則が見られるものがある。たとえば、まず、ファイルを一定の大きさに小分割し、test001,test002…といった連番のついたファイル名にして、ファイルの拡張子を適当な画像ファイルのものに変更し、ヘッダを画像ファイルのヘッダに書き換える、というものである。このために、ファイル名やファイルサイズを一覧できるツールを作成し、その規則性などから不正コンテンツの検出を試みる。

また、既に広く知られている偽装ツールに関しては、ヘッダなどの情報から偽装されていることが検出できる場合がある。このような、既によく知られたツールに個別対応する検出ツールの作成も有効である。

4. 実装

前節までの考察に基づき、現段階ではファイルのリスト表示ツールと、UG 語検出ツールの実装を試みた。いずれのツールもプログラミング言語としては Perl を使用している。チェックの際には、WWW サーバを直接利用でき、サーバにおいて各ユーザがインターネットに公開している全ファイルが直接参照可能であることを前提にした。

5. 1 ファイルリスト表示ツール

WWW において公開されているコンテンツのチェックに用いるツールであるため、各ユーザのホームディレクトリにある WWW 公開用のディレクトリ（本学の設定では public_html）以下に存在するファイルの一覧を表示するようにする。

まず事前に各ユーザのホームディレクトリの一覧をテキスト形式のファイルとして生成しておく。そのファイルを読み込みながら、ホームディレクトリに WWW 公開用のディレクトリが存在する場合において、そのディレクトリの内容をリスト形式で表示する。サブディレクトリが存在する場合も再帰的にディレクトリの内容を表示する。表示する項目としては、存在するファイルへの絶対パスと、ファイル名、ファイルサイズを＊を区切り文字として表示している。プログラムの結果（ファイルの一覧）を CSV 形式のファイルとして生成することにより、Excel での処理が可能となり、ファイルサイズでの並び変えや、ファイル名の規則性などを容易に見抜くことができるようになる。

5. 2 UG 語検出ツール

WWW サイト内で使用されている UG 語の検出には、茶筌 [2]を利用する。茶筌は日本語の文章に対して形態素解析を行う際、ユーザ辞書として、茶筌付属の辞書にはない単語を追加することが可能である。本研究では、収集した UG 語を詳しく品詞分類することはせず、UG 語そのものを「その他」分類の品詞の一部としてユーザ辞書に登録している。

実装は Perl を用いているため、茶筌付属の Chasen.pl を Perl モジュールとして使用している。

本ツールが UG 語検索の対象とするファイルは、日本語が使用されているテキスト形式のファイルである。html 形式のファイルだけではなく電子掲示板のログデータ等も対象となる。テキスト形式のファイルの判別には、File::Mmagic モジュールを利用した。プログラム本体は、チェック対象であるテキスト形式のファイルを読み込みながら、漢字コードの変換を行い、形態素解析を行う。その形態素解析の結果の中から UG 語である部分を数え上げ、チェック対象のファイルに含まれている UG 語の数とする。プログラムの実行結果として、チェックしたファイルの絶対パス付きのファイル名と、検出された UG 語の数を対にして一覧表示し、管理者はこの一覧を見て、UG 語が使用されている可能性がある WWW ページを知ることができる。不正コンテンツが配布されているかどうかの最終的な判断は、管理者が実際に当該 WWW ページを調べることによって行う。

5. 評価

今回実装した UG 語検出ツールを用い、実際に本学システム情報化センター内の 680 のユーザから得た WWW ページ（テキストファイルのみの積算で 10517 ファイル、約 30Mbytes）について調査をした。UG 語辞書として、約 200 単語を収集し使用した。茶筌では単語ごとにコストを設定する(数値が大きいほど出現頻度が低い)が、UG 語のコストは全て 100 で行い、その代わり通常の「その他 - UG 語」の品詞のコストを 2（つまり他の単語の 2 倍で計算する）とした。

調査には Intel Celeron500MHz のマシンを用いて、1225 秒の時間を要した。結果、UG 語が検出されなかったファイルが最も多く、10357 ファイルあった。以下、1 語～6 語 UG 語を含むファイルはそれぞれ 107、35、5、6、2、1 ファイルであった。それ以上の語数 UG 語を含むものは 4 ファイルしかなく、13 語および 17 語含むものがそれぞれ 2 であった（これらはそれぞれ同じファイルのコピーだったので実質で 2 ファイル）。

UG 語が 10 語以上検出された 4 ファイルについて、直接調べたところ、実際に UG 語が使われていたが、一般語としても使われるものであった（具体的には「アップ」「geocities」「lzh」の 3 つ）。実際にはこれらのページは不正コンテンツ配布には使われていなかった。

現在収集している UG 語は、文脈によっては一般語としても使われる語を多く含んでおり、茶筌での処理は厳密な文法や意味を考慮しないため、UG 語として判定されてしまうことがある。

しかし、そのような精度の悪さがあるとはいえ、UG 語が実際に多数使われている場合は、その WWW ページを高い確率で検出でき、システム管理上有用と考えられる。

6. 終わりに

本システムの実装により、WWW サイトにおいて公開されているコンテンツチェックという作業の効率を上げることができた。

今後の課題としては UG 語の検出精度を上げることが考えられる。具体的には、UG 語を一括して同じ品詞で扱っているため、形態素解析の精度が下がっているが、きちんと品詞分類することにより精度を上げられると考えられる。また、単語ごとの重み付けのチューニングも行うことによりいっそう精度は高まる。しかし、UG 語は日々新しいものが生まれ、古いものはすぐに使われなくなるので、UG 語チェックの精度を保つためには UG 語辞書の更新は重要な作業である。よって、UG 語辞書の更新作業を容易にするためにも、細かいチューニングを行わずにすむ適切な重み付けを模索したい。

もう一つの問題点として、一般的に用いられている言葉が、UG 語として辞書に存在する場合を検出していることである。これを避けるには、UG 語の中で一般にも使われる言葉を排除することであるが、これは逆に一般でも使われる語を数多く隠語の文脈で使っている WWW サイトを見落とす危険がある。この取捨選択について評価することも今後の課題として挙げられる。

参考文献

[1] 郵政省「インターネット上の情報流通の適正確保に関する研究会」報告書、2001 年 12 月 20 日

[2] 松本,北内,山下,平野,松田,高岡,浅原:形態素解析システム「茶筌」Version2.2.7 使用説明書,2001.6