2008－NL－184 (6)
2008／3／27

携帯端末向け地図情報検索・閲覧システムの提案

安川美智子† 綱川有希子†† 横尾 英俊†

† 群馬大学大学院工学研究科 情報工学専攻
376-8515 桐生市天神町 1–5–1
†† 群馬大学工学部 情報工学科

あらまし 携帯電話等を用いたモバイル環境における情報の検索・閲覧は，端末の画面の大きさや文字入力機能に制限があることから，パソコンでの情報検索とは異なる工夫が必要である．このため，ユーザが必要とする情報に対して精度の高い分類を行い，提示できることが望ましい．本稿では，地図情報閲覧支援のための分類型情報検索を提案する．また，分類に用いる単語数の調整やトピックと無関係な単語の除去により，分類の精度がどの程度向上するかを考察する．

キーワード クラスタリング，文書自動分類，ローカルサーチ，モバイル情報検索

Searching and Browsing Local Information via Mobile Phone

Michiko YASUKAWA†, Yukiko TSUNAKAWA††, and Hidetoshi YOKOO†

† Graduate School of Engineering, Gunma University, Kiryu 376-8515, Japan
†† Department of Computer Science, Gunma University

Abstract Mobile terminals are much more limited in terms of input/output capabilities. Some special techniques that are substantially different from those for PCs must be incorporated into them to be easily used for Web searching. In order to decrease user frustration, it is desirable to provide users with search results in clusters. We present a searching system for maps and text information tailored to mobile terminals, which builds comprehensible clusters with user-selectable link functions. We report on an experimental study that compares the effectiveness of noise reduction from text data.

Key words Clustering, classification, local search, mobile information retrieval

1. はじめに

携帯電話等のモバイル端末の普及に伴い，外出先での情報検索を可能にする，モバイル情報検索 [1] が現実的なものとなってきている．従来のモバイル情報検索に関する研究では，ユーザの位置情報の取得や，Web ページを位置情報と関連づける処理など，モバイル情報検索の実現に必要となる基本的な技術が主な研究課題となっていた．Takahashi らは，Web ページから住所表記を抽出し，モバイルユーザの現在位置との距離が近い Web ページを提示する，位置依存型情報検索 (Location Based Search) の実験サービス「モバイルインフォサーチ」[4] を提案している．現在，ユーザの位置情報の取得は，携帯電話などにおいて可能となっており，たとえば，GPS 方式や通信基地局を利用した測位方式が携帯電話の標準的な機能として装備されている．また，Web ページと位置情報の関連付けについては，住所名や施設名など位置を表現する文字列（非位置データ）から，緯度・経度（位置データ）に変換する処理である，ジオコーディングが，Web 経由で手軽に行えるようになっている [5]．このように，現在，携帯電話等のモバイル端末を用いて，情報検索を行う上での技術要素は整いつつあると言えるが，モバイル情報検索は，現状では，広く普及しているとは言えない．モバイル端末での情報検索は，パソコンでの情報検索と比べると，表示画面が小さく，マウスがないなど，入力機能が不十分であり，キー操作も，パソコンのキーボード入力と比べると自由度が低いという問題がある．このため，モバイル情報検索では，パソコンでの情報検索とは異なる工夫が必要になると考えられる．

以上のような背景をもとに，我々は，モバイル環境において位置依存型の情報検索をより効果的に行えるようにす

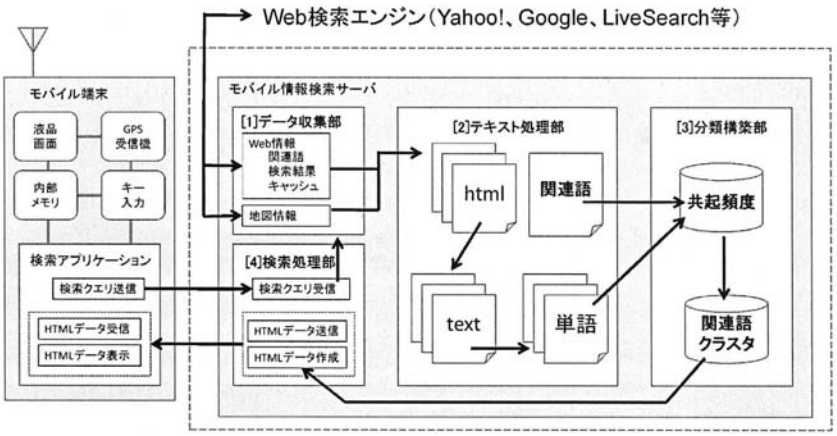

図1　システム構成

る，地図情報検索・閲覧システムを開発した．開発システムは，我々がこれまでに提案してきた分類型情報検索のモデル ([2], [3]) の地図情報検索への応用である．我々の提案する分類型検索のモデルは，検索ログから獲得した検索語の関連語で文書分類のためのクラスタラベルを作成し，作成したクラスタラベルに文書を対応付けることにより，ユーザにとって分かりやすい情報提示を可能とするものである．これまでの研究で，パソコン上で，一般的な名詞を検索語として，テキスト情報の検索を行う際の日本語の検索語・関連語 [2]，及び，英語の検索語・関連語 [3] について，分類型情報検索の有効性を確認している．

分類型情報検索は，簡潔で分かりやすい検索結果の提示が行えることから，モバイル端末における情報検索にも適している．しかし，地名・駅名など，固有名詞の周辺にはトピックに関連のある単語が少ないため，従来手法による検索語近傍抽出では，クラスタラベルがうまく作成できない．このため，一般名詞とは異なる，固有名詞の性質を考慮し，文書群から単語データをいかに精度よく抽出するかについての工夫が必要となる．

以下，2 章では，地図情報検索・閲覧システムと，分類のための単語絞り込みの工夫について述べ，3 章では，Web 検索結果の適合性に関する調査について，また，4 章では，分類に用いる単語の絞り込みに関する評価実験について述べ，最後にまとめと今後の課題について言及する．

2. 地図情報検索・閲覧システム

2.1 システム概要

地図情報検索・閲覧システムは，「データ収集部」「テキスト処理部」，「分類処理部」，及び，「検索結果表示部」の大きく 4 つの部分から構成される．各部の処理の概要を以下に説明する．

図2　地図情報検索・閲覧のイメージ (図中の地図は (株) アルプス社制作のプロアトラス SV3 の地図データを使用しました)

[1] データ収集部

Web 情報検索 ··· 検索語（地名・駅名）の関連語 [6] を取得する．取得した関連語と検索語で検索クエリを作成し，Web 検索 [7] を行い，検索結果（タイトル，URL，サマリ，キャッシュURL），及び，キャッシュページを取得する．

地図情報検索 ··· 検索語（地名・駅名）の周辺情報の検索 (ローカルサーチ) [5] を行い，住所の位置情報（緯度・経度），及び，周辺の施設名等の情報を取得する．

[2] テキスト処理部

テキストデータ抽出 ··· キャッシュページ (HTML ファイル) から，HTML タグを除去し，テキストデータ (EUC-JP テキストファイル) を抽出する．

単語データ抽出 ··· テキストデータを入力として，形態素解析 [8] を行い，名詞・未知語のみを抽出する．英語の場合は，単語の接辞処理 (stemming) を行い，語幹を抽出する．また，抽出した単語 (形態素または語幹) の絞り込みを行う．絞り込み処理の詳細は 2.2 節で説明する．

[3] 分類構築部

単語データ中の関連語間の共起頻度と関連語の文字列長に基づき，関連語のみに限定した単語クラスタリングを行い，分類型検索のクラスタラベルを作成する．このクラスタリング処理の詳細は，別途報告する予定であるので，本稿では説明を省略する．

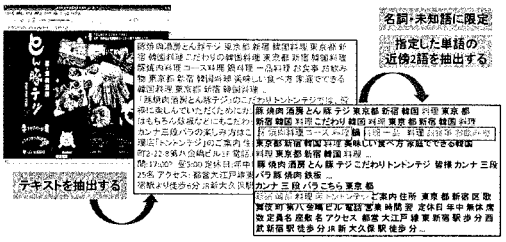

図3　単語抽出と近傍絞り込み

[4] 検索処理部

クラスタラベル中の単語からユーザが選択した任意の単語，もしくは，テキストフィールドに入力した単語で，テキストデータの全文検索を行い，検索されたテキストデータを表示する．テキストデータに住所相当文字列（例:「東京都新宿区○○丁目」等）が含まれる場合は，住所に対応する位置情報 (緯度・経度) の地図画像をテキストデータと共に表示する．また，ユーザが選択，もしくは，入力した単語で，連想語 (単語で検索される文書中において特徴的に出現頻度が高い語) の検索を行う．これにより，ユーザは，画面に表示された単語の中から好みの単語を選ぶだけで，効果的な検索を行うための検索クエリ精錬を行える (図 2).

2.2　単語データの絞り込み

Web ページには雑多な情報が多く含まれることから，単語の中にはノイズとなる (分類処理に役立たない) 単語が多数存在する．分類処理においてノイズとなるような単語は，単語抽出の段階で取り除いておくことが望ましい．取り除くべき単語として，以下のような単語が挙げられる．

- 重要な単語から離れた位置にある単語
- 文書群全体で低頻度の単語

一般に，検索結果中の検索語周辺のテキストデータには，検索語に関連のあるトピックの情報が記載されていることが多いことから，検索語の近傍の単語のみに限定して，関連語間の共起関係を抽出し，関連語のクラスタリングを行うことで，トピックに関連性の高い精度の良いクラスタリングが行える．しかし，地名・駅名等の固有名詞の検索語では，検索語の周辺は住所文字列など，トピックとは無関係な情報が多い．このため，検索語周辺で関連語間の共起関係を抽出すると，分類の精度が悪くなることが予想される．地名・駅名等の固有名詞の検索語では，トピックに関連のある単語は関連語の周辺の方にあることが多く，関連語周辺で関連語間の共起関係を抽出することで精度の良い分類が行えると考えられる (図 3).

また，文書群における単語のうち，頻度の高い単語のみを抽出することで，ノイズとなる低頻度の単語を除去できると考えられる．

3.　Web 検索結果の適合性

開発システムは，Web 検索エンジンの検索結果を収集し，分類提示を行うメタサーチエンジンであることから，Web 検索エンジンの検索結果に適合文書が含まれることが前提となる．この前提がどの程度成り立つのかを実際のデータに基づき検証する．

3.1　検索結果データの収集

東京の駅名 10 件 (秋葉原，恵比寿，銀座，原宿，池袋，目黒，表参道，渋谷，新宿，有楽町) と，駅名 10 件に共通して頻繁に検索される 5 つのトピック (グルメ，映画，ホテル，デパート，美容室)，及び，そのトピックに関連のある関連語で Web 検索を行った．具体的には，「(都道府県名) AND (地名／駅名) AND (トピックと関連語の OR 結合)」で，駅名とトピックの組み合わせごとに検索クエリを生成し，Web 検索を行い，各検索クエリにつき 1000 件の検索結果（タイトル，サマリ，URL，キャッシュURL），及び，キャッシュページをダウンロードした．検索クエリの生成に使用した関連語を以下に示す．

[1] グルメ … レストラン，イタリアン，フレンチ，中華，和食，洋食，焼肉，ケーキ，居酒屋，ラーメン

[2] 映画 … 映画館，シネマ，テアトル，上映，ロードショー，レイトショー

[3] ホテル … ビジネスホテル，宿泊，旅館

[4] デパート … 百貨店，ショッピング，買い物，買物

[5] 美容室 … 美容院，ヘアサロン，ネイルサロン，エステティックサロン

3.2　検索結果の適合性判定

収集した検索結果のうち，新宿の 5 トピックの上位 500 件について，タイトル，サマリ，キャッシュページを閲覧し，適合性の判定を行った．判定は，高適合のみを適合とした．検索クエリ中の単語を含んでいても，トピックとの関連性が低いと思われるページは不適合とした．上位 100 件までの 10 件ごと，及び，上位 500 件までの 100 件ごとの適合・不適合の判定結果を図 4〜図 13 に示す．全体的な傾向として，検索結果の上位 10 位までには不適合が少ない．また，上位 500 件までの分布では，検索結果の下位の方になると次第に不適合が増えるが，上位の方では比較的，適合文書が多い．また，グルメと美容室は，全体的に適合文書が多いが，映画，ホテル，デパートは適合文書が少ない．これは，グルメと美容室の関連語は他よりもトピックに対して適合性が高く，検索クエリが目的とする文書群を適切に説明できていたためと推測される．不適合文書は，検索クエリ中の単語を含んでおり，一見すると，適合文書のように見えるものも多かった．分類規則 (決定木) などを適用することで不適合文書を，適合文書と区別して扱えると考えられるが，これについては今後の課題である．

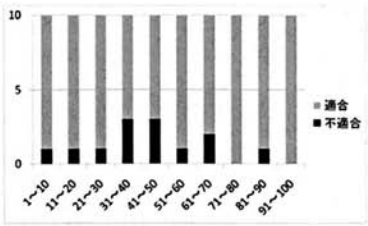

図 4　グルメ (上位 100 件)

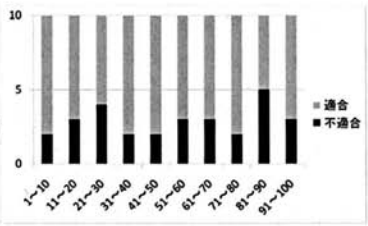

図 5　映画 (上位 100 件)

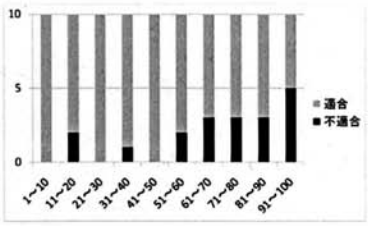

図 6　ホテル (上位 100 件)

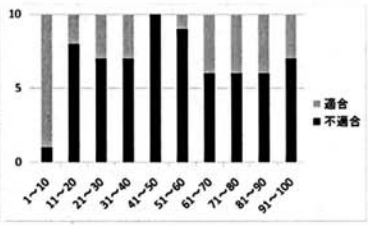

図 7　デパート (上位 100 件)

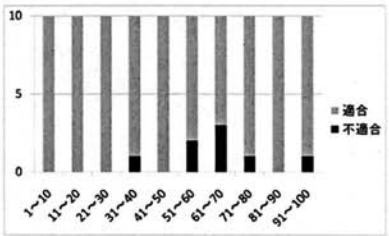

図 8　美容室 (上位 100 件)

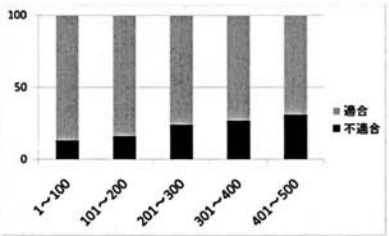

図 9　グルメ (上位 500 件)

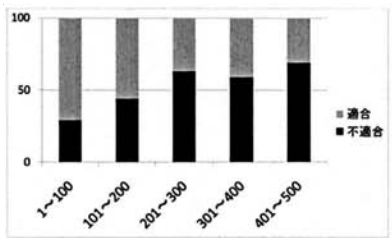

図 10　映画 (上位 500 件)

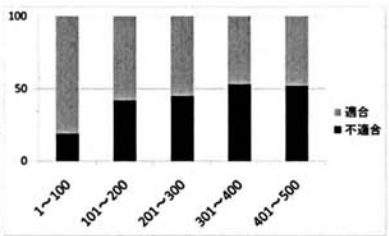

図 11　ホテル (上位 500 件)

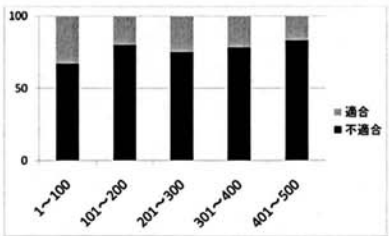

図 12　デパート (上位 500 件)

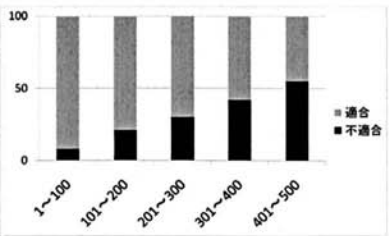

図 13　美容室 (上位 500 件)

トピックごとに観察された，適合・不適合の文書の詳細について以下に説明する．

[1] グルメ… 基本的には適合が多いが，後半になるにつれて不適合が増える．不適合の文書にはホテル関係のページや求人情報が多い．

[2] 映画… 不適合が多い．不動産関係や学校，俳優のプロフィール，ブログ，本など異なる分野のページが含まれている．不動産関係のページでは「近隣に映画館が…」のような表記が検索されている．

[3] ホテル… 後半になると不適合が多くなる．不動産関係やツアー，レンタルオフィスなどのページが含まれる．不動産関係のページは「ホテルのような設備の…」などの表記が検索されている．

[4] デパート… 不適合がとても多い．不動産関係やデパート以外の店，オンラインショッピング，会社の事務所などのページが多い．買い物という単語で検索されているものが多い．

[5] 美容室… 適合が多いが，後半になるにつれて，求人などの不適合の文書が多くなる．検索クエリに含まれているエステやネイルサロンのページを適合としたが，純粋に美容室関係のページだけを検索した場合は，不適合が増えると予想される．

4. 単語絞り込みの効果

Web 検索結果データと文書分類ソフトウエア TMSK [9] を用いて，分類に用いる単語の絞り込みの効果を検証する．以下では，まず，TMSK で実装されている線形モデルについて説明し，単語の絞り込みと単語数の調整を行った場合の TMSK の分類精度の向上を実験により確認する．

4.1 線形モデルに基づく文書分類

線形重み $(\hat{w}, \hat{b})$ は，文書ベクトル x とラベル $y \in \{-1, 1\}$ を学習データとして，学習データの平均エラーを最小化する値として，以下のように計算される [10].

$$(\hat{w}, \hat{b}) = \arg\min_{w} \frac{1}{n} \sum_{i=1}^{n} h(w^T x_i + b, y_i) \qquad (1)$$

ここで，h は，損失関数であり，次のように定義される．

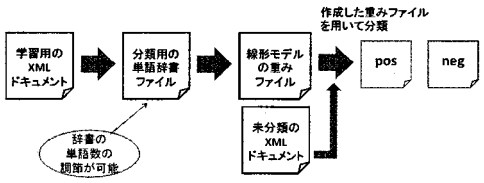

図 14 線形モデルに基づく文書分類

$$h(p, y) = \begin{cases} -2py & py < -1, \\ \frac{1}{2}(py - 1)^2 & py \in [-1, 1], \\ 0 & py > 1. \end{cases} \qquad (2)$$

式 (1) に基づき，線形重み $(\hat{w}, \hat{b})$ は，次のアルゴリズムで計算される．

Algorithm 1.

input: training data $(x_1, y_1), \cdots, (x_n, y_n)$
output: weight vactor(w,b)
let $\alpha_i = 0$ $(i = 1, \cdots, n)$
let $w = 0$ and $b = 0$
for $k = 1, \cdots, K$
  for $i = 1, \cdots, n$
    $p = (w^T x_i + b) y_i$
    $\Delta\alpha^i = \max(\min(2c - \alpha^i, \eta(\frac{c - \alpha^i}{c} - p)), -\alpha^i)$
    $w = w + \Delta\alpha^i x_i y_i$
    $b = b + \Delta\alpha^i y_i$
    $\alpha^i = \alpha^i + \Delta\alpha^i$
  end
end

線形重みの計算アルゴリズムは，大規模な文書群に対して，効率的な計算が行えるという特徴があり，F 値で測定される性能は，SVM(Support Vector Machine) と同じ程度の性能が得られることがデータ実験により確認されている [10]. 線形重み $(\hat{w}, \hat{b})$ を用いた文書分類の手順を図 14 に示す．

4.2 近傍絞り込みと辞書サイズの調整

TMSK(Text-Miner Software Kit) は，上に述べた線形モデルに基づく文書分類を行うソフトウエアである．XML 形式で記述された，ラベル付き文書を入力とし，文書ファイルからの分類用単語辞書ファイルの構築，文書ファイルのベクトル化などを，一括して行うことができる．TMSK では分類に用いる単語の個数 (辞書サイズ) を任意に定めることが出来る．

まず，3 章の適合判定で不適合の文書が少なかった「新宿」のトピック「グルメ」の文書を分類する実験を行った．検索結果上位には比較的適合文書が多いことが確認されていることから，「新宿」の各トピックの検索結果データの上位 100 件を学習データ，その次の 100 件を分類用テストデータとして使用し，以下の単語の近傍で絞り込みを行った．

(q) 検索語 (新宿)
(r) トピックと関連語 (例：グルメ, レストラン, …)
(f) 検索語，及び，トピックと関連語

絞り込み近傍サイズは，2 語，4 語，6 語とし，辞書サイズは，100 語，500 語，1000 語で線形重みの学習と分類を行った．結果を表 1 と図 15 に示す．図 15 の n2，n4，n6 は，それぞれ近傍 2 語，近傍 4 語，近傍 6 語の絞り込みを

表 1 「グルメ」の分類結果 (辞書サイズ 100 語)

	pos(正)	pos(誤)	neg(正)	neg(誤)
近傍絞り込み無し	66	58	342	34
近傍 2 (q)	39	36	364	61
近傍 4 (q)	49	54	346	51
近傍 6 (q)	46	37	363	54
近傍 2 (r)	95	8	392	5
近傍 4 (r)	97	6	394	3
近傍 6 (r)	88	8	392	12
近傍 2 (f)	86	11	389	14
近傍 4 (f)	81	16	384	19
近傍 6 (f)	75	25	375	25

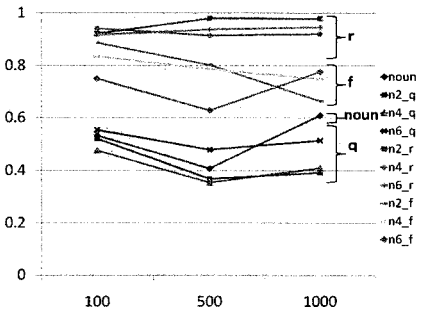

図 15 辞書サイズごとの分類精度 (precision)

表している．また，noun は絞り込み無しを表している．表 1 と図 15 から次のことが分かる．

- トピックと関連語の近傍で絞り込みを行うと，トピックと無関係な単語を除去でき，分類精度が良くなる傾向が見られる．この場合，辞書サイズの調整は，精度向上に特に効果は無い．
- 地名・駅名の検索語の近傍で絞り込みを行うとトピックと無関係な単語が多く含まれ，分類精度が非常に悪くなる．また，辞書サイズを大きくしても，分類精度の向上にあまり効果はない．
- 絞り込み無しの場合も，トピックと無関係な単語が多く含まれるが，辞書サイズを大きくすることで，精度向上が見込める．
- 検索語と関連語の両方の近傍を抽出すると，分類の性能は検索語の近傍，関連語の近傍の中間程度になる．また，辞書サイズを大きくしても分類精度の向上にあまり効果はないか，辞書サイズを大きくすると精度が悪くなる．

次に，10 駅，5 トピックの合計 50 件のデータで辞書サイズ 100 語の場合の (q) 検索語の近傍 2 語，及び，(r) トピックと関連語の近傍 2 語の絞り込みで分類を行い，F 値を調べた．F 値の計算式は次の通りである．

$$F = 2 \times (recall \times precision)/(recall + precision))$$

F 値の最小，最大，平均をそれぞれ表 2 に示す．表 2 より，検索語の近傍を抽出すると，トピックと無関係な単語が多くなるため，関連語近傍抽出の場合と比べて分類精度がかなり悪くなることが分かる．また，関連語の近傍を抽出した場合は，平均的に分類精度が良いことが分かる．

表 2 検索語近傍と関連語近傍の F 値

	最小	最大	平均
検索語 近傍 2 語以内	19.7	69.4	41.5
関連語 近傍 2 語以内	83.6	100.0	93.3

5. おわりに

本稿では，携帯電話等の端末で，地図情報を検索・閲覧するためのシステムを提案した．提案システムは，Web 検索エンジンの結果を収集し，分類提示するメタサーチシステムであることから，Web 検索エンジンの検索結果がどの程度，適合文書を含むかの検証を行った．また，検索結果のカテゴリでラベル付けした文書データと文書分類ソフトウエア TMSK を用いて，単語近傍の絞り込みと辞書サイズの調整により，分類の精度がどの程度向上するかを確認した．実験により，地名・駅名などの固有名詞の周辺ではなく，トピック・関連語の近傍で絞り込みを行うことが有効であり，絞り込みを行わない場合には，辞書サイズを大きめに取ることで，分類精度をやや改善できることが分かった．今後の課題として，Web 検索エンジンの検索結果に不適合な文書が多い場合には，分類規則 (決定木) の適用を行い，不適合文書を取り除くことを検討していきたいと考えている．

文 献

[1] 河野浩之, 山田誠二, 北村泰彦, 高橋克己: 情報検索とエージェント, 東京電機大学出版局, (2002).

[2] 安川美智子, 横尾英俊: クエリログから獲得した関連語のクラスタリングに基づく Web 検索電子情報通信学会論文誌, Vol.J90-D, No.2, pp. 269–280, (2007).

[3] Yasukawa, M., Yokoo, H.: Related Terms Clustering for Enhancing the Comprehensibility of Web Search Results. Proc. DEXA, pp. 359–368, (2007).

[4] Takahashi, K., Miura, N., Yokoji, S., Shima, K.: Mobile Info Search: Information Integration for Location-Aware Computing 情処学論, Vol.41, No.4, pp. 1192–1201, (2000).

[5] Yahoo! 地図情報: ⟨http:// developer.yahoo.co.jp/map/⟩

[6] Yahoo!検索 関連検索ワード: ⟨http:// developer.yahoo.co.jp/search/webunit/V1/webunitSearch.html⟩

[7] Yahoo!検索 ウェブ検索: ⟨http:// developer.yahoo.co.jp/search/web/V1/webSearch.html⟩

[8] 形態素解析システム茶筌: ⟨http://chasen.naist.jp/hiki/ChaSen/⟩

[9] Indurkhya, N., Zhang, T., Damerau, F. J., Weiss, S. M.: Text Mining: Predictive Methods For Analyzing Unstructured Information, Springer, (2004). ⟨http://www.data-miner.com/tmsk.pdf⟩

[10] Damerau, F. J., Zang, T., Weiss, S. M., Indurkhya, N.: Text categorization for a comprehensive time-dependent benchmark, Information Processing and Management, Vol. 40, pp. 209–221, (2004).