2004年度 修士論文

JPドメインにおける
茶筌を用いた中国語ページの抽出

提出日: 2005年2月2日
指導: 上田和紀 教授

早稲田大学大学院理工学研究科
情報・ネットワーク専攻

魏 小比
学籍番号：3603U053

概 要

JP ドメインの WEB サイトにも中国語ドキュメントが多く存在するが、現状ではあまり有効利用されていない。中国語ページを正確に集めたサービスが存在せず、せっかくの豊かな資源が台無しである。

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。しかしながら、その情報がターゲットに行き渡らず、また、その情報を欲しい人も辿り着かない事が多いと考えられる。

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期待される。文字コードの多様化につれ、ファイルから言語を判別するのは、文章の意味解析を切口にしなければならなくなった。

この研究では、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つきとめ、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、実行することに成功した。また、その延長線上にある様々な応用の可能性についても述べた。

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

1.1.1　インターネットの発展

アイルランドの Nua Internet Surveys は、世界のインターネット利用人口 (2002 年 9 月末現在) を地域別にまとめた調査レポート「How Many Online?」を発表した。全世界人口の 1 割を占める 6 億 560 万人がインターネットを利用しているものの、利用者の分布図には地域別の格差が歴然と表われる結果になっている。

世界のインターネット利用人口の地域別割合でトップに立ったのは欧州。利用者数は 1 億 9,091 万人に達し、いまや 3 人に 1 人がインターネットを利用していることになる。特に東ヨーロッパ諸国へのインターネットの普及が著しい。また、地域別割合で 2 位となった、1 億 8,724 万人がインターネットを利用するアジア太平洋地域も、急速な伸びを見せている。(図 1.1)

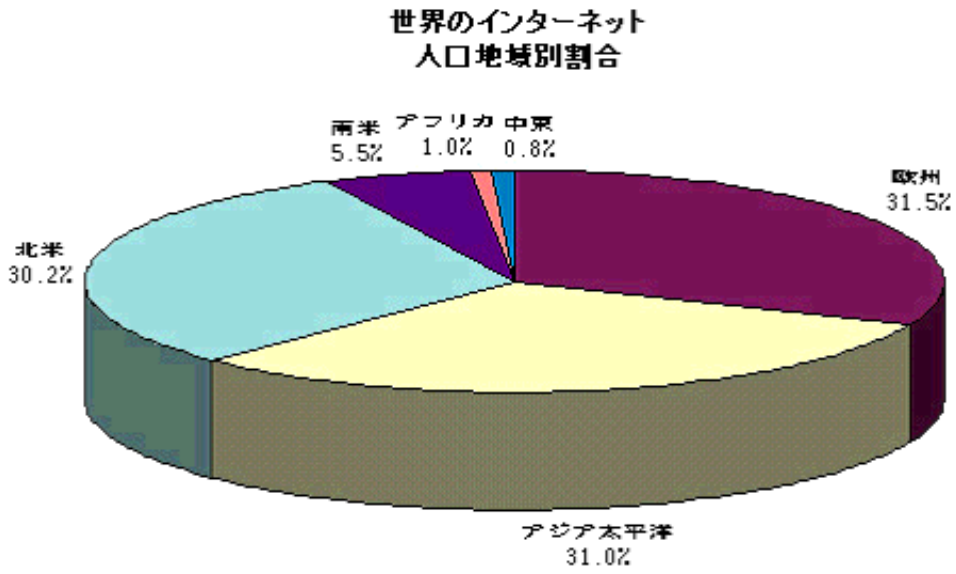

図 1.1: 世界のインターネット人口

また、中国ネットワークインフォメーションセンター（CNNIC : China Internet Network Information Center）は 2005 年 1 月 19 日、「第 15 回中国インターネット発展状況統計報告」を発表した。2004 年末までに、中国のインターネット人口が 9400 万に達した。アメリカに続いて、世界第二位になった。なお、ブロードバンド接続件数は 4280 万であり、CN ドメイン登録数は 43 万 2077、WEB サイト数は 66 万 8900 である。

そういった状況の中、中国語の豊かな WEB 資源を利用した研究やビジネスが数知れないほど存在している。

1.1.2 ドメイン

ネットワーク環境でのひとまとまりの管理対象をドメインと呼ぶ。ドメイン名は「インターネット上の住所表示」と言われ、URL(ホームページのアドレス) やメールアドレスなどの一部分として使われる。世界でただ一つの名前とするために、ルート (トップレベルドメイン) を頂点とした階層構造を持っており、名前の並びを“ . ”(ドット) でつなげた構造をしている。

「.jp」は日本を表すトップレベルドメインであり、「.cn」は中国のものである。しかしながら、日本のレジストラからも「.cn」を登録申請できる今現在、ドメインはもはや国別の意味を持たなくなった。

実際、他国のドメインにある WEB サイトをどれくらいアクセスしているかを見てみよう。CNNIC の調査データから香港のインターネット利用者の例（図 1.2）を挙げよう。

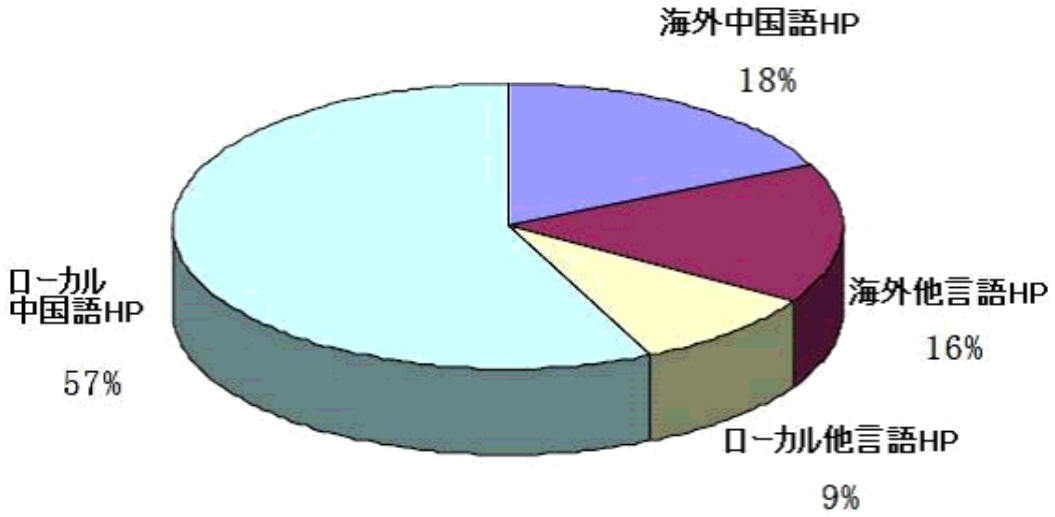

図 1.2: 香港のインターネット利用者が閲覧する WEB サイトの言語別割合

ご覧の通り、海外ドメインにある中国語ページは 18% も占めている。

JP ドメインの WEB サイトにも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。

1.1.3 日本と中国

1978 年に「日中平和友好条約」が調印され、現在に至り約 30 年近く良い関係を保っている。日本法務省入国管理局の最新統計データによると、2003 年の時点で、462,396 人の中国人が在日しており、在日外国人の全体の 24.1%を占めている。具体的な統計データがないが、利用金額の安いブロードバンドが普及している日本では、在日中国人のインターネット利用者がかなりの確率でいると思われる。

また、中国に関心を持つ日本人も多い。語学、ビジネス、旅行など、中国に関する情報を欲する人が大勢いると考えられる。

JP ドメインの WEB サイトにある中国語ページを正確に集めたサービスが存在せず、せっかくの豊かな資源が台無しである。

1.2 研究動機

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。しかしながら、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かない事が多いと考えられる。

JP ドメインの WEB サイトから中国語で書かれたホームページを正確に集められれば、それを解決できると考えられる。

以上の問題を踏まえて、自然言語処理の観点から、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、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正確に抽出する手法を考案し、実行した。

1.3 本論文の構成

本論文では、第 2 章で中国のインターネット事情を語りつつ、メジャーな中国語サーチエンジンやそれに関連する研究について紹介し、JP ドメインにある中国語ページについて分析する。第 3 章では、自然言語処理について言及し、中国語の特徴と中国語茶筌について紹介する。

第 4 章では、ページの収集から中国語形態素解析までの流れについて述べる。

第 5 章では、実際に北京大学と早稲田大学のページを中国語茶筌で処理し、そのデータを元に抽出ルールを決め、実行した。

第 6 章では、本研究の問題点について取り上げ、最後に、第 7 章で今後への課題などを取り上げつつまとめる。

1.4 研究環境

この研究に使ったコンピュータの構成は表 1.1 の通りである。

表 1.1: コンピュータの構成

ホスト名	morbier	cheddar
CPU	Pentium4 2.4GHz	Athlon64 3200+
Memory	512MB	1GB
HDD	45GB	160GB(システム用)+800GB(データ用)
OS	FreeBSD 4.7	SuSE 64bitOS 9.0
用途	ページ収集・テスト	形態素解析・各種統計・抽出処理

(2005 年 2 月現在)

第2章　中国の WWW 事情

2.1　中国のインターネット事情

2005 年 1 月 6 日、北京市婦産医院で男児が産声を上げ、中国の人口が１３億人を突破した。

2.1.1　インターネット人口の増加

CNNIC の調査によると、2004 年 12 月 31 日の時点で、中国のインターネット人口は 9400 万人に達し、2003 年の同時期に比べ 18.2% 増えた。更に、1997 年 10 月の一回目の調査に比べ、現在のインターネット人口は当初の 151.6 倍となった。(図 2.1)

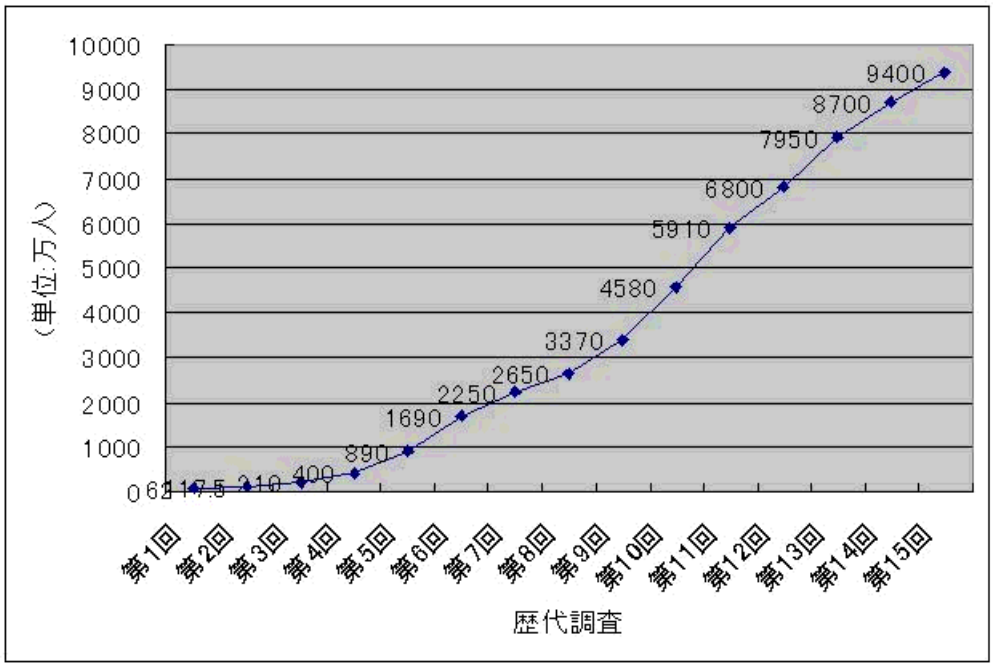

図 2.1: 中国インターネット人口の推移

2.1.2 インターネットに接続している PC 台数の増加

また、ネットに接続している PC の総台数も同じように毎年伸びている。PC の低価格化が激しく進んでいるため、中国では PC はもはや家電のような存在となった。2003 年の同時期に比べ 34.7% 増え、一回目の調査の 139.1 倍となった。(図 2.2)

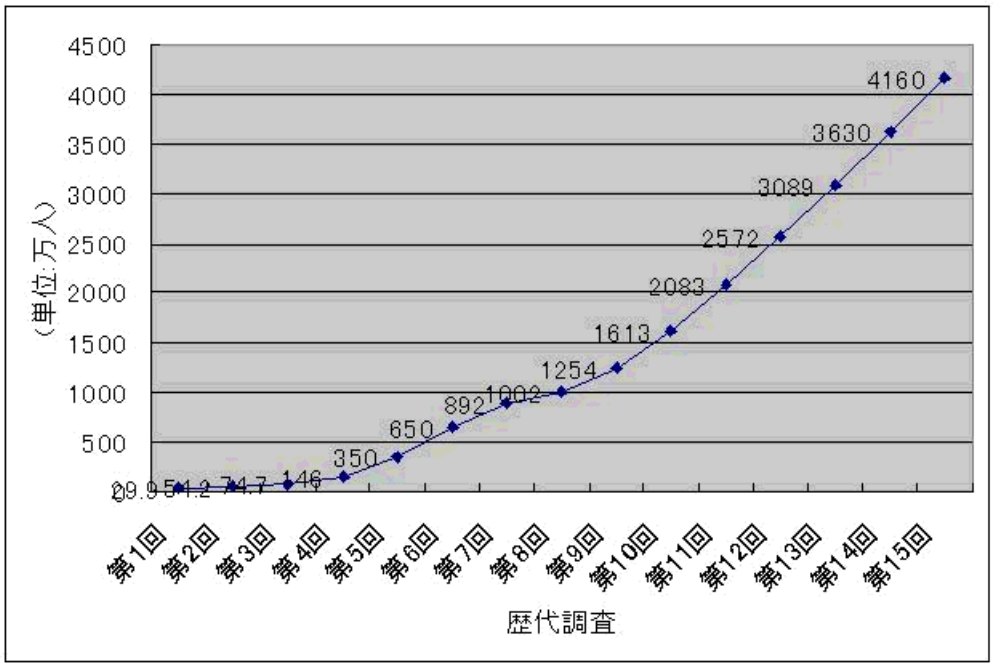

図 2.2: インターネットに接続している PC 総台数の推移

2.1.3 WEB サイトの事情

インターネット人口が増え続ける中、WEB サイトも急速に発展し、情報の大洪水が押し寄せた。

「WEB サイトを運営すれば金儲けができる」という信者が日々増え、インターネット広告の激しい取り合いとなり、ポップアップ広告が本来のページを覆うくらいであった。ダイヤルアップが主な接続方法だった時代、いったんページを表示させその後オフラインで見るという裏技もあったほどだ。

一見安全そうなサイトでも、アクセスするだけでソフトが勝手にインストールされることが珍しくない。

そういった状況の中、目的のサイトに少しでも早く辿り着くには、サーチエンジンが大きな役目を果たしている。

2.2 中国のサーチエンジン

CNNIC の調査データによれば、「インターネットに接続しどのようなサービスを使っているか」というアンケートでは「サーチエンジン」を選んだのは 65.0% であり、「電子メール」に続いて 2 位となっている。また、「WEB サイトの情報を得る手段」という問題には「サーチエンジン」が 86.6% で断然 1 位であった。何も中国に限った話ではないが、サーチエンジンの影響力は大きい。

2.2.1 メジャーなサーチエンジン

現在中国でサービスしているメジャーなサーチエンジンは、「百度 (Baidu)」「Google」「Yahoo」「3721」「捜狐 (Sohu)」「新浪 (Sina)」などがある。

中でも、“ 世界最大の中国語サーチエンジン ”と謳っている「百度[1]」(図 2.3) は大きなシェアを握っている。更に、2004 年 6 月、Google から融資を受け、事業を拡大している。

図 2.3: メジャーな中国語サーチエンジン-百度

「百度」はページランキングの算出によって順位を決めている「Google」と異なり、検索キーワードに広告主の出しているキャッチフレーズが含まれているならば、広告主が支払った金額でヒット順位を決めている。

商用サーチエンジンはあふれているが、研究目的として公開されているものはほとんどない。

[1]http://www.baidu.com/, http://www.baidu.jp/

2.3 中国語の文字コード

2.3.1 よく使われている文字コード

中国本土の国家規格は GB(GuojiaBiaozhun「国家標準」) 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。そのうち最もよく使われるのが、GB2312 という簡体字の規格である。

実際には GB2312 は GR(図形文字の領域) に呼び出されて EUC として使われるケースが多い。ちょうど日本語 EUC の G0・G1 と同じような形になる。実際、GB2321 は EUC_CN ともよばれている。

ただし、インターネットのメールやニュースでは 8 ビットデータの通過が必ずしも保証されないので、 alt.chinese.text では HZ(HanZi=漢字) とよばれる独自の 7 ビットエンコーディングを用いている。

また、1980 年に GB2321(漢字 6,763) が制定されてから今でも最も使われているが、1995 年に GBK(漢字 20,902 字)「漢字内碼拡張規範」が制定され、CJK 統合漢字の全てが収録された。また、2000 年より、GB2321 や GBK とも互換性を保ちつつ、今後の文字追加の可能性も考慮した GB18030(漢字 27,484 字) が制定された。「国家標準」ということもあり、内外のベンダーが採用を義務付けられている。今後少しずつ GB18030 にシフトしていくと思われる。

2.3.2 WEB ページに使われている文字コード

簡体字の WEB ページはたまに GBK と HZ も見かけるが、現状ではほぼ 100% といっていいほど、ほとんどのページが GB2321 に作られている。

2.4 N-gram と 形態素解析

検索システムにおいて、検索を行うためにはその文書の内容を表す要素、索引語 (index term) を抽出する必要がある。英語に関しては、空白やピリオドなどの記号によって単語を切り分けることができるので、比較的容易に索引語を抽出することが可能である。しかし、中国語の文には英語のように語と語との間に明白な区切りが存在しないため、何らかの索引付けの単位に文章を分割する必要がある。

2.4.1 形態素解析

中国語テキストの索引付けの方式には、「ワード型」と「N-gram 型」があり、形態素解析を利用した索引語の抽出は「ワード型」にあたる。ワード型は中国語

の文章を形態素の単位に分け、ある程度品詞情報をもった形で索引付けすることができるので、中国語の検索がよりしやすくなる。例えば検索キーワードにおいても形態素解析の処理を行うことで、文章が入力されても、単語ごとに分割しその各単語による検索を行うように処理することができるため、よりユーザーの意図に近い検索結果を返すことが可能になる。

しかし、一方で索引化、検索時共にその処理のさい、辞書を用いるため、新語・造語に対応させるには辞書の更新作業をしなければならないといった欠点がある。

2.4.2 N-gram 方式との比較

N-gram 方式には形態素解析に必要な辞書が不要であるため、索引作成が容易、索引語とユーザーが入力した検索キーワードの分かち書きによる検索の漏れがないといった利点がある。

- 形態素解析

  メリット 索キーワードがヒットした場合、一定水準以上の精度の検索キーワードがヒットした場合、一定水準以上の精度の高い検索が行なえる

  デメリット 検索精度を維持するためには辞書のメンテナンスが必要検索精度を維持するためには辞書のメンテナンスが必要

- N-gram

  メリット 辞書に依存せず、検索漏れは理論上ありえない辞書に依存せず、検索漏れは理論上ありえない

  デメリット 検索語によっては全く関係ない単語をヒットさせてしまう

形態素解析を用いた索引付けは全文検索に適用した場合、クライアントのリクエストに対する検索漏れが生じるが、ユーザーの意図により近い検索を行うためには役立つ。

2.4.3 形態素解析における日本語と中国語の比較

日本語では、単語の変形が存在する。

日本語キーワードに形態素解析を用いた例

検索語句：中国に行きたい
解析結果：中国 に 行く たい
検索語として採用：中国　行く

全文検索においては、上記の例では、「行きたい」という文字列を含む文書を検索する。従って、「行った」「行きます」といった文字列を含む文書は検索できない。これらの文字列を索引付けの際、形態素解析で「行く」という動詞として索引付けを行っておけば、実際の検索において同じような意味を示す言葉を同時に検索することが可能である。

中国語では、単語の変形が存在しないが、漢字を自由に組み合わせて新たな単語が作れるため、辞書に登録されていない未知語となる単語が多く存在する。そのため、日本語より形態素解析が困難である。

2.5　JP ドメインに存在する中国語ページ

日本には、中国語を扱う人々をターゲットとしたビジネスが多く存在する。大企業などのホームページに、「中国語版」へのリンクが貼られ、日本語版のコンテンツと全く遜色ない中国語のホームページが作られていることも多く見られる。また、日本人をターゲットとしない中国語 WEB サイトも JP ドメインに多く存在する。

日本では、JP ドメインを確保するのにかかる費用は、他のドメインより高価[2]な場合がほとんどである。発信側が、わざわざ値段の高い JP ドメインを取るということは、何らかの意図を持って中国語を扱う人々にアピールしていると言える。

[2]例：お名前.com ではもっとも値段が高い

第3章　自然言語処理

3.1　自然言語処理とは

3.1.1　歴史

自然言語処理とは、人が通常のコミュニケーションに用いている言語を、コンピュータにより理解したり生成したりする技術であり、人がコンピュータと直接コミュニケーションする場合の基本技術である。自然言語処理の研究は、1950 年代のコンピュータに翻訳をさせる試み（機械翻訳）の研究から始まった。米国において露英翻訳の研究が始まり、日本でも九州大学、電子技術総合研究所で開始された。

3.1.2　応用

一方、自然言語処理技術の応用として、キーワードによる検索技術が 1990 年代に開発され、文書の自動分類や検索が可能となった。この技術には、機械翻訳研究で確立された単語の切り出し品詞の割り当て技術（形態素解析）が使われている。また、WEB サイトの検索には WEB 検索サービスが欠かせないものとなっているが、この実現には、並列コンピュータによる高速単語サーチ技術と形態素解析が組み合わされている。

3.1.3　形態素解析

日本語、中国語、韓国語／朝鮮語、タイ語などアジアの言語で書かれた文章では、単語が空白で区切られていないため、そのままでは単語単位のインデックスを作成することはできない。これを可能にする技術が形態素解析である。つまり、形態素解析とは、文章から単語を切り出す技術であり、厳密に言えば、意味を持つ最小の言語単位（形態素）の範囲を抽出し、品詞や読みなど形態素の属性を同定する技術である。

「KAKASI」「MeCab」「ChaSen」「Juman」「Breakfast」「すもも」などと日本ではメジャーな形態素解析ツールはいくつもあるが、この研究では、奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 (松本研究室) が開発した

ChaSen「茶筌」を使用した。

3.2 中国語の形態素解析

3.2.1 中国国内の研究

「形態素解析」は中国語では「詞素解析」と言い、中国国内での研究は少ない。茶筌のように誰でもフリーにダウンロードできるツールは、公開されていない。検索エンジンに搭載される事が少なく、機械語翻訳もかなり精度が低い。メジャーな検索エンジンでサーチしても、ヒット数が少ない上に、「海外・日本では形態素解析はどのような研究をされているか」・「日本国立情報学研究所 NII ではどのような結果を出しているか」といった情報ばかりである。それほど注目度が低いのだ。

3.2.2 中国語の特徴

中国語が日本語と同じく単語間にスペースがないため、文節の区切りが自動的に見つけにくい。中国語の解析においては、わかち書きにより単語の境界を明らかにする必要があるが、未知語が多くなるにつれ、解析が非常に困難となる。

また、単語の活用（変化）形がない中国語の文章は、ほぼ全て漢字で構成されているため、自由に文字を組み合わせて新たな単語を作る事ができる上、略語・新語・造語などが多く存在するため、全ての未知語を辞書に登録するのは不可能に近いと言える。

事実上、形態素解析を搭載しているメジャーな中国語エンジンでも、長い検索キーワードを分析させた結果、一つの形態素となる文字列があまりにも長くなる事はしばしばある。

そのため、検索キーワードの意味合いを分析しヒット精度を向上させるための形態素解析が逆効果になってしまうので、中国語の検索エンジンでは、N-gram の方が盛んである。

3.3 茶筌と中国語茶筌

3.3.1 茶筌

茶筌システムの原形は、京都大学長尾研究室および奈良先端科学技術大学院大学情報科学研究科において開発された日本語形態素解析システム JUMAN(version2.0) である。JUMAN は、京都大学・奈良先端科学技術大学院大学のスタッフおよび多くの学生の協力を得て作成されたものだ。また、辞書に関しては、Wnn かな

漢字変換システムの辞書、および、ICOT から公開された日本語辞書を利用し、独自に修正を加えたものだ。

3.3.2 中国語茶筌

中国語茶筌は、同じく奈良先端科学技術大学院大学情報科学研究科松本研究室によって開発された茶せんを中国語に特化した形態素解析ツールである。

品詞の種類と名称は表 3.1 の通りである。

表 3.1: List of part-of-speech tags

POS	Description	POS	Description
a	Adjective	Ag	Morpheme used in adjective
b	Noun-modifier	ad	Deadjectival adverb
c	Conjunction	an	Deadjectival noun
d	Adverb	Bg	Morpheme used in noun-modifier
e	Interjection	Dg	Morpheme used in adverb
f	Localizer	Mg	Morpheme used in number
g	Morpheme	Ng	Morpheme used in noun
h	Head/Prefix	nt	Organization name
i	Idiom	nx	Foreign character (alphabet)
j	Abbreviation	nz	Other proper names
k	Tail/Suffix	Qg	Morpheme used in measure word
l	Collocation	Rg	Morpheme used in pronoun
m	Number	r	Pronoun/determiner
n	Noun	Tg	Morpheme used in temporal noun
nr	Person name	Ug	Morpheme used in particle
ns	Place name	Vg	Morpheme used in verb
o	Onomatopoeia	vd	Deverbal adjective
p	Preposition	w	Punctuation mark
q	Measure word	x	Non-morpheme character
s	Place noun	vn	Deverbal noun
u	Particle	t	Temporal noun
v	Verb	y	Modal/sentence-final particle
Yg	Morpheme used in modal/sentence-final particl e		
z	Stative adjective and adverb		

中国語茶筌で処理した中国語文章は、図 3.1 のように、「“ 単語 ” TAB ” 品詞情報 ” 改行 ”」というフォーマットになっている。

図 3.1: 中国語茶筌による解析の例-中国語

一方、日本語の文章をそのまま中国語茶筌で処理した場合、図 3.2 のように、「一文字で区切られた形態素」が殆どである上、「“ g ”に分類された形態素」の割合がかなり目立つ。

また、同じように、英語の文章も中国語茶筌で処理したところ、図 3.3 のようになった。

“ 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。“ g ”という品詞は、他の品詞種類に属しないもっとも小さな“ 単語の素 ”であるため、

図 3.2: 中国語茶筌による解析の例-日本語

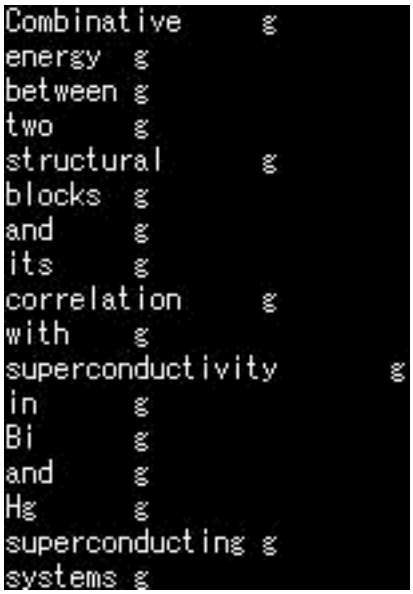

図 3.3: 中国語茶筌による解析の例-英語

“ 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なる。また、連続した“ g ”に分類された複数の形態素が未知語の元となる。

多くの文章を解析し、明らかになったのは、以下の 2 点である。意味のある中国語文章を中国語茶筌に渡せば、しっかり品詞付けし単語ごとに区切ってくれるが、もし中国語でない文章となれば、一文字ずつ区切ってしまうことが殆どである。また、辞書に存在しない単語は、文の前後関係によっては“ g ”に分類されることが多い。

その二つの特性を、中国語ページ抽出のポイントにした。

第4章　抽出までのプロセス

この章では、形態素解析の対象となる文章の元データ（HTML ファイル）の収集から、適切なフォーマットへのコーディングと最終の文章解析まで紹介する。

ステップは以下のようである。

1. ページ収集
2. 文字コードの統一
3. XML ファイルに修正
4. XML タグとダイジェストテキストの分離
5. テキスト解析

4.1　ページ収集

この研究に使用する三種類のデータベース、「北京大学 25,421 pages」及び「早稲田大学 88,634 pages」は上田研究室開発の WWW 全文検索システム Verno の収集ロボット Iron33 を利用した、また、「JP ドメイン 10,166,170 pages」は、フランス製の WEB 収集ロボット larbin を使用した。

北京大学と早稲田大学に関しては、どちらもドメイン内全ての HTML ページを収集した訳ではなく、収集ページ数の増加速度が著しく落ちてしまった時点で収集プログラムを止めることにした。収集時間は、北京大学において二日、早稲田大学において三日だった。上記のページ数は参考までに考えていただきたい。

また、JP ドメイン内で収集したページ数に関しては、ハードディスクの容量や後の解析処理時間を考えた上、1000 万ページとした。

集めたページの総サイズは、北京大学と早稲田大学の数百 MB に対して、JP ドメインは 100GB 程度となった。

4.2　文字コードの統一

多種類の言語が混在しているページを一斉に処理を仕掛ける際に、文字コードの違いが問題になってくる。デフォルトの入出力では、形態素解析後に文字化

けになってしまうページも多くみられたため、文字コードを適切に変える必要があった。

また、本研究に使用している中国語茶筌は、開発元では GB2312 encode でのみ検証が取れているが、事実上、EUC encode でも正常に解析できることは、本研究でたくさんの中国語文章データを解析し一つ一つ確認したことによって検証できた。

したがって、ここでは、nkf を利用し EUC encode に統一することにした。

4.3 XML ファイルに修正

ページ製作者のミスで、HTML にタグの記述違いとか多く見られる。終了タグがないなどといった、よくある間違いが、処理する際に致命的な原因（タグ内文字も形態素の対象となるケースがよくみられるが、最悪の場合、プロセスがエラーを出して終了してしまうこともある）となってしまう。解決するには HTML タグ修正ユーティリティ tidy を使用し間違いを修正した上、XML パーサーを利用し、互換性の高い XML ファイルにした。

例：HTML

```
...
<body>
<a href="./index.html">トップページへ戻る</a><br>
</body>
...
```

例：XML

```
<?xml version="1.0" encoding="euc-jp"?>
...
<body>
<a href="./index.html">トップページへ戻る</a><br />
</body>
...
```

4.4 XML タグとダイジェストテキストの分離

XML（または HTML）ファイルからタグを取り除いた部分を「ダイジェストテキスト」とする。

XML タグが文章に残っていては、茶筌では対応できないため、自然言語処理が行えない。そのため、形態素解析を行う前に、XML タグとダイジェストテキスト部分を分離する必要がある。そして、タグを取り除いた状態のダイジェストテキストファイルは、形態素解析の元となる。

上記の例で言うと、

例：プレインテキスト

トップページへ戻る
...

となる。

JP ドメインから収集してきた 1000 万ページを処理したところ、12 時間程度かかった。

4.5 テキスト解析

これまでの作業で得たデータは、HTML タグが混在しない素のテキストファイルである。中国語茶筌を使用し、三種類のデータベースをそれぞれ解析し、データを取った。

中国語ページ抽出の判別データとして、当初の予定では、未知語抽出した上、その数の多少で判別をすることにしていたが、「ChaSen 処理」→「YamCha 処理」→「未知語抽出処理」というハードな作業が時間をかけ過ぎるので、ChaSen 処理のステップから得たデータを分析し、中国語ページ判別のアルゴリズムを考えた。因みに、未知語抽出にかかる時間は 2 万ページで凡そ 3 時間であるのに対して、形態素解析だけならば、20 分もかからない。また、JP ドメインの 1000 万ページを形態素解析するのに、三日弱かかった。

それぞれの実行時間は、表 4.1 に示す。

また、作業時にいくつかのプロセスが同時に動作していたため、上記の時間データはあくまで参考程度にしていただきたい。

表 4.1: 実行時間データ

項目	時間
ページ収集	
北京大学	2 日
早稲田大学	3 日
JP ドメイン	5 日
タグとダイジェストテキストの分離	
北京大学	2 分
早稲田大学	8 分
JP ドメイン	12 時間
形態素解析	
北京大学	20 分
早稲田大学	54 分
JP ドメイン	58 時間 16 分
抽出	
北京大学	36 秒
早稲田大学	2 分 38 秒
JP ドメイン	5 時間 58 分

第5章　中国語ページの抽出

この章では、北京大学の 25,421 ページと早稲田大学の 88,634 ページを徹底的に比較し、考案した中国語文章抽出アルゴリズムについて紹介する。

5.1　北京大学 pku.edu.cn での試み

大半は中国語文章だと思われる北京大学から集めた 25,421 ページを中国語茶筌で処理したところ、およそ 20 分かかった。

5.1.1　一文字で区切られた形態素

もし文章が中国語でない言語で書かれているならば、中国語茶筌の辞書で文脈をチェックすることができないため、形態素が一文字一文字で区切られてしまう。

まず、その一文字で区切られた形態素は文章形態素総数のどれぐらいの割合を占めているかを見てみよう。ページごとで計算し、表 5.1 のように割合別に分けられた。

形態素総数が 0 であるページは、中国語ではないことが明らかなので、まず最初に分離した。ここでは、表記を「alphabet」とする。

形態素総数が 0 であるページは、約 24% を占めているが、それは、アルファベットと記号で始まる形態素をカウントしないというルールから発生したもので、何も書かれていないページかアルファベット（英語など）のページである可能性が高いと考えられる。

それとは別に、一文字で区切られた形態素が文章形態素総数は「0%」であるページ、「0 <= 割合 < 10%」〜「90 <= 割合 < 100%」の 10 種類のページと「100%」であるページ、という具合に、全部で 13 種類のブロックに分けて、図 5.1 のようなグラフが得られた。

勿論、中国語であっても、一文字である単語も多く存在するので、このグラフからは、10〜70% に集中している傾向があると分かった。しかし、文章に出現した単語は全部辞書に載っている可能性もあるため、0〜10% のブロックにあるページも、中国語であることが多いいので、抽出の対象とする。

また、「割合＝ 100%」であるページが少し目立つが、元々形態素総数が極めて少ないページや、意味のない漢字がただ並べられたページなどが考えられる。簡

表 5.1: 一文字で区切られた形態素の割合表-北京大学

ブロック	ページ数
alphabet	6091
percent=0%	294
$0 <= per < 10\%$	139
$10 <= per < 20\%$	847
$20 <= per < 30\%$	2485
$30 <= per < 40\%$	4146
$40 <= per < 50\%$	4523
$50 <= per < 60\%$	3053
$60 <= per < 70\%$	1034
$70 <= per < 80\%$	465
$80 <= per < 90\%$	678
$90 <= per < 100\%$	338
$percent = 100\%$	1328

(北京大学 25,421 ページ)

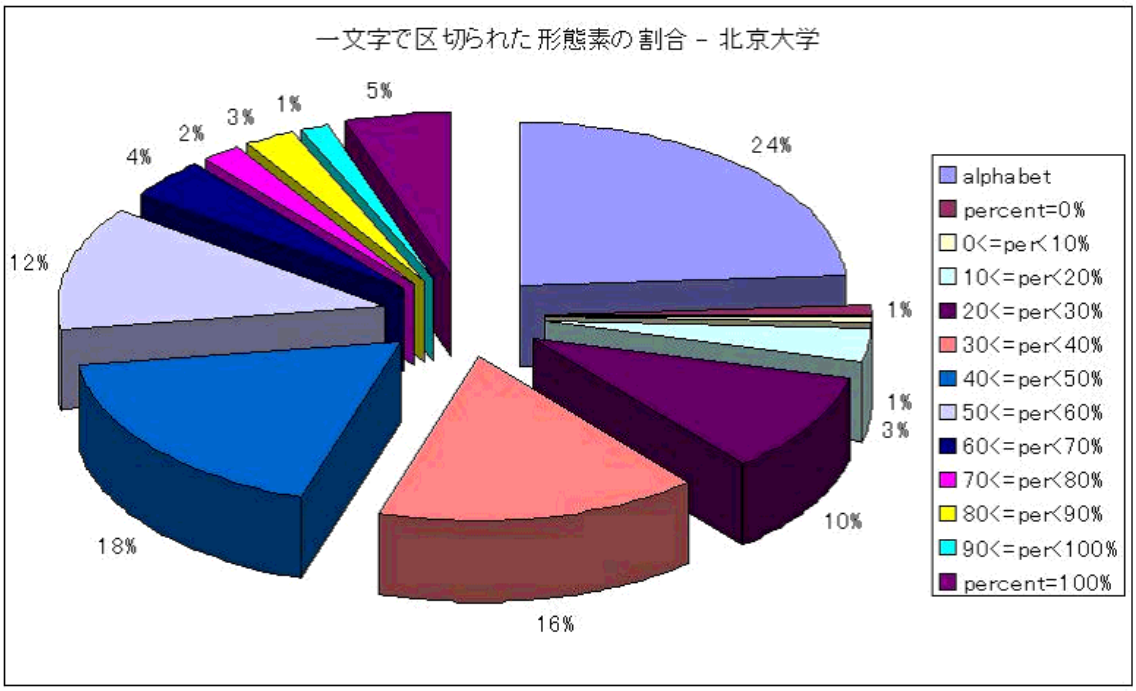

図 5.1: 一文字で区切られた形態素の割合図-北京大学

単な例を挙げると、形態素総数が 1 の場合、「$\frac{\text{一文字で区切られた形態素の数}}{\text{形態素総数}} = 100\%$」であることは言うまでもない。

5.1.2 “ g ”と分類された形態素

“ g ”に分類された形態素が文章形態素総数のどれぐらいの割合を占めているかを見てみる。表 5.2

表 5.2:“ g ”に分類された形態素の割合表-北京大学

ブロック	ページ数
alphabet	6091
percent=0%	8941
$0 <= per < 10\%$	7632
$10 <= per < 20\%$	783
$20 <= per < 30\%$	369
$30 <= per < 40\%$	126
$40 <= per < 50\%$	95
$50 <= per < 60\%$	112
$60 <= per < 70\%$	64
$70 <= per < 80\%$	120
$80 <= per < 90\%$	66
$90 <= per < 100\%$	45
$percent = 100\%$	977

(北京大学 25,421 ページ)

それを円グラフにすると、図 5.2 のようになる。

このグラフではっきり分かるように、アルファベットのページを除いて、“ g ”に分類された形態素が文章形態素総数の割合が「0〜20%」殆どである。特に、0〜10%の割合に集中している。

これは、「中国語の文章ならば、“ g ”に分類された形態素の割合が低い」ということだと推測できる。新語・造語など辞書に載っていない未知語が文章の一〜二割ほどあるということである。

事実上、連続した“ g ”に分類された複数の形態素が一つの未知語として認識されることが殆どなので、未知語識別を行えば、文章の一割に綺麗に収まるというデータも出ている。

また、このグラフにも、「割合=100%」であるページが目立つが、理由は 5.1.1 章で紹介した通りである。5.3 章では詳しく論じる。

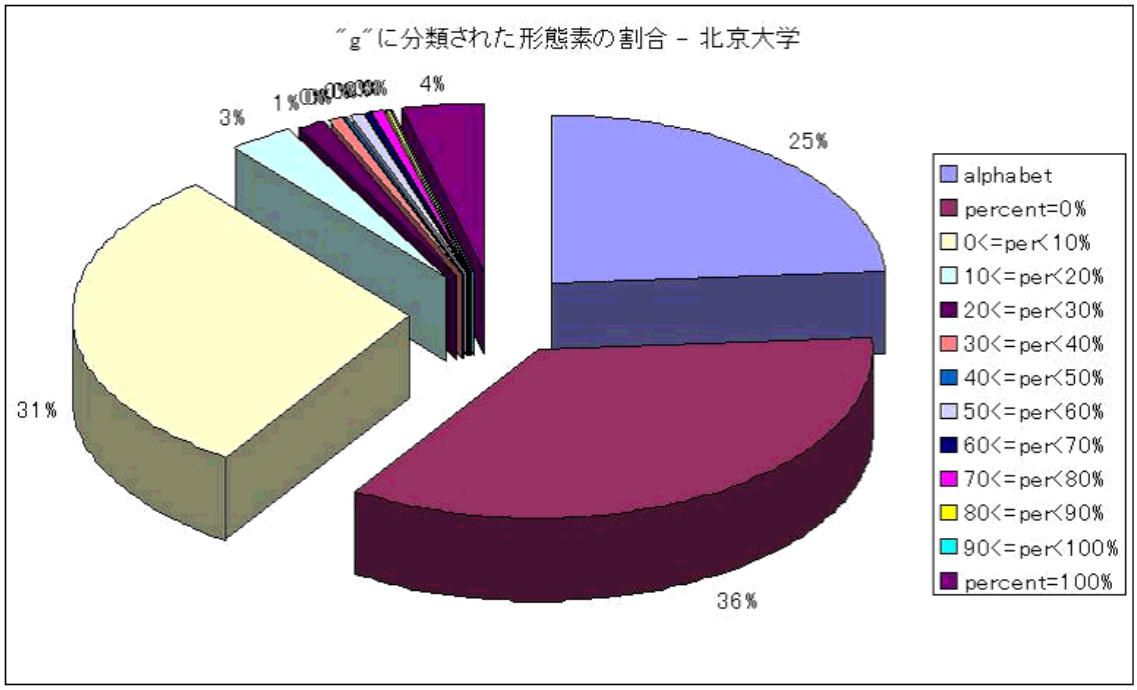

図 5.2:“ g ”に分類された形態素の割合図-北京大学

5.1.3　全体の位置グラフ

図 5.3 では、北京大学全体のページ分布を示した。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 5.1 と図 5.2 で分析したように、「一文字で区切られた形態素の割合が 10%〜70%」かつ「“ g ”に分類された形態素の割合が 20%以下」のブロックに集中していることが分かった。

中国語文章が多いという北京大学ではこのような結果が出たが、中国語文章の割合が低い早稲田大学のページを対象に分析してみる。

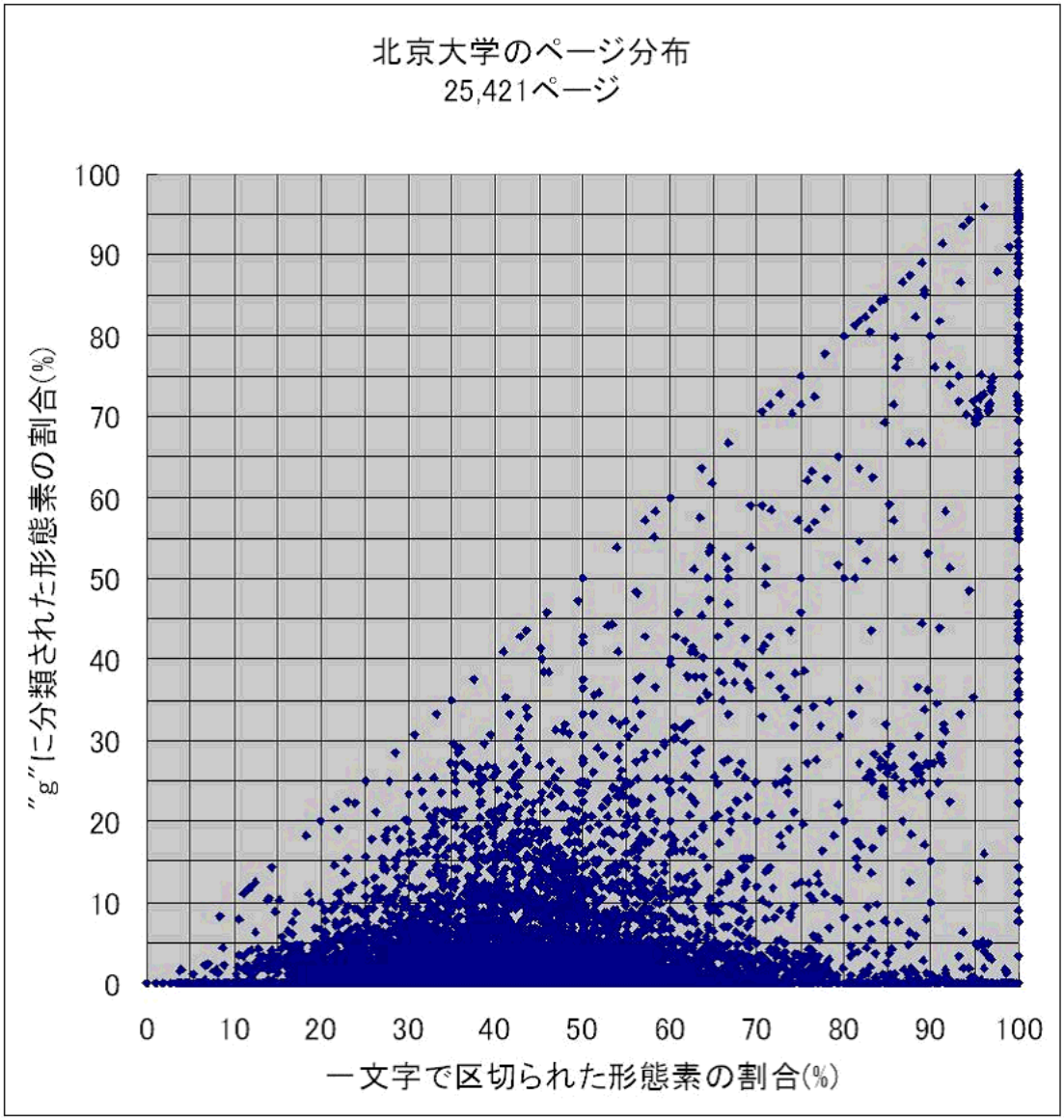

図 5.3: 北京大学のページ分布図

5.2 早稲田大学 waseda.jp/waseda.ac.jp での試み

北京大学で処理した時と同じ手法で、早稲田大学の 88,634 ページを中国語茶筌に処理したところ、54 分かかった。

5.2.1 一文字で区切られた形態素

アルファベットのページを除き、殆ど 90〜100%に集中していることは表 5.3 で明らかになった。

表 5.3: 一文字で区切られた形態素の割合表-早稲田大学

ブロック	ページ数
alphabet	10155
percent=0%	11
$0 <= per < 10\%$	0
$10 <= per < 20\%$	3
$20 <= per < 30\%$	23
$30 <= per < 40\%$	46
$40 <= per < 50\%$	24
$50 <= per < 60\%$	15
$60 <= per < 70\%$	14
$70 <= per < 80\%$	76
$80 <= per < 90\%$	173
$90 <= per < 100\%$	24593
$percent = 100\%$	53501

(早稲田大学 88,634 ページ)

特に、一文字で区切られた形態素の割合が 100%であるページは、53301 もあり、全体の 6 割以上を占めている。それを円グラフで表現したのは図 5.4 である。

中国語でない文章が出現した場合、一文字ずつ区切られると考えられるが、実は、文章に漢字を使っている日本語でも、中国語茶筌で処理する際に、漢字で構成される単語が中国語茶筌の辞書に存在していれば、単語単位で区切られることがあるため、一文字で区切られる形態素の割合が 100%ではない日本語文章も存在することになった。それは、「90 <= 割合 < 100%」のブロックにもページ数が多い理由である。

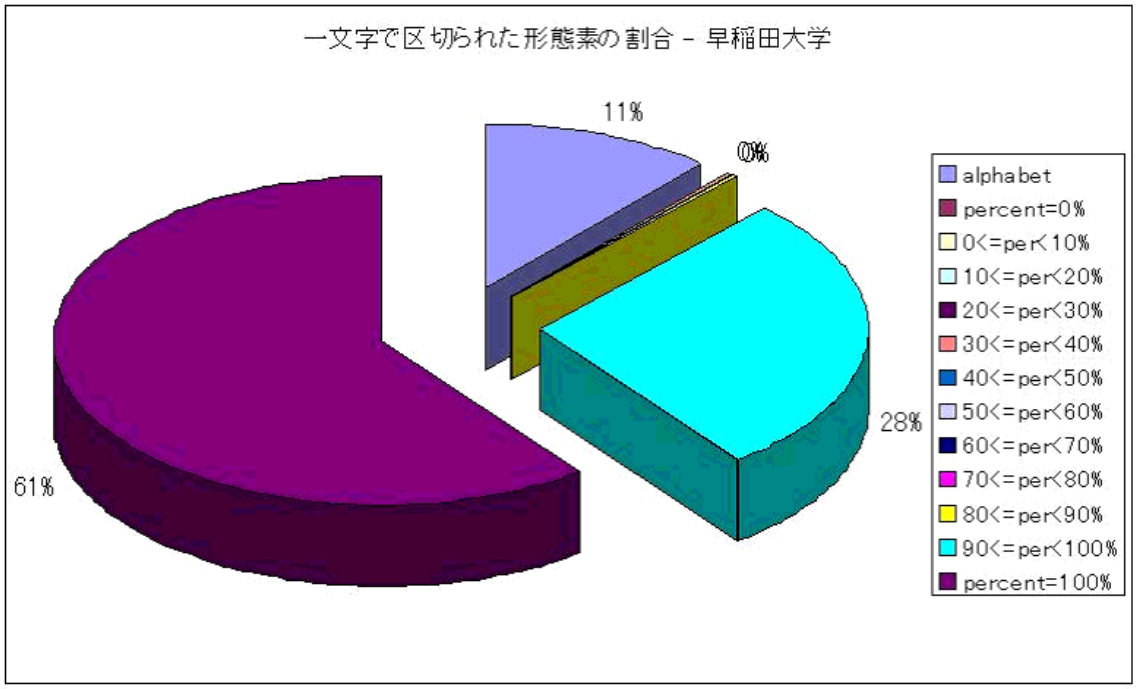

図 5.4: 一文字で区切られた形態素の割合図-早稲田大学

5.2.2 “ g ”と分類された形態素

それを円グラフにすると、図 5.5 のようになる。

「一文字で区切られた形態素の割合」が特定のブロックに集中しているのに対して、「“ g ”に分類された形態素の割合」が分散していることが分かった。

5.2.3 全体の位置グラフ

図 5.6 では、中国語ではないテキストが X 軸の右側に集まっている。つまり、アルファベットで書かれたページを除き、中国語でない文章は殆ど、一文字で区切られてしまうことになる。

表 5.4:“ g ”に分類された形態素の割合表-早稲田大学

ブロック	ページ数
alphabet	10155
percent=0%	1031
$0 <= per < 10\%$	876
$10 <= per < 20\%$	3617
$20 <= per < 30\%$	7590
$30 <= per < 40\%$	7981
$40 <= per < 50\%$	12292
$50 <= per < 60\%$	14376
$60 <= per < 70\%$	13423
$70 <= per < 80\%$	12042
$80 <= per < 90\%$	3372
$90 <= per < 100\%$	517
$percent = 100\%$	2262

(早稲田大学 88,634 ページ)

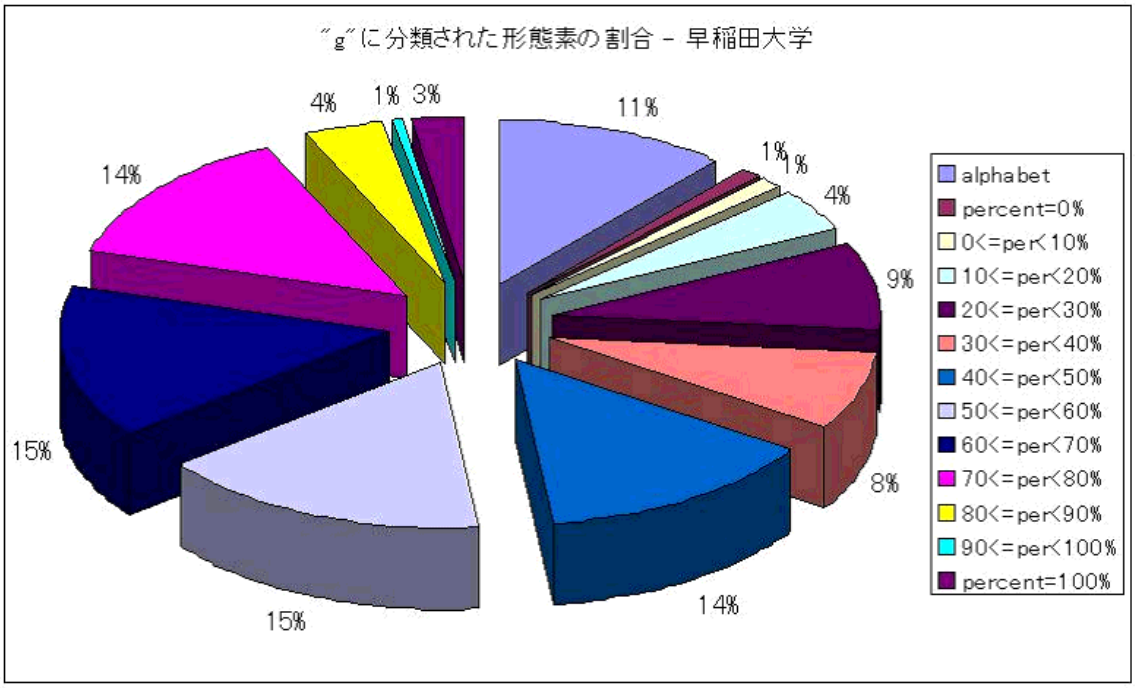

図 5.5:“ g ”に分類された形態素の割合図-早稲田大学

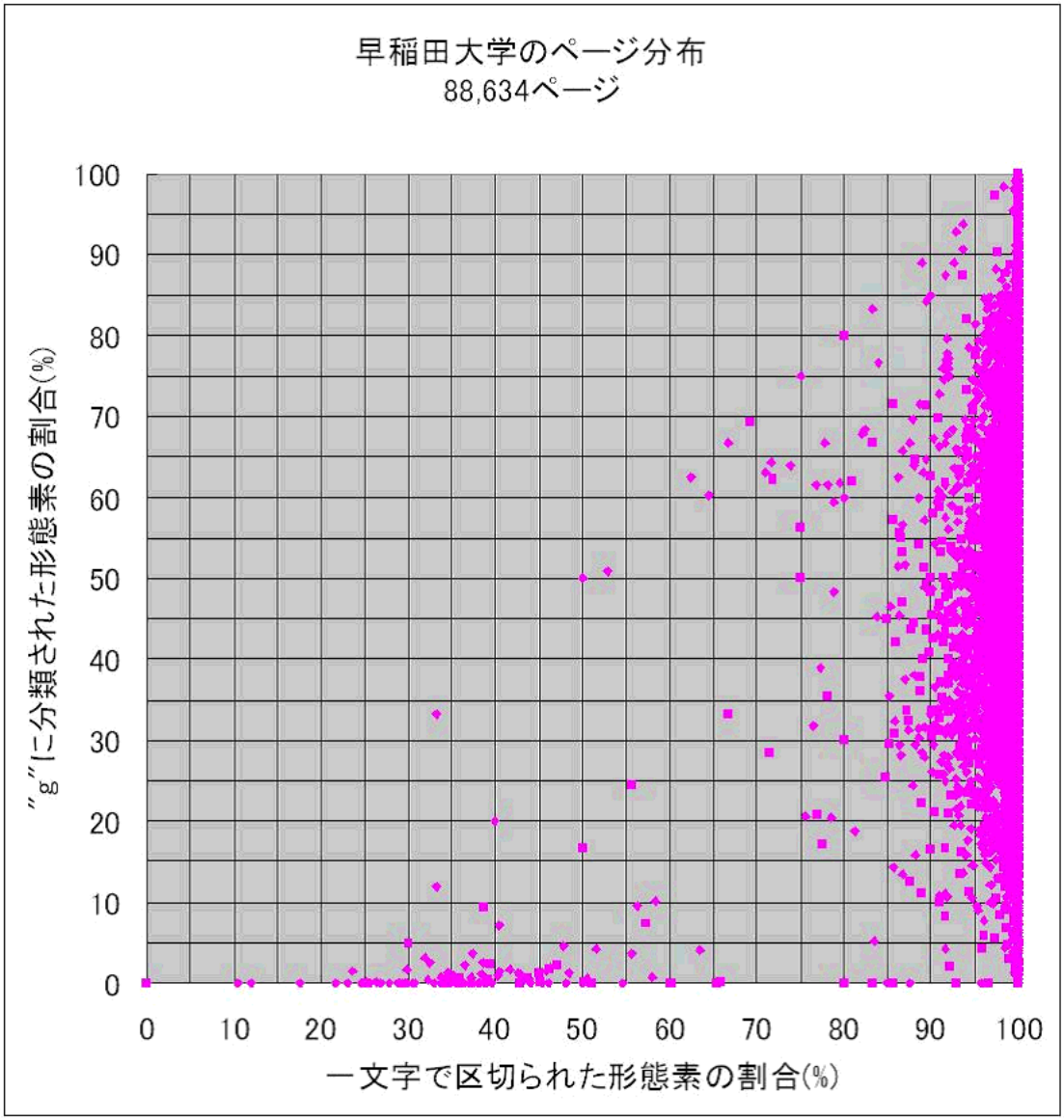

図 5.6: 早稲田大学のページ分布図

5.3 形態素総数

5.1.1 章で述べたように、一文字で区切られた形態素の割合が 100%であるのは、

A ページは中国語漢字以外の文字で構成されている

B ページは中国語漢字で構成されているが、文章として意味がない

C ページは極めて少ない形態素で構成されている

などのケースが考えられる。ケース A と B は簡単に理解できるが、ケース C について分析しよう。

まず、形態素総数のカウントは、アルファベットと記号などを取り除いた上で行ったため、形態素総数＝文章全体単語数という認識が正しくない。例えば、英語のページに何らかの理由で中国語（もしくは日本語漢字）を少しだけ載せられている、ということも考えられる。

それでも、ページを閲覧するユーザにとって、参考価値のある中国語の文ならば、中国語ページとして抽出する条件に満たしているが、その中国語の文があまり短ければ、参考価値がないとする。

実は、形態素総数が少なければ、一つの形態素の比重が重くなるので、“ g ”に分類された割合が 0% と 100% の数にも大きく影響する。例えば、日本語で「日本」という単語が書かれたページが存在するとする。勿論中国語茶筌の辞書にも同じ単語があるので、一つの単語として区切られる上、“ g ”ではない品詞付けに分類される場合もありえる。逆に、中国語の一文字しか書かれていないページがあるとすれば、「一文字で区切られた形態素」の割合が 100% になり、「“ g ”に分類された形態素」の割合も 100%になることも十分考えられる。従って、形態素総数が少なければ、茶筌で処理した後、「一文字で区切られた形態素の割合が 0%」もしくは「“ g ”に分類された形態素の割合が 0%」であっても、そのページが中国語で書かれていることかどうかは判断できない。

上述のことから、ユーザのページ参考価値を基準に、抽出のルール上、形態素総数が少ないページをフィルターする必要があると考えられる。

北京大学のページでは、形態素総数の少ないページがカウントされないとすれば、「一文字で区切られた形態素」と「“ g ”に分類された形態素」の割合がそれぞれ「0%」と「100%」であるページ数はどれぐらい変化するかを図 5.7 で示している。

X 軸は形態素総数を制限する値であり、1 から 10 まで十回計っている。Y 軸はページ数の変化を示している。形態素総数が t の場合、(t-1) の時との差を表している。各パラメーターの紹介は下記の通りである。「ページ数」とは、形態素総数の下限が t である時、(t-1) より多くカウントされなくなったページ総数。「一 per」と「“ g ”per」とは、それぞれ、「一文字で区切られた形態素の割合」と「“ g ”に分類された形態素の割合」の略であり、形態素総数が t である時、(t-1)

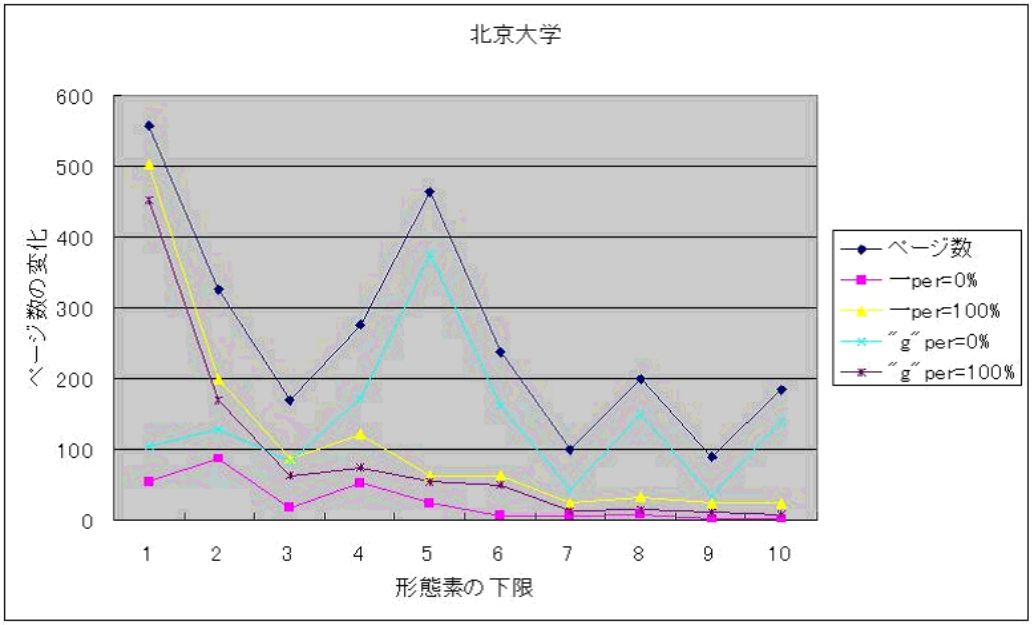

図 5.7: 形態素総数とページ数の変化

より少なくなったページ数を表している。つまり、「ページ数」＝「一 per=0%」＋「一 per=100%」＝「“ g ”per=0%」＋「“ g ”=100%」ということになる。

このグラフで明らかになったのは、形態素総数が「6」になった時点で、値を増やしてもパラメーターに激しい変化がなくなった。ということで、中国語ページを抽出することにあたって、形態素総数が「6」以下ならば、価値のないページとする。

5.4 JP ドメインのページ分布

これまでに北京大学と早稲田大学において、グラフ等で分析して来た。JP ドメインにあるページも同じような手法でデータ（表 5.5）を取ったところ、早稲田大学と似たグラフができた。図 5.8 5.9

また、JP ドメインのページ分布図は図 5.10 のようである。

分散しているブロックも存在するが、全体的には境界線がはっきり見えていることが分かった。

5.5 ルール設定

これまで導き出したルールを合わせると、このようになる。より正確な判別として、中国語テキストとは、「形態素総数は 6 以上ある」かつ「“ g ”に分類され

表 5.5: 「一文字」と「“ g ”」割合別表-JP ドメイン

ブロック	「一文字」ページ数	「“ g ”」ページ数
alphabet	1613425	1613425
percent=0%	1317	99017
$0 <= per < 10\%$	19	58171
$10 <= per < 20\%$	137	164191
$20 <= per < 30\%$	458	373558
$30 <= per < 40\%$	1887	751439
$40 <= per < 50\%$	3961	1379961
$50 <= per < 60\%$	4398	1821786
$60 <= per < 70\%$	4821	1804108
$70 <= per < 80\%$	10290	1268438
$80 <= per < 90\%$	56816	439425
$90 <= per < 100\%$	3994103	81218
$percent = 100\%$	4468545	305440

(JP ドメイン 10,160,177 ページ)

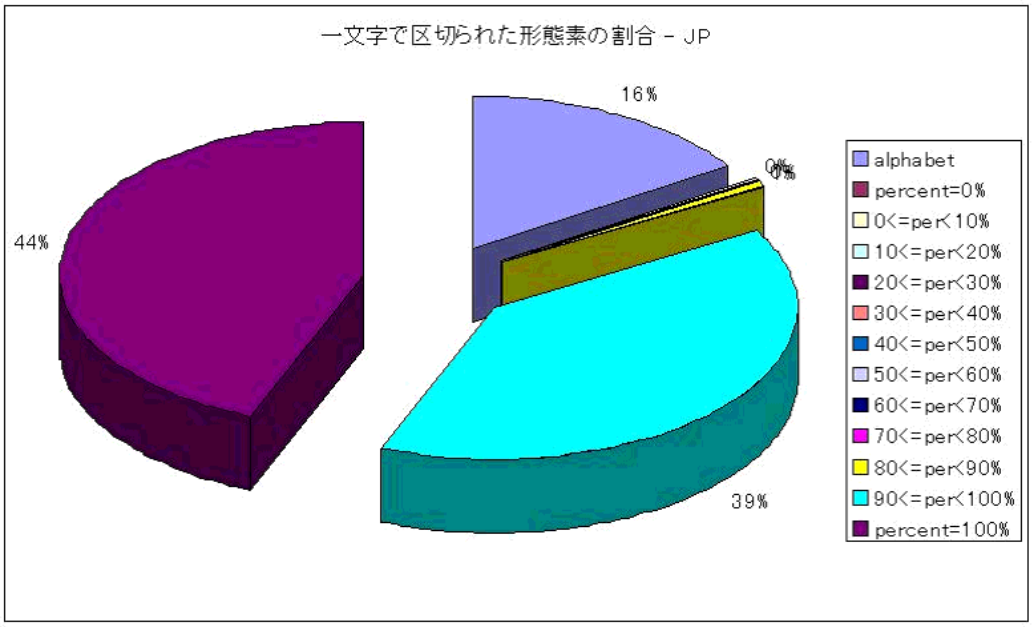

図 5.8: 一文字で区切られた形態素の割合図-JP ドメイン

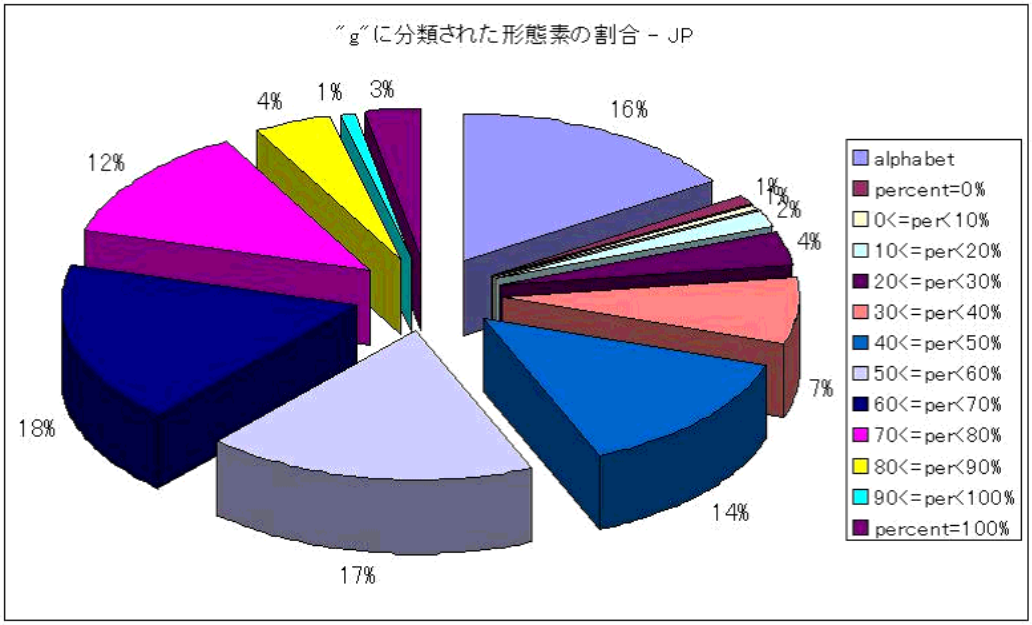

図 5.9:“ g ”に分類された形態素の割合図-JP ドメイン

た形態素の割合は 20% 以下である」なおかつ「一文字で区切られた形態素の割合は 70% 以下である」の条件を満たすものである。

5.6 ルール適用の結果

判別ルールを適用したところ、表 5.6 のデータが得られた。

表 5.6: 中国語ページ抽出の結果

判定範囲	中国語ページ数	全体のパーセンテージ
北京大学	15455	60.79619%
早稲田大学	110	0.12411%
JP ドメイン	10110	0.09951%

(最終結果)

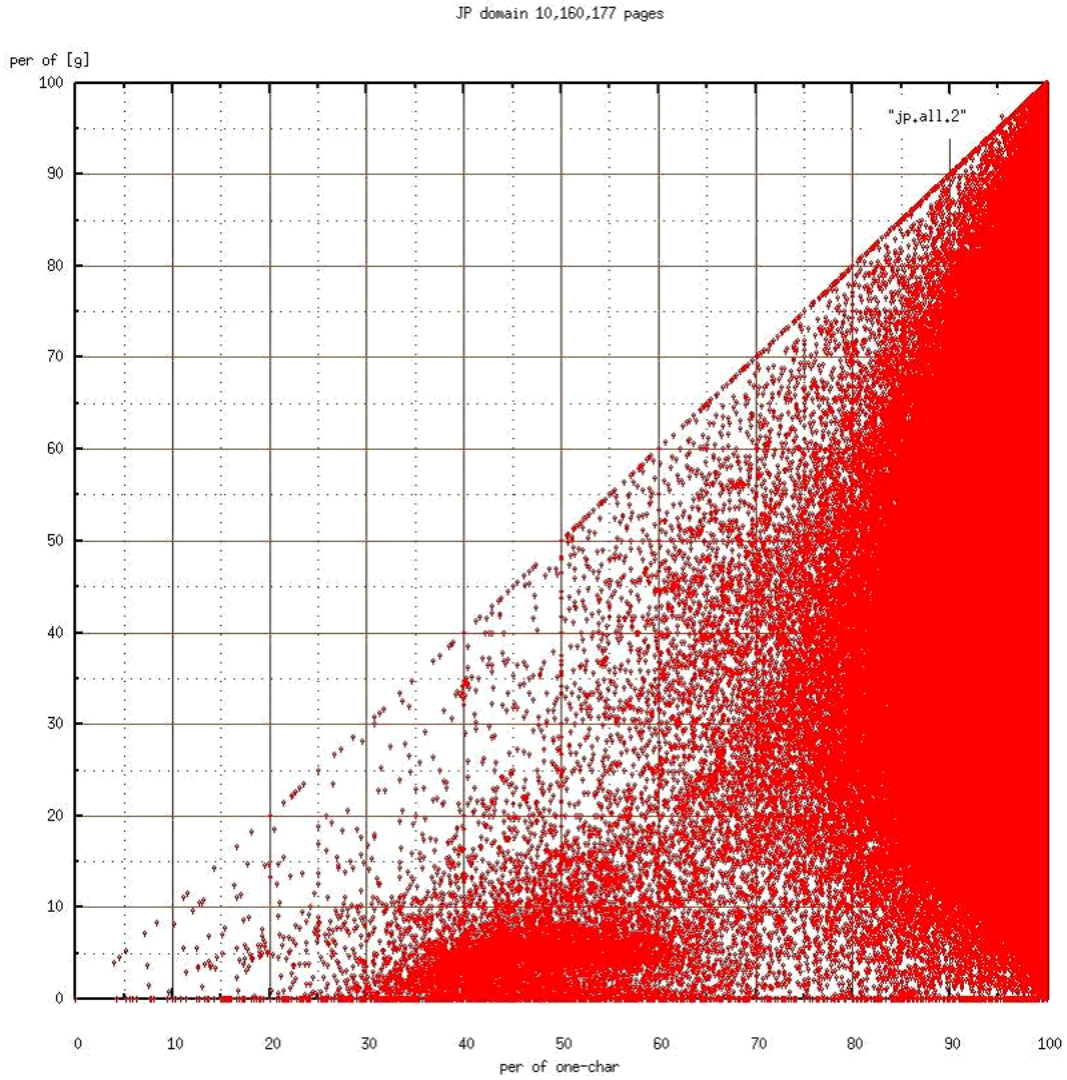

図 5.10: JP ドメイン 1000 万ページ分布図

第6章　考察

まず、第5章で得た三つのページ分布図の共通点に注目したい。全てのポイントは(0,0)と(100,100)を結ぶ斜線の右下にある。それは、つまり、

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

ということになる。

対象の文字が茶筌の辞書に存在しなければ、一文字として区切られる上に、“ g ”に分類されるが、辞書に存在していれば、同じく一文字として区切られるが、“ g ”には分類されず他の品詞情報を付き加えられる。そのため、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。第5章で紹介したように、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。

6.1　北京大学のページ分布図

図5.3のページ分布状況から、下のブロックから分散しているように見えるが、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、表5.1に示したように、(0,y)に付着しているポイントが5%を占めているゆえ、残りのページは凡そ1割である。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、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多く存在している。第5.3章でも紹介したように、形態素総数が少なく、意味の持った文になっていない中国語ファイルを、中国語茶筌が適切な結果を出さなくなることがある。また、フレームページセットなど、キーワードのいくつしか書かれていないページに関しても、同じ現象が起こりうる。簡単に言えば、文章としての意味が持たない文字列を組み合わせたページを中国語茶筌で適切な処理ができない場合があるため、グラフの中で発散するように見えると考えられる。

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、判断し難いページをカウントしないことにしているが、将来の課題としては、きちんと解決する方法を考えるべきである。

6.2 早稲田大学のページ分布図

図 5.6 に示したように、早稲田大学は北京大学の 3 倍以上のページ数を持つが、予測通り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%」というパタンは 61% も占めている。漢字も使われているだけに、Y 軸の値が集中しない。少しに「一文字で区切られた形態素の割合」は 90%〜100% のブロックにも存在しているが、日本語ページならが、X 軸の値がそれより小さいものはほとんどないと考えられる。

また、北京大学と早稲田大学の分布図は、Microsoft Office Excel で作成しているが、仕様上 Excel が 65,536 行しか読み込めたいため、早稲田大学の分布図を作成するさいに、二回分けて作成し合成したものである。

6.3 JP ドメインのページ分布図

図 5.10 は早稲田大学のページ分布のパタンに似ているが、異なるところもある。日本語やアルファベット以外にも中国語でない言語がたくさん存在するので、「一文字で区切られた形態素の割合」は 70%〜90% であるページ見られるのも珍しくない。また、北京大学の分布と同じように、「中国語ブロック」以外に分散した中国語ページも少し存在する。1000 万ページを超えると、様々な思い付かないパタンも潜んでいるため、完璧に説明できないページもあると考えられる。将来の課題としては、特殊パタンの識別も考えなければならない。

また、Microsoft Office Excel では、読み込み行数の制限があったため、図 5.10 は gnuplot によって作成したものである。また、Windows 上から 10,160,177 行のデータを読み込もうとしたところ、一時間以上待っていても応答がなかったため、表 1.1 に示した cheddar でグラフを出力した。「データ読み込み」と「描画」にかかった合計時間は 20 分程度であった。

第7章　まとめと今度の課題

7.1　まとめ

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 10,160,177 個の HTML ファイルを対象に、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。

しかしながら、その同時に研究の過程で様々な問題点や今後解決すべき点などが明らかになってきた。

7.2　問題点と改善策

7.2.1　速度

1000 万ページを対象にした ChaSen の実行時間は三日弱であった。「億」でも数え切れない JP ドメイン全体を持ってきたら、確実に一ヶ月以上かかってしまう。本研究では、ChaSen の処理時間はネックとなった。

ChaSen の処理時間を短縮するために、自然言語処理の分野をもっと深く研究する必要がある。

7.2.2　ファイルサイズ

本研究では、100GB ほどの元 HTML データを対象に処理・分析を行った。元データを削除しない方法では、形態素解析した時点で、データ総サイズが 230GB くらいに膨らんだ。その計算ならば、800GB のストレージを持つマシンでは、3000 万ページが限界である。

有限なストレージ容量でより多くのデータを処理するには、転置ファイルの作成・安全な元データの削除など、処理時間の関連も考えつつ改善案を練るべきである。

7.2.3 言語と文字コード

本研究では、「中国語」「日本語」「alphabet」という三種類の言語を考慮した上解析を行ったが、それ以外のケースはあまり想定していない。また、文字コードを EUC に固定するなど、エンコード特有の情報を保持していない。それに、多国語言語が混在しているページも存在しているが、それへの考慮も不十分である。

それらを解決するのに、文字コードの研究から強化すべきである。

7.2.4 ルール設定

本研究では、判別ルールを設定する元情報として、北京大学と早稲田大学の WEB ページを分析しデータを取った。ページ編集者の記述力の信頼度など、少し曖昧なところがあるが、最終抽出対象も WEB ページであったので、基本的に適切とは言える。

しかしながら、完全に信頼のできる新聞記事などを元に、基準データを取る必要がある。また、論文の中で提示した条件以外にも、判定のパラメーターとなるデータを考慮すべきである。

7.3 応用と今後の課題

7.3.1 日本に関する情報が書かれた中国語ページの抽出

本研究では、実現するまでに至らなかったが、延長研究として考えている。

7.3.2 繁体字中国語への対応

本研究では、「簡体字中国語」をベースとした処理であり、「繁体字」を「多国言語」と見なした。将来的にはケース分けで考えたい。

7.3.3 更なる大規模な処理への考慮

JP ドメイン内にある 1000 万ページを対象としたが、5000 万・1 億ページの処理なども視野に入れ、処理速度やファイル総サイズの難関をクリアし実現したいと考えている。

7.3.4 未知語データベース

本研究で手掛けたことでもあるが、形態素処理をした上に、未知語を抽出しデータベースを構築すれば、更なる研究や統計ができただろう。処理時間がネックになり、割愛してしまったが、今後の課題として引き続き研究したい。

7.3.5 サーチエンジンへの応用

また、本研究で得た成果を元にすれば、サーチエンジンへのユニークな応用も十分に考えられる。「JP ドメイン内にある全ての中国語を集めた検索サービス」など、様々な可能性がある。

謝辞

本研究を進めることにあたり、ご指導頂いた上田和紀教授に厚く御礼申し上げます。

研究には最初から最後まで付き合って頂いた検索班の方々に感謝しております。特に、様々な面からサポートして下さった内藤一兵衛氏と牧野知仁氏がいらっしゃらなければ、このような研究成果が得られなかったのでしょう。

また、この研究のきっかけとなった中国語茶筌を開発され、東京での学会の隙間に face to face で色々なアドバイスを頂いた奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座の呉翠玲氏に深い感謝の意を表したく存じます。

最後に、日頃様々なアドバイスを頂いた研究室の方々、私の友人、家族にも感謝しております。

参考文献

[1] Baidu : http://www.baidu.com/

[2] Google : http://www.google.com/

[3] Sohu : http://www.sohu.com/

[4] Sina : http://www.sina.com/

[5] Yahoo : http://www.yahoo.com/

[6] China Internet Network Information Center : http://www.cnnic.cn/

[7] The Web's Richest Source : http://www.nua.ie/

[8] The World Wide Web Consortium : http://www.w3c.org/

[9] HTML TIDY : http://www.w3.org/People/Raggett/tidy/

[10] XML and Scheme : http://okmij.org/ftp/Scheme/xml.html

[11] XSL Transformations : http://www.w3.org/TR/1999/REC-xslt-19991116

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

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

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

[15] Chooi-Ling GOH. Chinese Unknown Word Identification by Combining Statistical Models. 奈良先端科学技術大学院大学情報科学研究科 2003 年度修士論文, 29 August 2003.

[16] 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.

[17] K.J. Chen, C.R. Huang, L.P. Chang, and H.L. Hsu. Sinica corpus: Design methodology for balanced corpora. In PACLIC 11: Language, Information and Computation Selected Papers from the 11th Pacic Asia Conference on Language, Information and Computation -Seoul-, pp. 167-176, December 1996.

[18] 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.

[19] 浅原正幸, 米田隆一, 松田寛, 坪井祐太, 高岡一馬, 松本裕治. 統計的中日形態素解析のための品詞タグつきコーパス管理システム. 2001.

[20] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』version 2.2.9 使用説明書. Technical report, 奈良先端科学技術大学院大学, 2002.

[21] 日野大介, 形態素解析と TF・IDF 重み付けを用いた検索結果用スコアデータベースの構築とその評価. 早稲田大学理工学部情報学科 2001 年度卒業論文, 2002.

[22] 魏小比, WWW 全文検索システム Verno における S 式を用いた半構造化データの処理とその評価. 早稲田大学理工学部情報学科 2002 年度卒業論文, 2003.

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

[24] 日向 俊二, 独習 XML. 翔泳社, 2001.