

地図上の文書データを利用したモバイル情報獲得方式の検討

高林 靖[†] 川口 紗保理[†] 鷹野 孝典[‡]

^{†,‡} 神奈川工科大学情報学部情報工学科 〒243-0292 神奈川県厚木市下荻野 1030

E-mail: [†] {s065105, s075053}@cce.kanagawa-it.ac.jp, [‡] takano@ic.kanagawa-it.ac.jp

あらまし 本稿では、モバイルユーザの情報獲得支援を目的として、地図上に対応づけられた文書データ群を利用した、モバイル環境における情報獲得システムの実現方式について検討する。本システムの特徴は、利用者がデジタル地図インターフェース上で閲覧する文書の内容に応じて、異なる分野ドメインを持った特徴単語集合を選択することにより、それぞれの特徴単語集合に対応した検索ドメインを持つ文書ベクトル空間コンポーネントを自動的に構築する点にある。検索処理において、構築された複数の文書ベクトル空間コンポーネント上で文書データ間の関連性計量を行うことにより、利用者の検索意図に応じた情報獲得が実現可能となる。本稿では、プロトタイプシステムを用いた予備実験を行い、提案方式の実現可能性を確認する。

キーワード 情報検索, ベクトル空間モデル, モバイル, 地図インターフェース

Mobile Document Search System using Digital Map Interface

Yasushi TAKABAYSHI[†] Saori Kawaguchi[†] and Kosuke TAKANO[‡]

^{†,‡} Department of Computer and Information Sciences, Kanagawa Institute of Technology 1030 Shimo-ogino, Atsugi-shi, Kanagawa, 243-0292 Japan

E-mail: [†] {s065105, s075053}@cce.kanagawa-it.ac.jp, [‡] takano@ic.kanagawa-it.ac.jp

Abstract In this paper, we present a mobile information retrieval system with a digital map interface for supporting mobile users to obtain proper information in the mobile environment. Our system adaptively selects feature term sets with different search domains according to documents being viewed by the user on the digital map interface, and constructs vector space components with the corresponding search domains. By computing relevancy between document data on the search engine components constructed, our system allows the mobile users to retrieve documents required according to their search intentions. By preliminary experiments using our prototype system, we show the feasibility of the proposed method.

Keyword Document retrieval, vector space model, mobile, digital map interface

1. はじめに

1.1. 背景

利用者の多様な検索目的に対応するために、様々な分野のデータベースが独立に構築されている。このようなデータベース環境において、各分野のデータベースを効率的に利用するために、利用者の問い合わせに応じて適切な分野のデータベースを抽出するデータベース選択方式[2], [7], [16], [17]が考案されている。データベース選択方式は、問い合わせに関連の少ない分野のデータベース群を問い合わせ対象から除外することにより、通信コストや関連性計量コストを削減するとともに、検索精度を向上させるのに有効である。しかしながら、データベース選択方式の多くは、重み付けされた問い合わせ単語集合のみを用いてデータベースを選択するため、検索プロセスにおいて、利用者の移り変わる意図や関心に適応することができない。

モバイル環境における情報獲得においては、利用者

の移り変わる意図や関心を検索システムに反映することが重要である。例えば、鎌倉のような歴史のある街を散策している際に、携帯端末を用いてオンライン地図上に配置された寺院に関する文書を閲覧している状況を想定する。この場合、源頼朝という問い合わせ単語を検索システムに入力したとすると、利用者は源頼朝に関わりのある神社について関心があると判断することができる。

利用者の意図や関心を反映させるための直接的な方法として、現在利用者が閲覧している文書に含まれる単語群を用いて、問い合わせ単語集合を拡張する手法が考えられる。しかしながら、そのような閲覧文書には、利用者の意図や関心とは無関係なノイズ単語が多く含まれている場合がある。従って、閲覧文書を直接利用して拡張した問い合わせ単語集合を用いて検索を行うと、検索精度が低下してしまう可能性がある。

1.2. アプローチ

Wikipedia に掲載されている地理データに関する文書等の例に見られるように、文書の内容に対応づけられた緯度・経度情報を抽出可能な文書群が作成されており、インターネット上で利用可能となっている。このような緯度・経度情報を伴った文書データをデジタル地図上に対応づけることにより、モバイル環境における情報獲得のための空間情報として利用できると考えられる。

本研究では、文献[13]に示されている方式を適用し、このようなデジタル地図上に配置された文書データを利用して、その文書データの内容に応じた、異なる検索ドメインを持つ文書ベクトル空間を動的に生成する方式を示す。本方式により、モバイルユーザは、現在位置する地点等に関する文書データをデジタル地図上で閲覧している際に、その場所に関する文書データの内容を検索コンテキストとして用いた情報獲得が可能となる。

文献[13]に示される方式は、閲覧文書の内容に応じて、異なる分野ドメインを持った特徴単語集合を選択することにより、それぞれの特徴単語集合に対応した検索ドメインを持つ文書ベクトル空間コンポーネントを作成する方式であり、以下の2機能を特徴とする。

- (1) 異なる分野ドメインの特徴単語集合の選択機能
- (2) 複数の文書検索ベクトル空間の構築機能

上記(1)、(2)の機能において、例えば、一つの閲覧文書中で、2つのトピック「鎌倉の神社」と「鎌倉仏教」が記述されていれば、それぞれに対応した文書ベクトル空間コンポーネントが作成される。ここで、文書ベクトル空間コンポーネントを構築するための特徴単語集合は、あらかじめ専用リポジトリにおいて事前に定義されているものとする。

本稿では、この方式を拡張し、地図上に対応づけられた文書群を抽出し、文書ベクトル空間コンポーネントの自動構築を可能とする地図インターフェース[8],[9],[12]を備えた、モバイル環境における情報獲得システムの実現方式について検討する。さらに、実際に構築した実験システムを用いて予備実験を行い、提案方式の実現可能性を確認する。

2. 検索シナリオ

本章では、モバイル環境において、デジタル地図上に対応づけられた文書データ群を検索コンテキストとして利用する情報検索を行う際の検索シナリオの例について説明する。下記のシナリオは、本研究において実現するモバイル情報獲得システムの利用例として想定している。

【シナリオ例】図1はシナリオ例の概要図を示してい

る。このシナリオでは、「Aくんは鎌倉の(緯度 経度)=(35.32 139.55)付近の地点において、鎌倉の歴史について学びたいと思っている」という状況を想定している。この状況において、Aくんは、デジタル地図上で半径250mの範囲内にある鎌倉の神社を検索すると、鎌倉の神社(宝戒寺)を見つけることができた(Step-1)。そこで、デジタル地図上に対応づけられている宝戒寺の内容に関する文書を閲覧した(Step-2)。Aくんは、文書の閲覧中に「仏教」というキーワードに興味を持ち、本システム上で検索キーワードとして入力する(Step-3)。すると、本システムは「宝戒寺」の内容に関する文書を検索コンテキストとして利用して、「宝戒寺」に関連する仏教の守護神(多聞天)に関する写真や記事を検索結果として提示した(Step-4)。Aくんは歴史的な遺産彫刻について興味を持っていたので、宝戒寺に興味を持ち、訪れてみることにした(Step-5)。

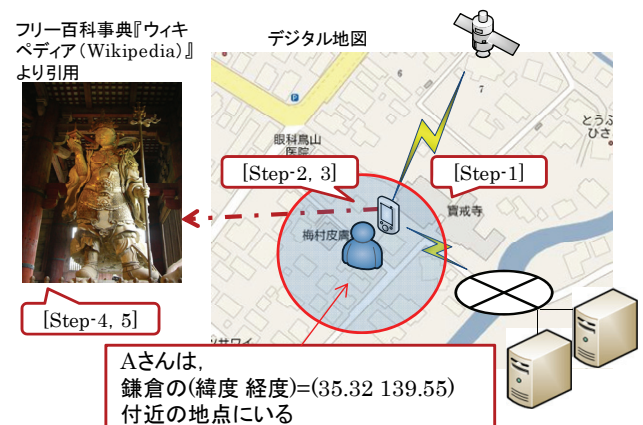


図1 検索シナリオの概要図

このように、本モバイル情報獲得システムを用いることにより、モバイルユーザは、現在位置する地点に関する文書データをデジタル地図上で閲覧している際に、その場所に関する文書データの内容を検索コンテキストとして用いた情報獲得が可能となる。上記の検索シナリオ例では、利用者は、検索をする際に検索キーワードを入力した。しかしながら、モバイルユーザにとって、このような検索キーワードを入力するのは手間であると考えられる。そのため、検索キーワードの履歴から、利用者の現在閲覧している文書に関連する検索キーワードを抽出し、自動的に検索を実行する方法等が考えられる。

3. 検索意図に応じた情報獲得方式

本章では、文献[13]に示されている異種の文書ベクトル空間を用いた、利用者の検索意図に応じたアダプティブな情報獲得方式の概要について述べる(図2)。本方式は、主に、以下の3機能、(i)対象文書からのキ

ワード抽出機能, (ii) 異なる検索ドメインを持った, 複数の文書ベクトル空間構築機能, (iii) 各文書ベクトル空間の返す検索結果の統合機能, より構成される。

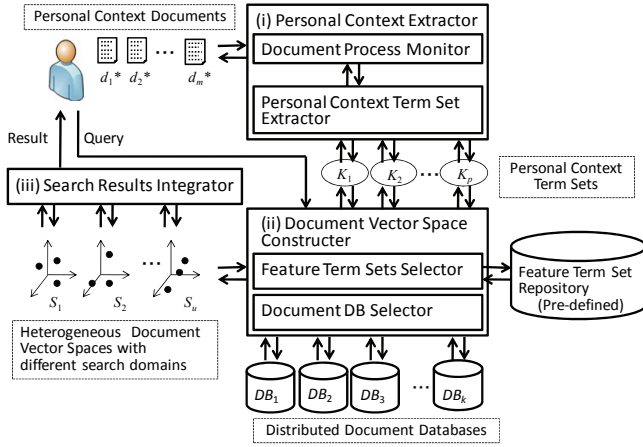


図2 システム構成図

3.1. 対象文書からのキーワード抽出

ここでは, 対象とする文書集合中に含まれるキーワード集合 K_p を抽出する。

まず, 各文書 d_x より単語集合 T_x が抽出される. T_x は d_x に含まれる m 個の単語を含む。

$$T_x = \{ t_{x1}, t_{x2}, t_{x3}, \dots, t_{xm} \}, \quad (1)$$

ここで, t_{xi} ($i = 1, 2, \dots, m$) は, T_x 中の各単語である. 各 T_x に含まれる全単語集合 T は, 次のように表わされる。

$$T = \bigcup_{i=1}^x T_i. \quad (2)$$

単語集合 T 中の単語群を用いて, 文書ベクトル空間を生成する. 文書 d_i^* は次のようにベクトル化される。

$$\mathbf{d}_i^* = [u_{i1}, u_{i2}, \dots, u_{i|T|}], \quad (3)$$

ここで, u_{ij} は文書 d_i^* における単語の出現頻度であり, $|T|$ は T に含まれる単語数である. 文書ベクトル空間において, 文書 $\mathbf{d}_i^*$ は, k -means 等の分類アルゴリズムを用いて, k 個のクラスターに分類される. 分類結果に基づいて, クラスター C_p に分類される文書 d_i 中に含まれるキーワード集合 K_p が抽出される。

$$K_p = \{ t \in d_i \wedge d_i \in C_p \} \quad (4)$$

3.2. 文書ベクトル空間の生成

キーワード集合 K_p を用いて, 特徴単語集合 F_q ($q = 1, 2, 3, \dots, u$) を選択し, ベクトル空間モデルに基づいた検索コンポーネント群を生成する. 検索コンポーネントの生成プロセスは, 次の 3 機能: (1) 特徴単語選択機能, (2) 文書データベース選択機能, (3) 文書の写像機能, より構成される. 以下, これらの機能の詳細について述べる。

3.2.1. 特徴単語選択

ここでは利用者の現在の意図や関心に基づいた相関量計算が可能な文書ベクトル空間の構築を目的として, 対象文書群から抽出されたキーワード集合 K_p を用いて, 特徴単語集合 F_q を選択する. 本研究では, 各特徴単語集合はあらかじめ定義されており, 専用のアプリケーション・サーバ等から抽出できるものと仮定している. 特徴単語集合 F_q は v 個の特徴単語を用いて, 下記のように表わされる。

$$F_q = \{ ft_{q1}, ft_{q2}, ft_{q3}, \dots, ft_{qv} \}, \quad (5)$$

ここで, ft_{qi} ($i = 1, 2, \dots, v$) は, F_q に含まれる単語要素である. 例えば, F_q の分野ドメインがバイオインフォマティクスであるとするれば, F_q は *DNA*, *RNA*, および *gene* などのバイオインフォマティクス分野に関連の強い単語群から構成される。

キーワード集合 K_p と特徴単語 F_q の関連性を計量するために, K_p と F_q を用いて, 文書ベクトル空間 S_q を構築する. S_q が v 次元のベクトル空間であるとするれば, K_p は, 文書ベクトル空間上 S_q で v 次元ベクトル $\mathbf{K}_p^{S_q}$ として表わされる。

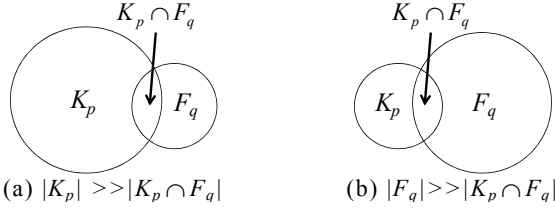
$$\mathbf{K}_p^{S_q} = [e_{p1}, e_{p2}, e_{p3}, \dots, e_{pv}], \quad (6)$$

ここで, e_{pi} ($i = 1, 2, \dots, v$) は, クラスター C_p における $f_{pi} \in F_p$ の出現頻度である. 例えば, F_{bio} と F_{mining} を, それぞれバイオインフォマティクス分野とデータマイニング分野を分野ドメインとする特徴単語集合であるとする. また, $F_{bio} = \{DNA, RNA, gene\}$, $F_{mining} = \{clustering, distance, attribute, Euclid\}$, および $K_p = \{DNA, RNA, clustering, computer, human, visualization\}$ であるとする. この場合, K_p はベクトル空間 S_{bio} と S_{mining} 上で, 次のようにベクトル化される: $\mathbf{K}_p^{S_{bio}} = [1, 1, 0]$, $\mathbf{K}_p^{S_{mining}} = [1, 0, 0, 0]$.

キーワード集合 K_p と特徴単語 F_q の関連性を計量するために, 次のようにベクトル $\mathbf{K}_p^{S_q}$ の 2 ノルム値を算出する。

$$\eta(\mathbf{K}_p^{S_q}) = \sqrt{e_{p1}^2 + e_{p2}^2 + e_{p3}^2 + \dots + e_{pv}^2} \quad (7)$$

K_p と S_q の全てのペアに対して、 $\mathbf{K}_p^{S_q}$ の 2 ノルム値を計算する。



$|K_p|$, $|F_q|$, および $|K_p \cap F_q|$ を、それぞれ K_p , F_q , および $K_p \cap F_q$ に含まれる単語数とする。式(7)において、 $\mathbf{K}_p^{S_q}$ の 2 ノルム値の大きさは、 $|K_p \cap F_q|$ に依存する。しかしながら、 $|K_p \cap F_q|$ が大きくとも、 $|K_p| \gg |K_p \cap F_q|$ の場合(図 3(a))には、 F_q に基づいて作成される文書ベクトル空間 S_q は検索プロセスにおいて、キーワード集合 K_p に内在するコンテキストを一部しか反映することができない。さらに、 $|F_q| \gg |K_p \cap F_q|$ の場合(図 3 (b))には、 S_q は K_p に内在するコンテキストの範囲内において、検索対象文書を精度良く計量することができない。このため、利用者の現在の作業コンテキストに適応するために、 $|K_p|$, $|F_q|$, および $|K_p \cap F_q|$ の比率に基づいて、特徴単語集合 F_q を選択する必要がある。

どの程度、 S_q がキーワード集合 K_p のコンテキストを反映しているかを評価するために、下記のように特徴単語の共有率 r_1 を定義する：

$$r_1(K_p, F_q) = |K_p \cap F_q| / |K_p| \quad (8)$$

さらに、どの程度、 S_q が K_p のコンテキスト領域で計量可能であるかを評価するために、下記の特徴単語の共有率 r_2 を定義する：

$$r_2(K_p, F_q) = |K_p \cap F_q| / |F_q| \quad (9)$$

特徴単語選択機能では、下記の条件を満たす u 個の特徴単語集合を選択する。

$$\eta(\mathbf{K}_p^{S_q}) > \varepsilon_1 \quad (10)$$

$$r_1(K_p, F_q) > \varepsilon_2 \quad (11)$$

$$r_2(K_p, F_q) > \varepsilon_3 \quad (12)$$

ここで、 ε_1 , ε_2 および ε_3 は、それぞれ、 $\eta(\mathbf{K}_p^{S_q})$, r_1 および r_2 の閾値である。これらの選択された特徴単語

集合を用いて、 u 個の文書ベクトル空間 $S_1, S_2, \dots, S_u$ を生成する。

3.2.2. データベース選択

本方式では、利用者の問い合わせ単語 $q_1, q_2, \dots, q_n$ に基づいて r 個のデータベース $D_1, D_2, \dots, D_r$ を選択する。データベース選択において、CORI[2], GLOSS[7], CVV[17]等の既存のアルゴリズムを用いる。現在実装しているプロトタイプシステムでは、CVV (Cue-Validity Variance) アルゴリズムを実装している。CVV 方式の詳細は、文献[17]に示されている。CVV 方式は、各データベース中の DF データに基づいて、問い合わせに対応したデータベースを選択する。コレクション集合 C が与えられると、CVV 方式は各コレクション $c_i \in C$ に対して、問い合わせ q に応じた goodness スコアを割り当てる。

$$G_{i,q} = \sum_{j=1}^M CVV_j \cdot DF_{i,j}, \quad (13)$$

ここで、 $DF_{i,j}$ は、 c_i の DF_j であり、 CVV_j は、単語 j のキューの妥当性を示す CV_j の分散である。 c_i に対する単語 j のキューの妥当性を示す $CV_{i,j}$ を、下記のように定義する。

$$CV_{i,j} = \frac{DF_{i,j} / N_i}{DF_{i,j} / N_i + \sum_{k \neq i}^{|C|} DF_{k,j} / \sum_{k \neq i}^{|C|} N_k}, \quad (14)$$

ここで、 N_i は c_i における文書数であり、 $|C|$ はシステムが対象とするコレクション数である。 $CV_{i,j}$ の母分散 CVV_j はコレクションに対する単語 j の分布の歪度を計量する。 CVV_j は次式のように計算される。

$$CVV_j = \frac{\sum_{i=1}^{|C|} (CV_{i,j} - \overline{CV_j})^2}{|C|}, \quad (15)$$

ここで、 $\overline{CV_j}$ は、全てのコレクションに対する $CV_{i,j}$ の母平均である。

3.2.3. 文書の写像

選択されたデータベース $D_1, D_2, \dots, D_r$ 中の各文書 d_i は、選択された文書ベクトル空間 $S_1, S_2, \dots, S_u$ 上に写像される。 d_i が v 次元のベクトル空間 S_q に写像されたとすると、 d_i は S_q 上で、次のようにベクトル化される：

$$\mathbf{d}_i^{S_q} = [s_{i1}, s_{i2}, s_{i3}, \dots, s_{iv}], \quad (16)$$

ここで, s_{ij} ($j = 1, 2, \dots, v$) は, 文書 $d_i \in D_1 \cup D_2 \cup \dots \cup D_r$ 中の特徴単語 f_{qi} の出現頻度である.

3.3. 文書獲得方式

文書獲得プロセスでは, 問い合わせ Q と文書 d_i の相関値を, 下記のステップで計算する.

Step-1: 選択されたそれぞれのベクトル空間 S_q において, 各文書ベクトル $\mathbf{d}_i^{S_q}$ のノルム値を計算する.

$$\eta(\mathbf{d}_i^{S_q}) = \sqrt{s_{i1}^2 + s_{i2}^2 + s_{i3}^2 + \dots + s_{iv}^2} \quad (17)$$

例えば, $\mathbf{d}_1^{S_{bio}} = [1, 1, 0]$, $\mathbf{d}_2^{S_{bio}} = [0, 1, 0]$, $\mathbf{d}_3^{S_{bio}} = [1, 1, 1]$ とすると, $\eta(\mathbf{d}_1^{S_{bio}}) = \sqrt{2} = 1.41$, $\eta(\mathbf{d}_2^{S_{bio}}) = 1$, $\eta(\mathbf{d}_3^{S_{bio}}) = \sqrt{3} = 1.73$ となる.

Step-2: 本方式では, 文書ベクトル空間 S_q において, 多くの文書が大きなノルム値を伴って写像される場合, S_q は問い合わせ Q にとって重要であると仮定する. このため, ベクトル空間 S_q 上の文書ベクトル $\mathbf{d}_i^{S_q}$ のノルム値の総和に応じて, 各 S_q に対して重み値 w_q を設定する. ノルム値の総和は, その最大値を用いて次のように正規化される.

$$w_q = \frac{\sum_{i=1}^N \eta(\mathbf{d}_i^{S_q})}{\Sigma_{\max}} \quad (18)$$

ここで, 選択されたデータベース $D_1, D_2, \dots, D_r$ 中の N は総文書数であり, $\Sigma_{\max}$ は, 各文書ベクトル空間

S_q におけるノルム値の総和 $\sum \eta(\mathbf{d}_i^{S_q})$ の最大値である. 例

えば, Step-1 の例を用いると, $w_{bio} = (1.41 + 1 + 1.73) / 1.73 = 2.39$ と計算される.

Step-3: 選択されたそれぞれのベクトル空間 S_q において, 文書ベクトル $\mathbf{d}_i^{S_q}$ の重み付ノルム値 W を計算する.

$$W(\mathbf{d}_i, S_q) = w_q \times \eta(\mathbf{d}_i^{S_q}) \quad (19)$$

Step-4: 問い合わせ Q と文書 d_i の関連性を評価するために, 選択されたそれぞれの文書ベクトル空間における $W(\mathbf{d}_i, S_q)$ の和をとることにより, 相関値を算出する.

$$V(Q, d_i) = \sum_{S_q \in S} W(\mathbf{d}_i, S_q) \quad (20)$$

ここで, S は選択されたベクトル空間集合である. 各文書 d_i は, 式(20)の結果に基づいてランキングされる. 例えば, $S_{bio}, S_{mining} \in S$, $W(\mathbf{d}_1, S_{bio}) = 3$, $W(\mathbf{d}_1, S_{mining}) = 2$

であるとする, $V(Q, d_1) = 3 + 2 = 5$ と計算される.

4. モバイル情報獲得システムの実現方式

本章では, 3 章で述べた方式を適用することにより実現される, モバイル環境における情報獲得システムの実現方式について述べる. 図 4 は, 本モバイル情報獲得システムの構成図を示している.

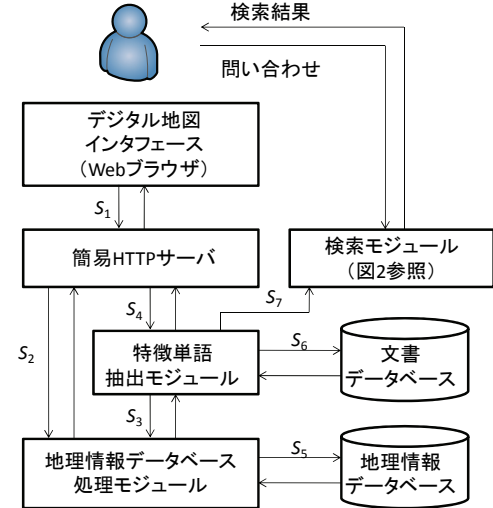


図4 モバイル情報獲得システムの構成図

本システムでは, デジタル地図上に対応づけられる文書データから, 3 章で示した情報獲得方式において, 特徴単語集合を選択するためのキーワード集合 K_p を抽出する機能を実現している. この機能は, 以下のステップにより実現される.

ステップ 1: デジタル地図インターフェース上で, 利用者が選択した地域範囲を簡易 HTTP サーバに送信する (S_1). 地域範囲は, 指定した中心座標 (緯度・経度) と半径 (m) からなる円の内部として表される. 円の中心座標, および, 半径は GET メソッドにより, URL に付加して送信される.

`http://127.0.0.1:8080/la=latitude&lo=longitude&r=radius
&m=mode`

ここで, latitude, longitude, radius, mode は, それぞれ緯度, 経度, 半径, 検索モードを表している. 検索モードに応じて, 独立に構築した地理情報データベースから, それぞれに対応したデータ群を検索することができる.

ステップ 2: 地理情報データベース処理モジュールは, 受信した緯度・経度情報, および, 中心点座標からなる円の内部に位置するデータ群を, 地理情報データベースより抽出する (S_2, S_5).

ステップ 3: 特徴単語処理モジュールは, ステップ 2

において得られたデータ群に対応する文書データ群を、文書データベースから選択する。次に、選択された文書データ群から、その文書中に出現するキーワード集合 K_p を抽出する (S_3 , S_6)。

ステップ 4: ステップ 3 で抽出されたキーワード集合 K_p を、3 章で示した情報獲得システムに送信する (S_7)。情報獲得システムでは、このキーワード集合 K_p に基づいて、複数の異なる分野の特徴単語群を抽出し、それぞれに対応した検索ドメインを持った文書ベクトル空間を構築する。

5. 予備実験

本章では、実際に構築したプロトタイプシステムを用いた予備実験について述べる。本実験では、提案システムを用いて、デジタル地図上の文書データの内容を検索コンテキストとして利用した情報獲得が実現可能であることを検証する。

5.1. 実験環境

プロトタイプシステムでは、3.2.1 節で述べた特徴単語選択機能を評価するために、芸術、自然科学、工学、日本仏教等の様々な分野に関する 129 種類の特徴単語集合を用意した。また、3.3 節で述べた文書獲得機能を評価するために文書テストコレクションとして広く利用されている CACM, MED, NPL, REUTERS21578 に加え、西欧芸術、鎌倉の神社、映画、東京の娯楽施設、ワシントン DC 建築等の 14 種類の分野別データベースを独自に構築した。データベース中の文書数は、合計 23,207 件である。なお、特徴単語群、および文書の収集には、それぞれ Wikipedia のカテゴリリスト、および Wikipedia 記事や観光ガイドを主に利用した。



図5 地図インターフェース

また、構築したデータベースのうち、鎌倉の神社、東京の娯楽施設、ワシントン DC 建築のそれぞれのデータベース中の文書については、緯度・経度情報を付加している。緯度・経度を付加することにより地図上に対応づけられた文書群は、プロトタイプシステムに

実装されている地図インターフェース（図 5）を用いて抽出することが可能である。本地図インターフェースは、Google map を利用しており、半径を指定した上で地図上の点をマウスクリックして緯度・経度を指定することにより、クリック点を中心とした範囲円に含まれる、地図に対応づけられた文書データ群を抽出することができる。

5.2. 実験方法

モバイル環境における情報獲得を想定した検索実験を行うために、下記の 3 つのシナリオを用意する。

[シナリオ 1] 利用者は、(緯度, 経度)=(35.32, 139.55), 半径=250m の範囲にある、鎌倉の神社に関する文書 d_1 (宝戒寺)を閲覧している。

[シナリオ 2] 利用者は、(緯度, 経度)=(35.66, 139.70), 半径=30m の範囲にある、渋谷の映画館(ユーロスペース)で上映されている映画(「天使と悪魔」)に関する文書 d_3 を閲覧している。

[シナリオ 3] 利用者は、(緯度, 経度)=(38.89, -77.55), 半径=100m の範囲にある、ワシントン D.C.の建物に関する文書 d_3 (Robert F. Kennedy 司法省ビル)を閲覧している。

上記の 3 つのシナリオにおいて、それぞれ下記の問い合わせ $q_1 \sim q_3$ を用いて、提案方式による検索を行う。

- q_1 : Buddhism
- q_2 : Rome, art
- q_3 : Washington, movie

さらに、下記の 2 つの検索方式を用いた検索実験を行い、提案方式との検索結果を比較する。

[類似文書検索方式] 各シナリオで、利用者が閲覧していると想定する文書 $d_1 \sim d_3$ を用いて、コサイン尺度により、類似文書検索を行う。

[データベース選択方式] 問い合わせ $q_1 \sim q_3$ を用いて、3.2.2 節で述べたデータベース選択方式である CVV 方式により、検索を行う。

5.3. 実験結果

表 1 は、3.2.1 節で述べた特徴単語抽出方式に基づき、シナリオ 1~3 における閲覧文書 $d_1 \sim d_3$ を用いて、抽出された特徴単語集合を示している。表 1 より、各閲覧文書に応じて、適切な特徴単語集合が選択されることがわかる。例えば、 d_1 は鎌倉にある神社である宝戒寺に関する文書であるので、宝戒寺に深い関連のある真言仏教や日本建築に関する特徴単語集合が選択されていることが確認できる。

また、表 2 は、問い合わせ $q_1 \sim q_3$ を用いて、CVV 方式に基づいて選択されたデータベースを示している。提案方式、および、データベース選択方式を用いた検索実験において、表 2 に示されるように、高い CVV 値が算出された上位 3 件のデータベースを選択している。

表 1 選択された特徴単語集合

文書 ID	選択された特徴単語集合 (算出されたノルム値)				閾値	
					r_1	r_2
d_1	1	真言仏教 (10.82)	2	日本建築 (10.68)	0.03	0.10
d_2	1	ローマ教皇 庁(28.53)	2	Pope Leo XIII (17.92)	0.04	0.14
d_3	1	ワシントン D.C. の米 国 政 府 ビ ル (11.83)	2	ワ シ ン ト ン D.C. における オフィスピル (10.10)	0.30	0.10

表 2 選択されたデータベース

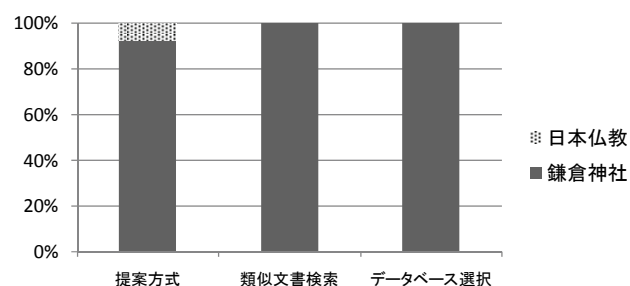
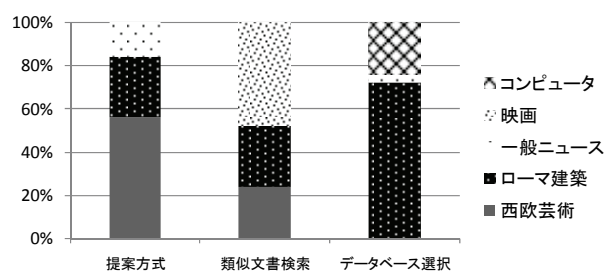
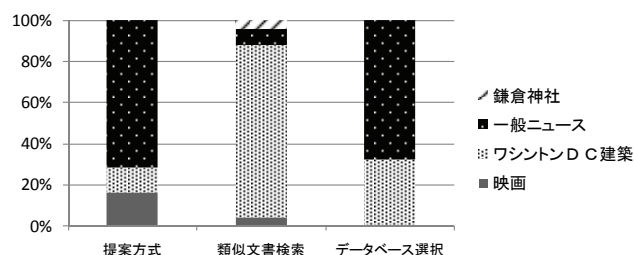
問い合 わせ	選択されたデータベース (CVV 値)		
	1	2	3
q_1	鎌倉神社 (5.58)	日本仏教 (2.19)	日本建築 (0.80)
q_2	西欧芸術 (6.95)	ローマ建築 (5.89)	CISI (4.93)
q_3	REUTERS (18.72)	ワシントン DC 建築 (7.42)	映画 (4.34)

図 6～8 は、それぞれ $q_1 \sim q_3$ を用いた検索結果における上位 25 件中の文書ジャンルの分布について、提案方式、類似文書検索方式、およびデータベース選択方式で比較した結果を示している。図 6 より、類似文書検索方式とデータベース選択方式を用いた検索では、鎌倉神社に関する文書しか上位 25 件中に検索できていない。これは、類似文書検索方式では、入力文書と出現単語の類似した文書の検索を行い、また、データベース選択方式では、表 2 に示されている選択されたデータベースにおいて、検索単語 Buddhism を含む文書を検索しているためである。これに対して、提案方式では、表 1 に示される特徴単語集合（真言仏教、日本建築）を用いて構築される文書ベクトル空間上で検索単語 Buddhism に関連のある文書を計量することにより、鎌倉神社に関連のある日本仏教の文書 2 件(15 位, 23 位)を検索できていることがわかる。

同様に、図 7 より、提案方式では、映画「天使と悪魔」中に登場するローマ建築やルネサンス期西洋芸術に深く関連する文書を上位に検索できていることがわかる。これに対して、データベース選択方式では、ローマ建築に関する文書は検索されているが、西洋芸術に関する文書は検索されていない。また、類似文書検索方式では、ローマ建築や西洋芸術に関する文書がともに検索されているが、問い合わせに無関係な映画に関する文書も多く検索されている。

図 8 に示す結果では、提案方式では、少数ではあるが、問い合わせに関連のある、ワシントン D.C. における建物、および映画に関する文書を検索できていることが確認できる。しかしながら、検索意図である「ワシントン D.C. に関連のある映画」については上位に検索することができなかった。なお、検索結果に一般ニ

ュースに関する文書が多いのは、これらの文書に U.S. や office など、ワシントン D.C. における建築物についての特徴単語群との共通単語が多く含まれているためである。また、図 8 に結果については、各分野の特徴単語集合の質を高めることにより、検索精度を改善することができると考えられる。

図 6 検索結果上位 25 件の文書分布比較(q_1)図 7 検索結果上位 25 件の文書分布比較(q_2)図 8 検索結果上位 25 件の文書分布比較(q_3)

6. まとめ

本稿では、地図上の文書データを利用したモバイル情報獲得方式の検討を行った。本方式は、地図上に対応づけられた文書群を抽出し、その抽出結果に基づいて文書ベクトル空間コンポーネントの自動構築を行うものである。構築された複数の文書ベクトル空間を用いて、抽出文書に応じた複数分野の検索ドメインでの文書検索が可能となる。

今後の課題として、モバイルユーザが地図インターフェースで閲覧する文書内容の変化が、そのモバイルユーザの検索意図や関心の移り変わりを反映するもの

であると捉えることにより，モバイルユーザの検索要求に沿った状況適応的な問い合わせ生成機能を実現していくことが挙げられる．さらに，提案方式の検索性能に関する有効性を検証するとともに，屋内外における実証実験を行うことにより，モバイル環境における情報獲得機能の有用性評価を行っていく予定である．

[謝辞] 本研究の一部は，科研費(20700100)の助成により行なわれた．

文 献

- [1] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley, 1999.
- [2] Callan, J. P., Lu, Z. and Croft, W. B. Searching distributed collections with inference networks. In SIGIR, 1995.
- [3] Chen, X. and Kiyoki, Y., A Dynamic Retrieval Space Creation Method for Semantic Information Retrieval, Information Modelling and Knowledge Bases, Vol.XVI, IOS Press, pp.46-63, 2005.
- [4] Dasgupta, A., Drineas P., Harb, B., Kumar, R., Josifovsky, V. and Mahoney M. W., Feature Selection Methods for Text Classification, Proceedings of the Thirteenth Annual ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2007, San Jose, California, USA, August 2007.
- [5] Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K. and Harshman, R., Indexing by latent semantic analysis, Journal of the American Society for Information Science, Vol. 41, No. 6, pp.391-407, 1990.
- [6] Forman, G. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3:1289-1305, 2003.
- [7] Gravano, L., Garcia-Molina, H. and Tomasic, A. GLOSS:text-source discovery over the Internet. ACM Transactions on Database Systems, 24(2):229-264, 1999.
- [8] 平松 薫, 小林堅治, 石田亨, 赤埴淳一, "デジタルシティにおける情報検索のための地図インタフェース," 情報処理学会, Vol.41, No.12, pp.3314-3322, 2000.
- [9] 平元 綾子, 角谷和俊, "地図インタフェースにおけるユーザ操作からの Web 検索方式", 日本データベース学会, Vol.5, No.1, pp. 85~88, 2006.
- [10] Kiyoki, Y., Kitagawa, T. and Hayama, T., A metadatabase system for semantic image search by a mathematical model of meaning, ACM SIGMOD Record, Vol.23, No. 4, pp.34-41, Dec. 1994.
- [11] Last, M., Kandel, A., Maimon, O., and Eberbach, E., Anytime Algorithm for Feature Selection, Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science, pp.532-539, 2001.
- [12] Sasaki, S., Takahashi, Y. and Kiyoki, Y., "The 4D World Map System with Semantic and Spatio-temporal Analyzers," the 19th European-Japanese Conference on Information Modelling and Knowledge Bases, pp.11-24, June 1-5th, Maribor, Slovenia, 2009.
- [13] Takano, K., Kurabayashi, S., Chen, X., Kiyoki, Y., An Adaptive Search System using Heterogeneous Document Vector Spaces, Proceedings of the 2009 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (Pacrim2009), pp.193-198, 2009.
- [14] Wong, S. K. M., Ziarko, W., Wong, P. C. N., Generalized Vector Space Model in Information Retrieval, Proceedings of SIGIR, pp.18-25, 1985.
- [15] Yu, B., Li, G., Sollins, K. and Tung, A. K. H.: Effective keyword-based selection of relational databases, Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp.139-150, 2007.
- [16] Yu, C., Meng, W., Wu, W. and Liu, K. L. Efficient and effective metasearch for text databases incorporating linkages among documents, In SIGMOD, 2001.
- [17] Yuwono, B. and Lee, D. L. Server ranking for distributed text retrieval systems on the Internet, Proceedings of the Fifth International Conference on Database Systems for Advanced Applications (DASFAA), 1997.
- [18] 有川 正俊, 太田 守重, GIS のためのモデリング入門 地理空間データの設計と応用, ソフトバンククリエイティブ, 2007.