第 1 回横幹連合コンファレンス
2005 年 11 月 25 日，26 日　JA 長野県ビル

日常生活をより豊かにする Web マイニング

○相良 毅，喜連川 優（東京大学 生産技術研究所）

Web Mining Application for Better Daily Life

* T. Sagara, M. Kitsuregawa (Institute of Industrial Science, the University of Tokyo)

Abstract－ Real world information on the Web can be utilize as a rich, polyphenic information source when they are linked and referred to real world objects, such as persons and locations. In this presentation, a new information service will be introduced which collect and summarize information of shops, especially restaurants to provide various dinner information.
Key Words: Web Mining, Geographical Information Retrieval, Search Engine

1　はじめに

World Wide Web（以下，Web）は今日幅広く日常的に用いられており，さまざまな情報を検索する際の情報源としてなくてはならないものとなりつつある．その一方で，Web上の情報量が爆発的に増加しているために，従来のキーワード検索技術を用いたサーチエンジンでは検索結果の絞り込みが困難になっていることや，その膨大な情報をノイズとして除去するのではなくより効果的に用いる技術が少ないといった問題が研究課題として注目されている[1)]．これらの問題の主な原因の1つとして，Webページに含まれる単語に意味情報が欠落していることが指摘されている．これを解決するため，Webページの著者が意味情報を付与できるようにHTMLを拡張しようというSemantic Webに関する研究が盛んに行われているが[2)]，まだ広く普及するには至っていない．もう一つのアプローチとして，自然言語処理技術を用いてHTMLに自動的に意味情報を付与し，よりきめ細かい検索を可能にするという研究も行われているが[3)4)]，限定されたデータセットに対する検索でも精度が7割程度と実用的なレベルには達していない．

そこで本研究では，日常的にWebで検索される地域のレストランやホテルなどの店舗に関する情報を，高い精度で検索するシステムの開発を行っている．ここで店舗に関する情報とは，営業時間や定休日といった定型の情報だけではなく，ポータルサイトや個人サイト，ブログ（Weblog），電子掲示板などに分散している店舗利用者による評価やコメントを含む，Webに特徴的で価値の高い情報を指す．われわれは，これらの情報を高い精度で収集するため，電話帳（タウンページ）を地域の店舗情報の辞書として利用し，サーチエンジンを用いて多くのサイトに分散するWebページを収集する手法を開発した．また，ある店舗に関する多数のWebページから，特徴的な文を抽出し要約を作成することで，店舗の全体的な情報を簡潔に得られるシステムを開発した．

2　提案手法の詳細

2.1.　処理の流れ

提案手法では，まず一般的な Web クローラによってページを収集し，キーワードインデックスを付与した Web アーカイブを作成する．次に，電話帳の各レコードに記載されている店舗 1 軒ごとにアーカイブから情報を抽出するという処理を繰り返す．各店舗に対する処理の流れは以下の通りである．

(1) Web アーカイブから対象店舗の情報を含む可能性のあるページを抽出する．
(2) HTML 構造の解析により，他店舗の情報やバナー広告・サイドメニューなどを除去し，対象とする店舗に直接関係のある部分のテキストを抽出する．
(3) 複数の Web ページから抽出されたテキストの集合から，索引語と要約テキストを作成する．
(4) 得られた対象店舗に関する URL のリスト・索引語・要約テキストと，電話帳に記載されている住所から計算した経緯度を RDBMS に登録する．

以下，(1)～(3)の各処理について詳細に説明する．

2.2.　Web アーカイブからのページ抽出

まず電話帳に記載されている情報（店舗名・住所・電話番号）から，適切な検索キーワードを作成する．

(1) 店舗名称

電話帳には正式な店舗名称が記載されるため，一般に利用者が店名として意識しない文字列が含まれていることが多い．たとえば「株式会社」「有限会社」などの会社種別や，「東口店」「渋谷駅前店」などの支店名，「炉端居酒屋」「ダイニングキッチン」のような業態などがある．これらの文字列はユーザが作成する Web ページなどには記載されていないことも多いため，記載名称をそのまま用いると検索できない．不要な文字列の除去はヒューリスティックなルールの組み合わせとなるので，詳細は省略する．

(2) 住所表記

検索対象店舗と同じ名前で，非常に有名な別の店舗が存在する場合，有名度によるランク付けが行われている Web アーカイブでは対象としている店舗に関する情報の優先度が低くなってしまう．そこで，住所を店舗名とともに検索キーワードとして用いることで，比較的無名な店舗に関するページを収集す

ることができる．ただし住所の表記には都道府県名の省略などによる揺れがあるため，住所表記のうち市区町村名と大字名の部分を地域名として取り出し，キーワードとして利用する（以下,「地域」と呼ぶ）．たとえば「千代田区大手町」「福岡市中央区天神」「大阪市北区梅田」のようになる．

(3) 電話番号

電話番号は重複のない一意な識別子と考えられるため，強力なキーワードとなる．ただし，Web ページには市外局番が記載されていない場合が多いため，市外局番を除いたものを店舗名と組み合わせて検索キーワードとする（以下，「市内番号」と呼ぶ）．

以上の処理によって得られたキーワード群を用いて次のような条件を満たす Web ページをアーカイブから検索する．住所録の記載情報をそのまま利用する場合よりも広くページを収集できる．

検索条件：（店舗名称）OR（店舗名称 AND 地域）OR（店舗名称 AND 市内番号）を含む

2.3. HTML 構造の解析

前節の検索条件に一致するページには，系列店のリストや商店街加盟店一覧のページなど，1 ページに複数の店舗情報が含まれているものがある．また，バナー広告やリンクメニューなど，店舗の情報以外の文章も含まれている．このようなページに含まれるテキストを，HTML のタグ構造を用いて次のアルゴリズムによってブロックに分割する．

ページのブロック分割

(1) HTML 要素をノードとする木構造を作成する

各ノードは，タグ名（<table>など），属性リスト（{width=”100%”, bgcolor=”#ffffff”}など），タグで囲まれる文字列，および下位のノードのリストを属性として持つ(Fig 1).

(2) 住所表記数をカウント

それぞれのノードに対し，タグで囲まれる文字列に含まれる住所表記の数をカウントする（住所表記の検出には，[アドレスマッチング]で示した手法を利用する）．子ノードに含まれる住所表記の数も合計する．ページ全体に住所表記が複数見つかった場合，(5)へ.

(3) 電話番号の数をカウント

それぞれのノードに対し，タグで囲まれる文字列に含まれる電話番号の数をカウントする（電話番号の検出は，まず 5～11 桁の数字と記号（’-‘, ‘(‘, ‘)’ ）の並びを見つけ，市外局番・市内局番のリスト[総務省]を用いて妥当性を検証する）．子ノードに含まれる電話番号の数も合計する．ページ全体に電話番号が複数見つかった場合，(5)へ.

(4) テーブルタグの抽出

（ページ内に住所も電話番号も 1 つ以下しか存在しない場合）ページ内に現れるキーワードを全て含む

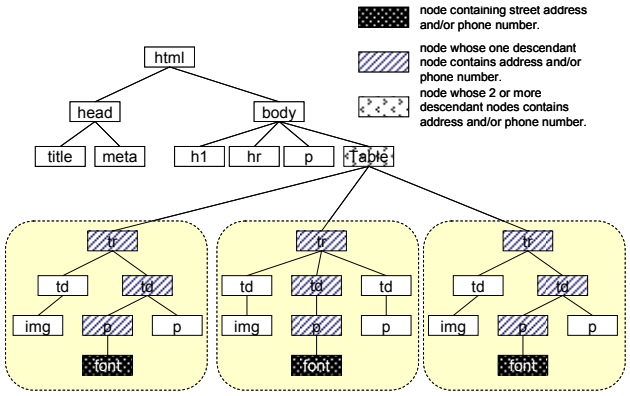

Fig.1: Text Block Extraction using HTML structure

最小の’<table>’ブロックを探し，見つかった場合にはそのブロックを抽出する．見つからなかった場合には’<body>’ブロックを抽出する．

(5) 分割するノード木のレベルを決定

あるノードから見て，複数の子ノードが住所表記または電話番号を含むとき，その子ノード以下を別々のブロックとして分割する．

なお(4)は，バナー広告やリンクメニューなどのノイズが含まれているページでは，<table>タグによってページの構成要素が分割されているケースが多く11), 構成要素のうち店舗の名称が含まれているものを取り出せば，ノイズを効率的に除去できるためである．

次に，上記のアルゴリズムを用いて抽出したテキストブロックに，以下の基準による確度のランク付けを行う．

確度 r の算定基準

r=5: ブロックに住所と電話番号が含まれている

r=4: ブロックに市外局番を含む完全な電話番号が含まれている

r=3: ブロックに店舗名と詳細な住所（地番レベル）が含まれている

r=2: ブロックに店舗名と地域名（丁目・字レベル）が含まれている

r=1: 上記のいずれかの条件に一致するが，ブロックに含まれる文字数が閾値（128 文字）以下

r=0: 上記のいずれの条件にも一致しない

3の実験により，このうち $r \geqq 3$ となるブロックを対象店舗に関するテキストとして採択する．

2.4. テキスト要約

採択された複数のテキストブロックに含まれる単語に対して，*tf*（term frequency）を計算する．その際，店舗名や住所に含まれる単語は除外する．次に *idf*（inverted document frequency）を計算するが，idf の計算は対象店舗以外の全店舗に対して抽出したテキストブロックに含まれる単語数を母数とする．*tf*idf* 値の高いものを対象店舗の特徴を表す索引語とする．

次にこの索引語を用いて重要文を抽出する．各文のスコア s は次式で計算する．直感的には特徴的な索引語を多く含む短い文が抽出される．

$$s = \frac{\sum_{i=1..n} tf(i) * idf(i)}{n}$$ (n は文中の単語数)

3　検索精度に関する実験

3.1.　ランク r の分布

NTT タウンページに含まれる都内のレストラン 107,127 店舗に対し，Web アーカイブから2.2で示した検索条件を満たす 3,845,183 ページを抜き出し，2.3で定義した確度 r の分布を調べた（Table 1）．

r	#pages
5	106,853
4	20,318
3	24,271
2	16,751
1	166,005
0	3,510,985

Table 1: Distribution of confidential value r

3.2.　目視によるサンプリング検査

提案手法では店舗と無関係として除去される r=0,1,2 と判定されたページについて，それぞれ 300, 200, 200 ページをランダムにサンプルとして抽出し，目視によって確認を行った．また，関係があるとされる r=3 以上のページについてはより詳しく 8 人の被験者により 8,348 ページを目視によって検査した（Table 2）．

Rank	All	Error	Correct	Sample	Est. Err.	Est. Corr.
5	106,853	635	5275	5910	11481	95372
4	20,318	333	699	1032	6556	13762
3	24,271	420	986	1406	7250	17021
2	16,751	181	19	200	15160	1591
1	166,005	21	179	200	17431	148574
0	3,510,985	291	9	300	3405655	105330

Table 2: Visual Check Result

表中，error は目視によって対象店舗と無関係と判断されたページ数，correct は対象店舗に関係すると判断されたページ数，sample はサンプリング数である．また，est.err.と est.corr.は，そのランクのページ数をサンプリング結果により比例配分した，無関係なページ数と関係のあるページ数の予測値である．

r=0 で対象店舗と関係があると判定された 9 件のうち，8 件は住所も電話番号も一切書かれていないケース，1 件は店舗移転前のページが残っていたため電話帳記載の住所と電話番号が一致しなかったケースである．

3.3.　Web ページ種別の分類

r≧3 と判定されたページについては，目視検査の際に以下の 9 つのカテゴリに分類した（Fig.2）．

（対象店舗と無関係なページ）
- e1：同名他店舗など，無関係なページ
- e2：系列店のページ
- e3：電話番号は正しいが店名が変わっている
- e4：レストランの情報以外がテーマのページ

（対象店舗と関係するページ）
- c1：表形式で一覧にまとめられたページ
- c2：口コミ情報など利用者が独自に作成したページ
- c3：雑誌・テレビなどによる特集記事
- c4：レストランのポータルサイト
- c5：レストランが直接作成，管理しているページ

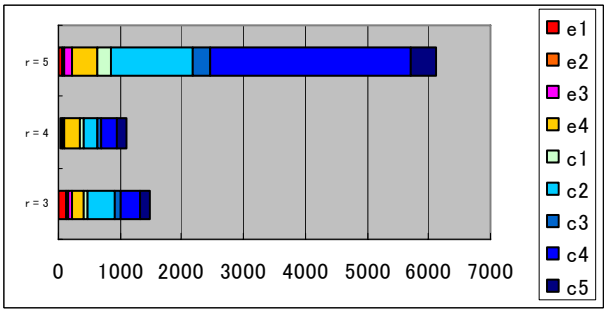

Fig.2: Distribution of Page Category

r ≧3 と判定されても人間が見ると対象店舗と無関係なページは約 15%であり，そのうち e4 が 2/3 近い 9.6%を占めている．e4 はたとえば同窓会のお知らせのページ内に，会場の情報として中華料理店が記載されている場合などである．

3.4.　考察

まず Table 1 と Table 2 より，店舗名などをキーワードとして検索した場合，約 9 割はその店舗と無関係なページであり，店舗情報の網羅的な検索が既存のサーチエンジンでは難しいことが分かる．

次に 3.3 より，提案手法の検索精度は適合率（precision）0.84 であり，実用に近いレベルにあるといえる．誤って検索されたページのうち 9.6%を占める e4 に分類されたページには，同窓会や講演会のお知らせ，ライブや写真展の開催情報など，いくつかの代表的なパターンが認められる．これらのパターンを検出することができれば適合率を大きく改善することができるため，今後の重要な研究課題である．

一方，再現率（recall）は Table 1, 2 より 0.330 と低い値となった．ただしこれには r = 1 に分類される一覧表のページが含まれており，これらのページが検索できてもあまり利用者にとって価値がない．そこで r =1 を母数から除外すると，再現率は 0.541 となる（適合率は変わらない）．再現率が低い原因は，住所も電話番号も記載されていないが対象店舗に関する情報を含むページが r = 0 と判定されてしまうためである．住所や電話番号以外に対象店舗を特定できる基準（たとえば業種固有のキーワードなど）によって再現率を改善できる可能性もあり，今後検討が必要である．

4　プロトタイプシステムの実装

提案手法によって収集した東京都内のレストランに関する情報を検索するプロトタイプシステムを UNIX 上に実装し，Web から検索できる試験サービスを公開した[5]（Fig.3）．

5　関連研究

本研究に密接に関連する研究としては，(1) Webから特定の話題を含むWebページを選択的に収集する手法[6)7)]，(2)Webページ内の人名・地名などの固有表現をマークアップする手法[8)9)10)]，(3)Webページからバナー広告やリンクメニューなど不要な情報を除去し，本来の内容部分（コンテンツ）を抽出する手法[11)12)]，(4)評価・評判表現を抽出する手法[13)14)15)16)]などが挙げられる．

(1)については，本研究ではWebアーカイブを作成する手法については特に検討していないため，既存の研究成果と組み合わせることでアーカイブの作成・更新効率を高めることができると考えられる．

(2)については自然言語処理の分野で多くの研究が進められているが，特に地名に注目した研究では同名地名の問題による多義性の解消が課題となっている[10)]．提案手法ではより詳細な住所のみに着目するため，地名の多義性は問題とはならない．

(3)については，提案手法では住所や電話番号という手がかりを用いてブロックを抽出している．しかしページ内に店舗が1つしか含まれない場合については精度が低いため，既存研究と組み合わせることで精度の向上が期待できる．

(4)については，「良い」「美しい」などの評価表現と呼ばれる単語や，それらの組み合わせを含む文を抽出し，教師情報を用いた学習によって肯定的・否定的な度合いを判定する手法が提案されている．本研究では対象とする店舗の業種を限定しないために評価表現を定義するのが難しいという問題があるが，今後検討するべき課題である．

6　おわりに

Webから店舗の評判やコメントなどを含むWebページを収集し，特徴的な文を抽出して簡潔に提示する検索手法を開発した．提案手法では，テキストからの店舗名の抽出，同名他店の検出，1ページ内に含まれる複数店舗に関する情報の分割，索引語に含まれる店舗名や店舗住所の除去などの比較的難しい処理を，電話帳を辞書として用いることで高い精度で行うことができる．

既存研究との組み合わせによる効率化や，さらなる検索精度（適合率・再現率）の向上が今後の課題である．

参考文献

1) http://itkaken.ex.nii.ac.jp/i-explosion/
2) http://www.w3.org/2001/sw/
3) http://www.almaden.ibm.com/webfountain/
4) Jens Graupmann, Ralf Schenkel, Garhard Weikum, The SphereSearch Engine forUnified Ranked Retrieval of Heterogeneous XML and Web Documents, Proc. of 31th VLDB, pp. 529-540, 2005.
5) http://157.82.157.42/~sagara/cgi-bin/search/restaurant.cgi
6) 横路誠司，高橋克巳，三浦伸幸，島健一，“位置指向の情報の収集，構造化および検索手法”，情報処理学会論文誌，Vol.41, No.7, pp.1987-1998, 2000
7) T. Tezuka, R. Lee, H. Takakura and Y. Kambayashi, ”Integrated Model for a Region Specific Search Systems and Its Implementation,” in Proceedings of 2003 IRC International Conference on Internet Information Retrieval, pp. 243-248, Koyang, Korea, (Oct. 2003)
8) 関根聡，“テキストからの情報抽出”，情報処理学会誌，Vol.40, No.4, pp.370-373，1999
9) 竹元義美，福島俊一，山田洋志，“辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出”，情報処理学会論文誌，Vol.42，No.6, pp.1580-1591，2001
10) Einat Amitay, Nadav Har’El, Ron Sivan, Aya Soffer, Web-a-Where: Geotagging Web Content, SIGIR2004, pp.273-280, 2004
11) D. Cai, S. Yu, J.-R. Web, W.-Y. Ma, Extracting content structure for web pages based on visual representation, Proc. 5th Asia Pacific Web Conference, Xi’an China, 2003.
12) D.Ikeda, Y.Yamada, S.Hirokawa. Expressive power of tree and string based wrappers Proc. UCAf Workshop on Information Integration on the Web pp.21-26 2-3
13) 立石 健二，石黒 義英，福島 俊一，インターネットからの評判情報検索，情報処理学会研究報告，NL-144-11, pp.75-82, 2001.
14) Kushual Dave, Steve Lawrence, David M. Pennock, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, International World Wide Web Conference(WWW2003), pp.519-528, 2003.
15) Jeonghee Yi, Wayne Niblack, Sentiment Mining in WebFountain, ICDE2005, pp.1073-1083, 2005
16) 長谷川 博之，工藤 峰一，中村 篤祥，構造と内容に基づく Web ページからの評判抽出におけるパターンの構成法，DEWS2005, 5-C-o4，2005

Fig.3: Implemented Prototype System