E-026

JPドメインにおける茶筌を用いた中国語ページの抽出

Extraction of Chinese pages from JP domain using ChaSen

魏 小比　内藤 一兵衛　上田 和紀
Xiaobi WEI　Ichibe NAITO　Kazunori UEDA
早稲田大学大学院理工学研究科　早稲田大学大学院理工学研究科　早稲田大学理工学部

概要: JPドメインには中国語ドキュメントが多く存在するが、あまり有効利用されていないのが現実である。中国語ページを抽出できれば、中国語を扱う人々に便利さをもたらすと共に、統計・語学・検索エンジンのデータベースなど様々な応用研究もできる。

本研究では、形態素解析ツール茶筌を使用し、早稲田大学（88,634 pages）と北京大学（25,421 pages）のWEBページを全面的に分析し、単語の品詞種類と一文字で区切られる形態素の割合から中国語文章の特徴を突き止め、JPドメインにある多国語の混在しているHTMLファイルから中国語で書かれたページを抽出する手法を考案し、実行する事に成功した。また、その延長である様々な応用の可能性についても述べる。

1. 研究背景と目的

JPドメインのWEBサイトに中国語ページが作られているということは、何らかの意味を持って中国語を扱う人々にアピールしていると考えられる。しかし, その情報がターゲットに行き渡らず、また、ターゲットが欲しがるその情報に辿り着けない事が多い。そこで、本研究ではユーザーにとって参考価値のある中国語ページの抽出を目的とする。

今まで、文字コードで判定する方法等が提案されていたが、近年、文字コードの多様化につれ、一つのファイルの中で複数の文字コードを混在させる事や、他言語の文字を自国の文字コードだけで表現する事ができる。そのため、言語の判定が困難である。本研究ではファイルに書かれている言語を判別する為に、文章の意味解析を使用することを提案する。

2. 中国語文字コード

中国の国家規格は「国家標準」(GuojiaBiaozhun;GB)といい、2バイト文字コードの規格も簡体字・繁体字それぞれにいくつか定められている。

現在、最も使われている文字コードは、GB 2312-80(以下GB2312)、GB 13000.1-93(以下GBK)とGB 18030-2000(以下GB18030)の三つである。

GB2321(漢字6,763字)は1980年に制定されてから最も使われている。1995年にGB2312の拡張であるGBK(漢字20,902字)「漢字内碼拡張規範」が制定された。GBKは、Microsoft Windows95の簡体字中国語コードページ(CodePage 936)に使われていたため、広く普及している。また、GB2312の全ての文字がGBK上でも同じ符号位置にあるため、互換性を保っている。

更に、2000年にGBKコードの後方に定義されていない漢字を追加することで互換性を保ちつつ、今後の文字追加の可能性も考慮した拡張コードGB18030(漢字27,484字)が制定された。

2.1 漢字コードの問題点

GBコードにおいて2バイトで一つの漢字を表現しているところは、日本語などのマルチバイト文字と同じである。

GBコードに片仮名、平仮名やハングル文字も定義され、逆に日本語や韓国語コードにも中国の漢字の一部が定義されている。中国語と日本語や韓国語などと、字形違いを含む同じ漢字のコードが重なったり、同じ漢字でもコードが異なったりすることで、文字コードだけでは言語種別を判断することが難しい。

2.2 中国語 WEB ページの文字コード

中国ではGBコードを国家標準としているため、国内で公開されるソフトやコンテンツは、それに準じなければならない。現在、GB18030(GB2312)が政令により内外のベンダーが採用を義務付けられている。

また、常用漢字はGB2312に収まっているため、現状ではWEBページに書かれている中国語文字はGB2312だけで表現できると言える。

この研究の元となるWEBページの中、中国語で書かれているページは、ほぼ100% GB2312に沿っている。

3. 中国語ページの抽出

本研究の流れとしては、WEBページを収集し、HTMLタグを取り除き、中国語茶筌でテキストを解析し、最後に統計を行う、といった手順である。

本研究に使用する三種類のデータは、WEB収集ロボットを使用して集めた「北京大学 25,421 pages」,「早稲田大学 88,634 pages」,「JPドメイン 10,166,170 pages」である。

HTMLタグが文章に残っていては、正確な自然言語処理ができない。そのため、形態素解析を行う前に、HTMLタグとそれ以外のテキスト部分を分離する必要がある。そして、HTMLタグを取り除いたテキストファイルを形態素解析の元データとした。

中国語茶筌を使用し、三種類のデータをそれぞれ解析し、データを取った。

3.1 中国語茶筌

中国語茶筌は、日本語形態素解析ツール「茶筌」をベースに、中国語に特化した形態素解析ツールである。

日本語の形態素解析器においては、連続した複数の未知語が一つまとまった未知語として解析されるが、本研究では、処理速度を向上させるため、未知語抽出部分のプログラムを使用せず、中国語の文書においても連続した未知語は一文字ずつ区切る。

意味のある中国語文章を中国語茶筌に渡せば、品詞付けし単語ごとに区切ってくれる。しかし、中国語でない文章となれば、中国語茶筌の辞書に存在しない文字列ばかりなので、一文字ずつ区切ってしまうことになる。

形態素とは、大まかに言えば、意味を持つ最小の言語単位である。中国語茶筌の品詞テーブルには、“g”という品詞があり、他の品詞種類に属しない最も小さな単位であるため、“g”に分類されるということは、その単語は中国語茶筌で扱われている辞書に存在しないということになる。また、アルファベットは、仕様上全て“g”に分類され、結果に大きな影響が与える可能性があるため、カウントしないものとする。

また、中国語、日本語と英語の文書を中国語茶筌で解析した結果のサンプルをそれぞれ図1、図2、図3に示す。

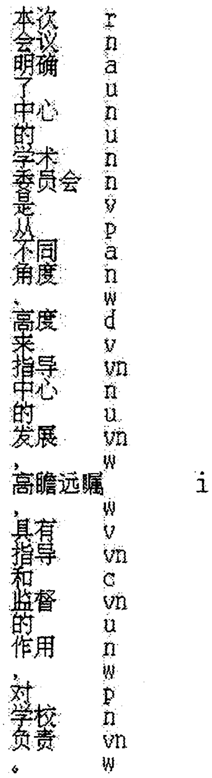

図 1: 中国語文書の解析結果

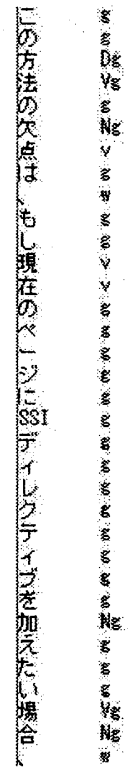

図 2: 日本語文書の解析結果

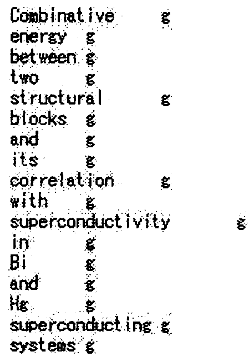

図 3: 英語文書の解析結果

上述にあった二つの特性、「一文字で区切られた形態素」と「“g”に分類された形態素」がそれぞれ全体形態素総数に占める割合を中国語判定のデータにし、中国語ページ抽出のポイントにした。

3.2 中国語ページが多いサイトの分析

上記の二つのパラメーターをページごとに測定し、北京大学ページの分布図を図4に示す。

X軸は「一文字で区切られた形態素の割合」であり、Y軸は「“g”に分類された形態素の割合」である。図4から分かるように、ページが特定のブロックに集まっ

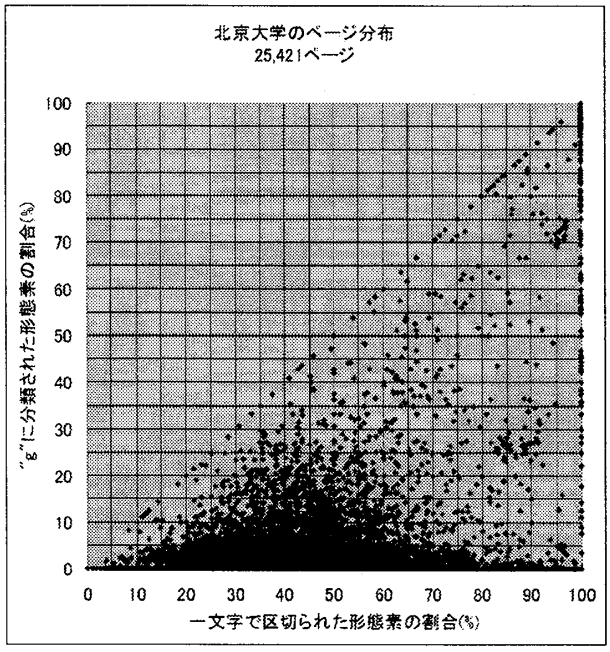

図 4: 北京大学ページの分布図

ている。その特性から、中国語ページを判定するポイントが導き出せる。

X 軸と Y 軸それぞれのデータから得た情報は、下記の通りである。

- 「一文字で区切られた形態素の割合」は 10% ～ 70% に集中
- 「"g"に分類された形態素の割合」は 20% に集中

また、集中ブロックから発散しているポイントも多く存在するように見えるが、数は極めて少ない。「中国語である集合」にあるページは全体の 61% を占め、アルファベットで書かれたページは 24% である。また、図 4 に示したように、(0,y) に付着しているポイントは 5% を占めているゆえ、残りのページはおよそ 10% である。アルファベット以外の「中国語でないページ」もその 10% の中に含まれているが、その他、まれに中国語で書かれたページも含まれていると考えられる。

事実上、発散しているファイルを個別に見たところ、その中は中国語で書かれたページも存在している。文章としての意味を持たない文字列を組み合わせたページは中国語茶筌で適切な処理ができない場合があるため、グラフの中で集中ブロックから発散するように見えていると考えられる。

本研究では、「ユーザーにとっては参考価値のある中国語ページ」をコンセプトに抽出しているので、判断しがたいページをカウントしないこととする。

3.3 中国語ページが少ないサイトの分析

同様の手法で早稲田大学のページにも処理を行ったところ、「一文字で区切られた形態素」が殆んどである上、「"g"に分類された形態素」の割合がかなり目立つ。その結果を図 5 に示す。

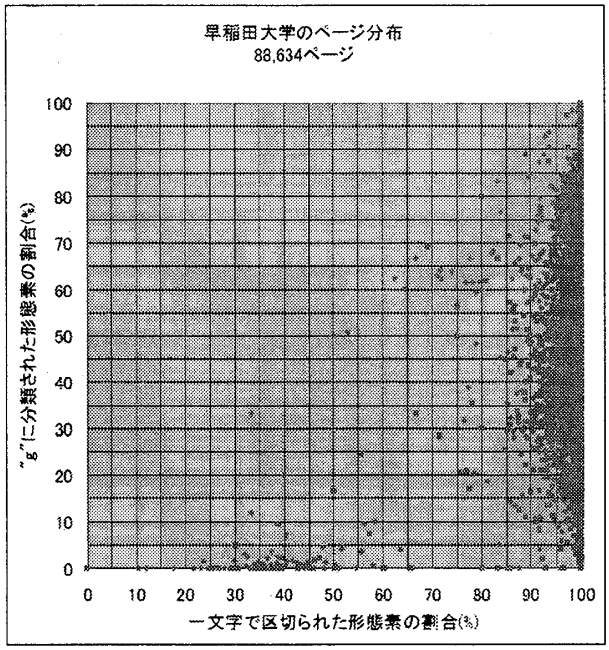

図 5: 早稲田大学ページの分布図

図 5 に示したように、早稲田大学のページが、(90～100,y) のブロックに集まっている。日本語で書かれたページがほとんどだという理由からである。

アルファベットで書かれたページが全体の 11% を占め、「一文字で区切られた形態素の割合＝ 100%」というパターンは 61% も占めている。Y 軸の値が集中しないのは漢字が多く使われている為である。

なお、形態素総数が極めて少ない場合、中国語文章として意味が持たないので、適切に形態素解析ができないため、判定上にずれが生じる。

また、実験データにより、形態素総数が「6」になった時点で、値を増やしても当てはまるページ数の変化がなくなったため、下記のようなルールを決めた。

3.4 ルール設定

北京大学と早稲田大学の WEB ページの分析結果より、次のような判定ルールを導き出した。中国語テキストとは、下記の三つ条件を同時に満たすものである。

- 形態素総数は 6 以上
- "g"に分類された形態素の割合は 20% 以下
- 一文字で区切られた形態素の割合は 70%以下

4. JP ドメインへのルール適用

4.5 節の導出ルールに基づいて JP ドメインから収集してきたページを分析した結果を図 6 に示す。

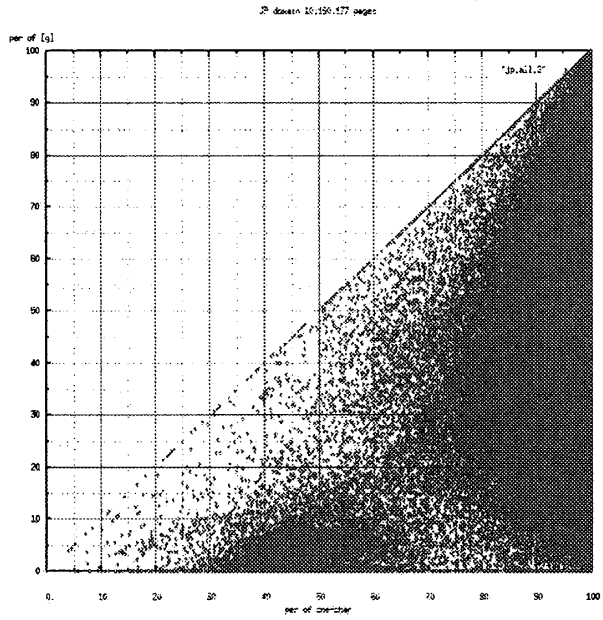

図 6: JP ドメインページの分布図

早稲田大学のページで取ったグラフと同じように、中国語であるページとそうでないページが異なったブロックに集まっていることが明らかになった。また、日本語やアルファベット以外にも中国語でない言語が多く存在するので、「一文字で区切られた形態素の割合」が 70% ～ 90% であるページも珍しくない。なお、北京大学の分布と同じように、「中国語ブロック」より発散した中国語ページも少し存在すると考えられる。

抽出ルールを適用した結果を表 1 に示す。

表 1: 中国語ページ抽出の結果

判定範囲	総数	中国語	全体の割合
北京大学	25,421	15455	60.80%
早稲田大学	88,634	110	0.12%
JP ドメイン	10,160,177	10110	0.10%

5. 考察

これまでに得た三つのページ分布図の共通点に注目する。全てのポイントは (0,0) と (100,100) を結ぶ斜線の右下にある。つまり、

$$\frac{\text{一文字で区切られた形態素の割合}}{\text{“}g\text{”に分類された形態素の割合}} \geq 1$$

表 2: 分析に要する時間

サイト／項目	形態素解析	抽出
北京大学	20 分	36 秒
早稲田大学	54 分	2 分 38 秒
JP ドメイン	58 時間 16 分	5 時間 58 分

ということになる。

中国語の文章は英数字を除き、全て漢字で構成され、中国語形態素の最小単位は一つの漢字であるため、“g”に分類された形態素は全て一文字となる。しかし、一文字の漢字でも文章の中で意味を持っている場合が多いので、一文字で区切られた形態素は“g”に分類されていない形態素も含まれている。本研究では、英数字をカウントしていないため、理論上“g”に分類された形態素の割合が一文字で区切られた形態素の割合より多い。

また、割合の少ないページを除き、ページが二つのブロックに集まっている。その二つのブロックとは、「中国語である集合」と「そうでない集合」と考えられる。

6. まとめと今後の課題

本研究では、JP ドメインにある 1000 万ページを対象に、中国語ページの抽出を行った。しかしながら、処理速度や特殊なパターンへの配慮など、まだまだ不備なところがあり、改善策を練らなければならない。

また、今回構築したデータは、検索エンジンへの拡張（JP ドメイン内にある全ての中国語ページを集めた検索サービスなど）や更なる自然言語処理（未知語データベースの構築など）にも応用できる可能性があり高い研究価値があるため、研究を続けていきたい。

参考文献

[1] Chooi-Ling GOH : Chinese unknown word identification based on morphological analysis and chunking, 自然言語処理研究会 26 May 2003.

[2] Chooi-Ling GOH, Masayuki ASAHARA, Yuji MATSUMOTO. Chinese unknown word identification using characted-based tagging and chunking. In Companion Volumn to the Proceedings of ACL 2003, Interactive Poster/Demo Sessions, pages 197-200. 7-12 July 2003.

[3] Ken Lunde. CJKV 日中韓越情報処理. オライリージャパン, December 2002.

[4] ChaSen : http://chasen.naist.jp/

[5] Verno : http://verno.ueda.info.waseda.ac.jp/

[6] CNNIC : http://www.cnnic.cn/