WEB ページの知的探索・統合・加工
Intelligent Search and Information Extraction of Web Data

廣川佐千男 *
Sachio Hirokawa

1. まえがき

インターネット上のホームページ群は世界際大の知識の書物と呼べる。人類はかってこれだけの知識の素を共有したことはない。しかも我々はそこから膨大な量の情報を瞬時に集めることができるという状況にある。増え続ける Web 空間から効率良く知識を獲得する手法の開発は、現在の情報社会における最も重要な研究テーマといえる。本発表ではその中で特に、同系統情報の収集について我々の研究室で行なっている研究を紹介する。

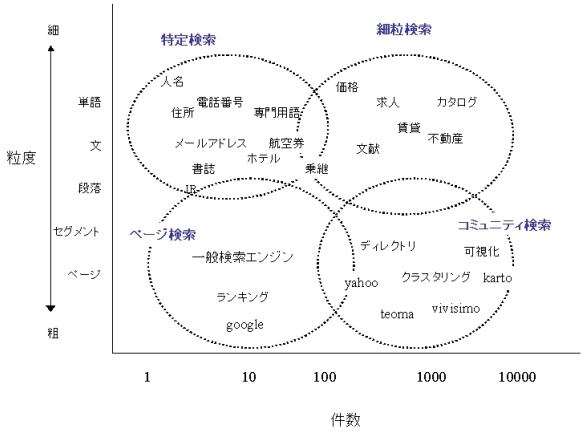


図 1: 同系統情報による細粒検索

検索結果の件数と検索対象の粒度という二つの観点から捉えることにより、従来の Web 検索と同系統情報に基づく新しい検索の方向を比較することができる (図 1) を、我々は提案している。通常の検索エンジンは、キーワードに応じた少数の Web ページを返すことがその第一義的目的なので、本稿ではページ検索と呼び、図 1 では左下の部分に位置付けられる。そこでは得られたページを個別に閲覧しなければならないので、より適切なページを上位に提示するために、ランキングが重要となっている。同様にページを対象としていても、検索結果として単独のページではなく、複数のページ群が期待される場合には、yahoo のように予め分類されたディレクトリ構造として表示する方法や、teoma や vivisimo のように関連するページ群をクラスタリングして表示する方法、あるいはグラフィカルな可視化を用いた kartoo などがあり、コミュニティ検索と呼ぶことができる。図 1 の左上の部分は、細粒度の単一情報を求める検索サービスであり、特定検索と呼ぶことにする。人名、住所、電話番号、メールアドレス、書誌情報、専門用語、企業決算公告データなどがある。これらは個別の DB を持っていて、検索結果として Web 情報を返すわけではないので Web 検索と呼ぶには広すぎるかもしれないが、そのような DB はなんらかの形で Web 情報を収集した結果として構築されたものといえる。図 1 の右上の細粒検索と表した部分では、例えば、関連研究を調査するときに利用する文献検索 siteseer のように、理想的には関連するすべての論文の情報が検索結果として要求される。このような高品質の検索を実現するために必要となるのが、同系統の情報の発見と収集、そして統合のための技術である。本発表では、「多量な同系統情報は高品質である」というヒューリスティックに基づき、我々が行なっている 5 項目の研究を紹介する。

*九州大学情報基盤センター, Computing and Communications Center, Kyushu University

2. 交代数によるパターン発見 [1, 2, 3]

半構造化テキストデータからコンテンツ部分を抽出するプログラムは一般的にラッパーと呼ばれる。増え続ける Web 上のデータ群から必要な情報を発見し、データベースのように活用するためには、ラッパー自動生成の技術が必須とされる。部分文字列の長さと出現頻度を決めると、入力テキストは高頻度部分と低頻度部分に識別でき、それらが交互に現れることになる。部分文字列の長さと出現頻度をパラメータとして、この変化の個数を交代数とよび、「交代数が極小となるパラメータに対する部分文字列がパターンを記述する」というヒューリステックスを導入した。この手法により各種 Web データに対するラッパーが構成できることを示した。図 2 は、高頻度部分を薄く表示した出力結果である。

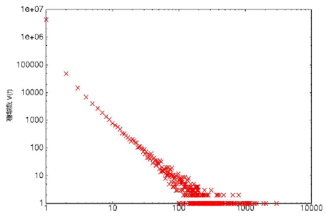
```
<html> <head> <title> Yomiuri On-Line/□□□5250□□□<font size="+2"><b> 小泉首相、今国会への補正予算案提出を否定 </b></font><br> <img src="/g/d.gif" width="1" height="5"><br> <img src="/g/db.gif" width="465" height="1"><br><img src="/g/d.gif" width="1" height="15"><br> <!– photo start –> <!– NO PHOTO –> <!– photo end –> <!– honbun start –> <p> 小泉首相は一日、首相官邸で記者団に対し、景気対策として二〇〇一年度補正予算案を今国会に提出する可能性について、「考えていない」と否定した。 </p> <p> 首相はこれまで、従来の公共事業中心の景気対策には否定的な立場をとっている。また、財政再建に向け、国債発行額を毎年三十兆円以下に抑える考えも示しており、補正予算編成への消極姿勢も、こうした基本方針を反映したものだ。 </p> <p> さらに、政府・与党が緊急経済対策に盛り込んだ、財政出動を伴う可能性のある「銀行保有株式取得機構」にも、首相は「早急につくるのではなく、もう少し専門家に意見を聞き、より充実したものにすべきだ」と、慎重に内容を検討する意向を示している。 </p> (5 月 1 日 21: 19)<br> <!– honbun end –> <div align="right"><img src="/gif6/arw.gif" width="11" height="11">□□□5800□□□ <LAYER SRC="/srcfiles/specials.htm" VISIBILITY=hidden ONLOAD="moveToAbsolute(specials.pageX, specials.pageY); visibility=true;"></LAYER> </body> </html>
```

図 2: 高頻度部分と低頻度部分への分割の例

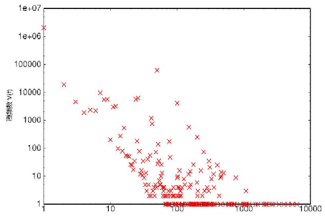
3. 頻度の頻度によるパターン発見 [4]

入力テキスト中に現れる部分文字列について、その出現回数をカウントするという単純な手法で、頻出するパ

ターンの発見が可能であることを、理論的にも実証的にも示した。自然言語文においては、部分文字列の頻度の頻度がベキ分布に従うことがジップの法則として知られている。HTML などの人工的なテキストでは、自然言語の文章以外にタグなどにより構造を示す文字列が含まれる。同系統のデータを表す複数の Web ページ群では、これのパターンを記述する文字列の頻度の頻度については、ベキ分布から乖離する点として現れる。このような乖離点とその頻度を持つ部分文字列を線形時間で発見するアルゴリズムを提案し、繰り返しパターンを高い精度で発見できることを示した。



夏目漱石 こころ



日経記事

図 3: 頻度の頻度に関するベキ分布

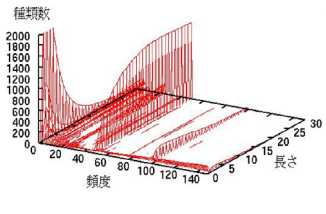


図 4: n グラム長・頻度・頻度の頻度

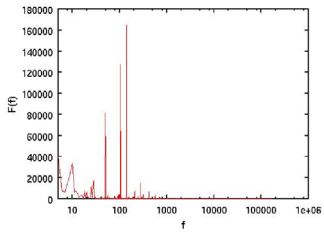


図 5: 複数テンプレートの発見

4. シリーズ型文書、知的探索、メタデータ構築 [5, 6]

リンク構造と構造類似性で特徴付けられる「シリーズ型文書群」という概念を提案し, そのような文書群を効率的に発見収集する Web ロボットの実装を行ない, その収集効率を定量的に評価した. 例えば、大学のシラバス、料理のレシピ、不動産の物件情報など同一サイト上にあり同一テンプレートで記述された Web ページ群は一覧ページとそこからリンクされた個別ページ群からなるシリーズ型文書の典型である。データの属性を表す２～３の特徴的キーワードを与えることで、対象となる Web ページを効率よく収集する。例えばシラバスについての実験では、科目、担当、概要、評価などが特徴的キーワードとして与え、国内の大学ページについてリンクをたどり、150 万ページ収集し、その中の約 3 がシラバスであった。効率的に収集できていることが実証できた。また, シリーズ型文書群の個別ページから共通のテンプレートを抽出することにより、メタデータを自動的に構成する手法を開発した.

一覧ページ

個別ページ

図 6: 「一覧・個別」構造

表 1: フィールド・インスタンス出現頻度

No.	頻度	単語
1	41	シラバス
2	41	電気工学科
3	41	学年
4	41	授業科目名
5	41	科目コード
6	41	担当教官名
7	41	開講期
8	41	単位数
9	41	必・選
10	10	５年
	15	４年
14	28	前期 後期
15	14	１
...	...	...
25	41	授業形式
27	41	成績評価 の方法
29	41	キーワード その他

5. Web 上の表検索 [7, 8]

ある概念についての例を多数集めたいとき、検索エンジンにその概念をキーワードとして与えても、得られるのはそれに関連するページであり、個別に単語を抜きだしまとめ直す作業が必要となる。一方、Web 上には、表形式で記述されたページが多数存在する。それらのページでは、同じ列には内容的に同系統の情報が書かれている。このようなページの表に集めたい具体的な例が２～３含まれていれば、一度に多数の他の例をその列に含まれるデータから得ることができる。Web 上の表形式のページを発見し、表情報を抽出する方式を考案した。その情報を用いて多数の同系統単語知識を収集するシステムを開発した。

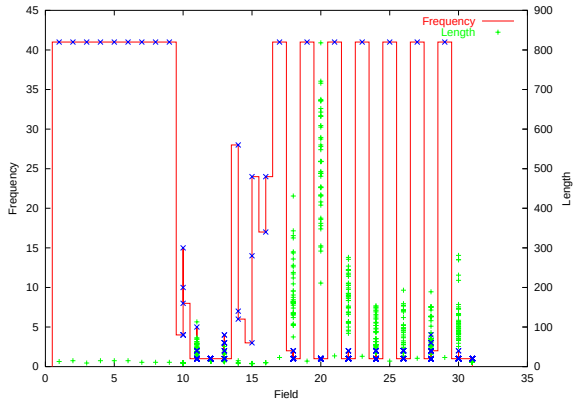


図 7: 出現頻度による属性名・属性値判定

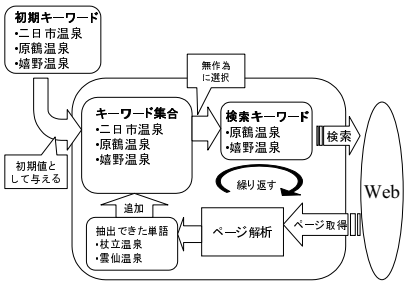


図 8: 同系統単語収集システム

キノコ, プログラミング言語, SMAP のメンバー, 小説作家, インターネットサービスプロバイダ, 大学, JR 九州の駅, 声優, 自動車メーカー, 鞄などのブランド, プロテニスプレーヤー, 漱石の作品, 有名な映画, 国名, 温泉, あるシリーズ小説のサブタイトル, お笑い芸人, 競争馬, あるアーティストの曲, あるゲームのキャラクター, 麻雀の役, プロレスラー, 学会名 (分野関係なし), コンピュータウイルス, 教科名 (小・中学校), 星, 冬のスポーツ, 寿司ネタ, トランジスタの型番, コンピュータ雑誌, テレビアニメ, CPU の種類, 目薬の名前, あるテレビアニメシリーズのサブタイトル, テレショップ, ポーカーの役, 電器店, 北欧神話の神, 三国志の武将, ある漫画の登場人物, あるテレビアニメの登場人物, 通販の会社, 日付, あるゲームのアイテム, あるアニメシリーズに出てくる型番, 検索エンジン

参考文献

[1] Y. Yamada, D. Ikeda, S. Hirokawa, Automatic Wrapper Generation for Multilingual Web Resources, Springer LNCS 2534,332-339, 2002

[2] D. Ikeda, Y. Yamada, S. Hirokawa, Eliminating Useless Parts in Semi-structured Documents using Alternation Counts, Springer LNCS 2226, 113-127, 2001

[3] 山田泰寛, 池田大輔, 廣川佐千男, n-gram 交代数を用いた半構造化データの不要部分削除, 第 144 回自然言語処理研究会, 信学技報 101(190), 53-60, 2001

[4] 池田大輔, 山田泰寛, 廣川佐千男: 部分文字列増幅法による共通パタン発見アルゴリズム, 情報処理学会論文誌「数理モデル化と応用」(採録決定)

[5] 山田信太郎, 松永吉広, 伊東栄典, 廣川佐千男 Web シラバス情報収集エージェントの試作電子情報通信学会論文誌 D1, J86-D1(8),566-574, 2003

[6] S. Hirokawa, E. Itoh, T. Miyahara, Semi-Automatic Construction of Metadata from A Series of Web Documents, Springer LNCS 2903, 942-953, 2003

[7] 野口正人, 廣川佐千男, Web からの同系統単語知識獲得方式, 2003 年情報学シンポジウム講演論文集, pp.21-24, 2003

[8] 野口正人, 廣川佐千男, Web からの同系統単語知識獲得についての実験, 情報処理学会第 65 回全国大会講演予稿集 第 5 分冊 pp.223-226,2003