

## インターネット中の特定情報識別技術の研究

于 浩, 王 主龍, 支 東霖  
哈爾濱工業大学, 哈爾濱, 15001, 中国  
{yu;dragon;zdl}@mmlab.hit.edu.cn

西野 文人, 松井 くにお  
富士通研究所,  
{nishino;matsui.kunio}@jp.fujitsu.com\*

### 1. はじめに

インターネット上にある著作権やプライバシーなどを侵害した不当な情報を監視したいという要望が高まっている。我々は、インターネット上の特定情報、特に“歌詞”を識別するシステムを開発した。ここでは識別効率とシステム構築コストを考え、キーワードと逆キーワードの認識、記号特徴認識、文章スタイル認識などの算法を導入し、さらに言語汎用性を向上させるために言語依存部と非依存部を分離した。中国語および日本語の Web ページでテストした結果、日本語の方がやや劣るものの十分な精度で歌詞発見ができることを確認した。

### 2. Web 情報特徴

本特定情報識別技術は Web ページを対象としている。Web ページの特徴には以下のものがある。

#### 1) 多様な書式

HTML におけるタグは表示方法を規定しているだけであり、それが意味するところは決まっていない。特に最近の HTML エディタで作られた HTML 文書の中には無意味なタグを多く含むことがある。

#### 2) 多テーマ

一つの Web ページの中にいろいろなテーマが含まれることがある。

#### 3) 自由な情報表現

情報は Web ページの中で自由に表現できる。またリンク情報や画像、音声などの情報を含めることもできる。Web ページの情報表現はとて自由で豊富である。

以上のような特徴から Web ページ情報の識別や処理は普通のテキスト情報と比べて多くの困難さがある。つまり、情報の位置の確定の問題のほかに、書式の変換や多テーマ情報の分離の処理も考えなければならない。

### 3. システム概要

#### 3.1 言語依存部と非依存部との分離

本システムでは、言語依存部と言語非依存部とに分離することで言語間の移植を容易にできるようにした。

システムの構成を図1に示す。

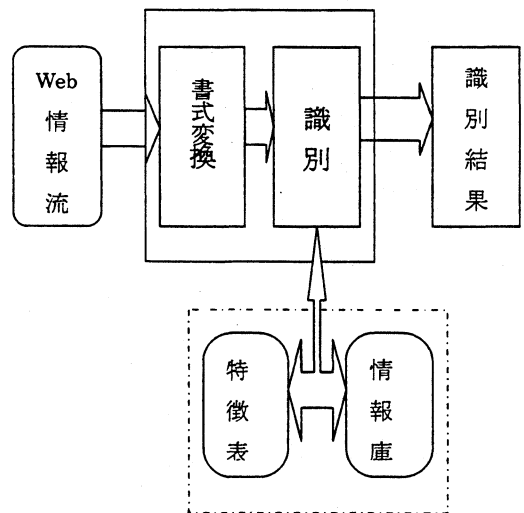


図1 システム構成

図1の中の実線の正方形で囲った部分が言語非依存部である。この部分は Web の書式の変換と分類（識別処理）を行うものである。書式変換モジュールは HTML 言語で表現された文書を一般の平文テキスト情報に変換し、その出力に対して識別モジュールが処理を行う。図1の点線の正方形で囲った部分は言語依存部分である。言語依存の特定情報タイプに関する情報を特徴表と情報庫の中に保存しておき、識別部ではこれらを利用して識別処理を行う。また識別部での処理結果は動的に情報庫にもフィ

ードバックされ情報が格納される。

### 3.2 情報のレベル別フィルタリング制御

Web ページには様々な種類の情報がある。また、一つのページの中に複数のテーマが含まれることもあるし、表現方式も様々である。我々は効率とシステム構築コストを考え段階的に制御を行うことにした。

段階的識別のシステムを図2に示す。

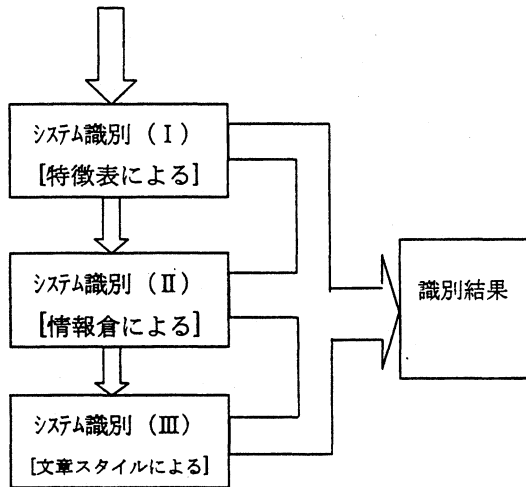


図2 システムの段階的識別

### 3.3 段階的識別の設計

Web 情報ストリームに対してまず第一段階で識別するが、出力は二つある。つまり、明確に識別できるものと明確に判断できない識別結果である（しきい値で判断する）。第二段階では明確に判断できない識別を処理する。このように順順に処理する。各段階の識別アルゴリズムは別々の方法を用いている。

## 4. 特定情報タイプ“歌詞ページ”の識別モジュールの設計

### 4.1 キーワードマッチング

インターネットの特定情報タイプを識別するために、実際の大量のホームページを分析した後を得たキーワードで特徴表を作った。特徴表により第一レベルの識別を行った。キーワードを利用し、以下の三つの特徴ベクトルを構築した。

$$T_c = (T_{c1}; T_{c2}; T_{c3}; \dots; T_{cn});$$

$$T_f = (T_{f1}; T_{f2}; T_{f3}; \dots; T_{fn});$$

$$T_w = s(T_{w1}; T_{w2}; T_{w3}; \dots; T_{wn});$$

但し、 $T_c$  は特徴ベクトル空間、 $T_f$  はそのキーワードの Web 中で出現の頻度数、 $T_w$  はキーワードの重み値ベクトル空間、 $N$  はキーワードの数である。

$T_a$  は三つのベクトルの各項をそれぞれ掛け算したものである。

$$T_a = T_c * T_f * T_w$$

一定のトレーニングを行うことで、比較的合理的なキーワード合計値を確立することができた。

次に歌詞ページという特定タイプ情報に対しての例を挙げる。大量の実際のスタイルが違うページを調査した結果、次のキーワードを得た。例えば：“歌詞：”、“歌手：”、“作曲：”、“歌詞”である。つまり、

$$T_c = \{ \text{“歌詞：”}; \text{“歌手：”}; \text{“作曲：”}; \text{“歌詞”}; \dots; T_{cn} \}$$

ところが、実際には自然言語の複雑性のため、あるキーワードは異なる言語環境においても、他の言葉と合わせて、異なった意味をもって利用されることもよくある。単なる特徴検索アルゴリズムだけを利用したのでは、誤判断率が高くなってしまったことがわかった。そこで、逆キーワード検索アルゴリズムを提案する。つまり、Web 中の明らかなキーワードは逆キーワードに含められる場合にまずシールドする。例えば、システムの特徴詞表に“詞：”というキーワードがあったとき、Web ページの中に“关键词：”のような言葉があれば、当ページの値は高くなってしまふ。だから、“关键词：”のような言葉

を逆キーワードとしてシールドすることにした。

#### 4.2 情報庫検索と情報庫フィードバック制御

情報庫は言語と特定情報タイプに関連した語彙データベースである。歌詞判定の場合には、歌手や曲名、歌い出しなどのデータを情報庫に格納することが考えられる。識別する際には、識別対象を情報庫から検索することになる。

しかし、はじめは情報庫にある情報は少ないから、フィードバック制御を導入した。つまり、もしシステムがあるページを歌詞ページと識別した時、そのページから重要な情報（例えば、歌手、曲名、歌詞の先頭1フレーズなど）を抽出し、動的に情報庫に格納する。

#### 4.3 文章スタイル識別

歌詞ページは普通のニュース、スポーツ、広告、電子書籍、フォーラムページなどと比べ、文章スタイルが大きく異なる。普通のページは段落の終わりでの行長は他の行に比べて短い。歌詞ページではセンテンス長の変動は一般的に非歌詞ページより小さい。つまり段落における各行の長さや行長の変動値などの情報を利用して識別ができる。

具体的なアルゴリズム：大段落のテキストを含めたページをまず分ける。ほかのページ中の各段落の長さを計測する。例えば、段落に付けた番号を横座標にし、各段落の長さを縦座標にし、線を描く（図3に示す）。

ただし、各行の平均値を歌詞ページの標準長とする。それから、機械学習の方法を利用し、適当な波動値と波動値以外の段落数を確定する。獲得した波動値と段落数でページを識別する。

### 5. 実験結果

中国語と日本語を対象言語として、識別システムを開発し、テストを行った。処理速度としてはPIII450Hzのパソコンで1日に100万のWebページを処理することができ、基本的に実用化の要求を満足した。

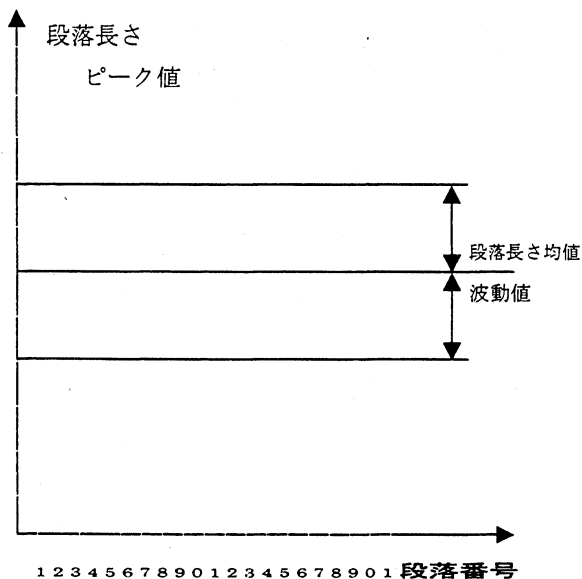


図3 文章スタイル図

#### 5.1 中国語版

2000（80種類）以上の歌詞ページと約1000の非歌詞ページ（基本的に歌詞らしさを示すキーワードを含むページと文芸ページを収集した）でテストした。

システムによる識別の評価結果を以下に示す。

	歌詞ページ	非歌詞ページ
実際	2000	1000
判断結果	2022	978
分析結果	非歌詞に判断	歌詞に判断
	14	36
適合率	1986/2022 =98.2%	942/978 =96.3%
再現率	1986/2000 =99.3%	942/1000 =94.2%

## 5.2 日本語版

700 (80 種類) 以上の歌詞ページと 200 の非歌詞ページ (キーワードを含めたページと同サイトページ) でテストした。

システム識別結果は以下のよう示す:

	歌詞ページ	非歌詞ページ
実際	700	200
判断結果	727	173
分析結果	非歌詞に判断	歌詞に判断
	9	36
適合率	691/727 =95.1%	164/173 =94.8%
再現率	691/700 =98.7%	164/200 =82.0%

## 5.3 実用性について

実際の歌詞判定の運用では、ある程度歌詞らしい手がかりのあるページ (今回の非歌詞ページはこのようにして収集したものである) に対して歌詞ページを発見するというものを想定している。このとき、歌詞を含むページに対して歌詞であることを見逃してしまう割合 (1-歌詞の再現率) は、中国語版 0.7% に対して、日本語版では 1.3% ということになる。一方、システムが歌詞であると判定したページの中に歌詞でないページ (雑音) が含まれる割合は実際に歌詞と非歌詞との割合が問題になるが、(1-非歌詞の適合率) で見ると、中国語版 3.7% に対して日本語版 5.2% ということになる。これは日本語のページの方が中国語に比べて形式的な特徴が少ないことが原因である。

## 6. まとめ

インターネット特定情報タイプ識別するために Linux 上で歌詞ページ識別実用システムを開発した。Web 情報の表示方法と普通のテキスト表示方法は大きな異なりがあることや汎用性やシステムコストなどを考え、システムレベル別識別制御と言語依存・言語依存しない部分に分けた識別方法を提案した。テストから、開発したシステムの十分な識別速度と効果を得た事が分かった。

現在、システムの対象は HTML 言語で書かれたページだけであり、XML を対象とする事が今後の課題である。

### 参考文献

1. An Binbin, Wang Peikang. Research on the Recognizing Algorithm of Web Pages. QingBaoXueBao 2001,2(77-81)
2. Craven M, Distasco D, Freitag D, et al. Learning to Extract Symbolic Knowledge from the World Wide Web. Technical Report. CMU-CS-98-122. School of Computer Science, Carnegie Mellon University, 1998-09
3. ZHU Ming, WANG Jun, WANG Junpu, The Study of Feature Selection in the Web Page Classification. Computer Engineering Vol.26. No8 35-37
4. Mark Cracem. Learning to extract symbolic knowledge from the world wide web. In Proc. Of the AAAI Fifteen National conf. On Artificial Intelligence, 1998.
5. Soumen Chakrabarti, Byron Dom, Piotr Indyk. Enhanced hypertext categorization using hyperlink. Proc. of ACM SIGMOD 98 Seattle, Washington, 1998
6. Craven M, Distasco D, Freitag D, Freitag D, et al. Learning to Extract Symbolic Knowledge from the World Wide Web Technical Report. CMU-CS-98-122. School of Computer Science, Carnegie Mellon University, 1998-09
7. Naveen Ashish and Craig A. Knoblock. Wrapper generation for semi-structured internet sources. SIGMOD Record, 26(4):8-15, 1997.