社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－DBS－140（II）（35）
2006／7／13

# ブログからのビジターの代表的な行動経路とそのコンテキストの抽出

郡　宏志†　服部　峻†　手塚　太郎†
田島　敬史†　田中　克己†

近年，ユーザが Blog の中で自らの行動を日記として発信することが盛んになってきている．そのような Blog の中には，ユーザの行動経路が地名を含むことにより記述されていることも多い．そこで，我々は Blog からユーザの行動時の代表的な経路とその文脈を抽出し，それらを地図上にマッピングすることにより集約して提示するシステムを提案する．ユーザの行動経路の文脈は，その経路を経由したユーザにおける，行動のテーマを表すキーワードという形で表される．また，ユーザの代表的な行動経路については，代表的なシーケンシャルパターンマイニングである PrefixSpan を用いて抽出する．また，マイニングを行うために，各 Blog エントリから個々の Blog の書き手の行動経路を抽出する．その際に，我々は地名が指す場所におけるビジターの行動に着目することで各 Blog エントリからビジターの行動経路を抽出する．こうしてマイニングした代表的な経路に対して，システムはその経路の文脈であるキーワードを抽出する．このような，ビジターの経路及びそのコンテキストを提示するシステムは，ユーザの実空間における行動計画の立案を支援すると我々は考える．

# Extraction of Visitors' Typical Route and its Context from Local Blogs

Hiroshi Kori,† Shun Hattori,† Taro Tezuka,† Keishi Tajima†
and Katsumi Tanaka †

Recently, it is common that users release their sightseeing through internet by blog. Route that the user passed is often described in blog. Then, we propose a system that extracts visitors' typical geographical route and its context and shows on map collectively. Context of typical route is described by keywords which express the theme of the visitors. We extract typical route by representative sequential pattern mining method "PrefixSpan". Moreover, for mining the typical route, We extract each visitor's route for one blog entry.Then we focus on whether the visitor did some action at the place. Thus, for extracted typical route we extract the keyword which is the context of the route. This system supports user's plan to visit some place.

## 1. はじめに

人は，主に観光等を目的として，様々な場所を訪れる．その際に，どの場所をどのような順番で訪れるかという計画を立てることが多い．そのような場合に参考とするべき情報としてガイドブック等が存在するが，一方で近年のインターネットの普及に伴い，Web 上の情報を参考とすることも多い．このような情報の中には，どのような場所をどのような順番で訪れるかを紹介しているサイトも存在する[?),1)]．このようなサイトで紹介されているコースは，効率的に訪れるという点を考慮していたり，あるいは訪れる場所をその目的別に紹介するという工夫を行っているが，実際にその場所を訪れた人がどのような経路で，さらにどのようなテーマで訪れているか，あるいは最新の話題をなるべくリアルタイムに伝えることについては考慮していない．我々は，このような実際のビジターの行動及び地理的な最新の話題が，ユーザが実際に行動計画を立案する際の支援になると考え，それらを地図上にマッピングすることにより集約して提示するシステムを提案する．

我々は，このようなシステムを構築するために，地域 Blog の収集を行い，収集した各 Blog エントリから地名が指す場所でのビジターの行動に着目し，さらにエントリ中の地名の出現順序に基づき，Blog の書き手の行動経路を抽出する．このようにして得られた行動経路を地名の順序つきリストと考え，代表的なシーケンシャルパターンマイニング手法である PrefixSpan により頻出するシーケンシャルパターンを抽出する．

† 京都大学大学院情報学研究科
Graduate School of Infomatics, Kyoto University

これをビジターの代表的な行動経路と考え，それぞれの経路からユーザ行動のテーマを表すキーワードを抽出しそれを経路の文脈として，その行動経路とともに地図上にマッピングすることによりユーザに提示する．

本論文においては，まず 2 章において関連研究について述べる．そして，3 章においてユーザの代表的な行動経路及びそのコンテキスト抽出手法の詳細を説明する．4 章では抽出した行動経路をユーザに提示するシステムのインタフェースとシステムに対するユーザの問い合わせ手法について述べ，5 章でまとめと今後の課題について述べる．

## 2. 関 連 事 項

### 2.1 PrefixSpan

データマイニングにおける技術としてアイテムの組合せの頻出パターンを発見する技術が提案されている[2] が，一方で，アイテム間の順序を保ったままで頻出するパターンを発見するシーケンシャルパターンマイニングがいくつか提案されている[3)~7)]．我々は，その内で深さ優先探索で多頻度パターンを抽出する手法であり，非常に高速なマイニングが可能である PrefixSpan を利用し，頻出する地名のシーケンシャルパターンを抽出する．

### 2.2 関 連 研 究

Blog と地理情報の統合に関する研究及びサービスとして，上松ら[8] の「場 log」や検索サービスの maplog[9] が存在する．しかし，これらは複数の Blog から情報を集約して提示するものではない．我々のシステムでは，このような地域性を持った複数の Blog エントリを解析し，その結果を地図上に集約して提示することにより，新たな知識発見を試みている．

一方で，Hurst[10] らは Blog エントリを地名により地図に対してマッピングし，複数の Blog ホスティングサービス間の相違点の発見を試みている．また，倉島ら[11] は，Blog から人々の体験を相関ルールマイニングによって抽出することで，それらを集約して提示するシステムを提案している．本システムでは，Blog から人々の行動経路を抽出し，その中で代表的なものを地図上にマッピングして提示する．その際の行動経路の抽出において，動作動詞に着目するという点で倉島らの手法を参考にする．

## 3. 行動経路及びそのコンテキスト抽出手法

### 3.1 地域 Blog クローラ

頻出するシーケンシャルパターンを代表的なユーザの行動経路として抽出するためには，大量のシーケンシャルパターンを対象としてマイニングを行う必要がある．そのためには大量の地域 Blog を解析することが必要となるが，システムがユーザからの問い合わせを受信する度に大量の Blog を収集するのは現実的ではない．そこで，我々は地域 Blog を定期的に収集する地域 Blog クローラを作成した．この地域 Blog クローラは，人手で作成した地名リストをクエリとして既存の Blog 検索エンジンを利用して定期的に Blog の検索を行う．そして，検索結果である RSS を取得し，RSS における Blog 情報から { タイトル，URL，本文，投稿時刻 } の組をデータベースに蓄積する．

| | |
|---|---|
| Blog エントリの数 | 100 |
| 地名数 | 197 |
| ビジターが訪れた地名の割合 | 49% |

**表 1** 地域 Blog 予備実験

### 3.2 行動経路抽出手法

本節では，ひとつの Blog エントリに対してひとつの行動経路を抽出することを考え，地域 Blog クローラによって取得した各 Blog エントリから Blog の書き手の行動経路を取得する手法について説明する．Blog 中に出現する地名のすべてを実際に Blog の書き手が訪れたとは考えにくい．そこで，我々はまず Blog の本文中で出現する地名の内，どの程度 Blog の書き手が実際に訪れているかについての予備実験を行った．2005 年 5 月 22 日〜6 月 13 日の期間に収集した 7960 件の Blog エントリから無作為に 100 件の Blog エントリを取り出し，その中に含まれる地名の内，どれ程の割合で Blog の書き手が実際に地名の指す場所を訪れているかを調べた．その結果を表 1 に示す．実験の結果，実際には 49%の地名しか Blog の書き手は訪れていないことが分かった．そこで我々は，各地名が，実際にその場所を訪れたという文脈で利用されているかを判定する地名フィルタを作成する．このフィルタにより，ビジターが実際に訪れた地名群が取得される．この地名群は，シーケンシャルパターンの構成要素となる．その後，Blog 本文中の出現順序に基づき，その地名群から地名のシーケンシャルパターンを生成する．以下，経路（ルート）の構成要素となりうる，実際にビジターが訪れた地名のことを「ルート要素」と呼ぶ．

そこで，本システムではビジターの行動経路，すなわち各 Blog エントリに対する地名のシーケンシャルパターンを抽出するために以下の 2 つのフェーズを適用する．

**Step1:** 地名フィルタ

**Step2:** シーケンシャルパターン生成

以下，その詳細について説明する．

### 3.2.1 地名フィルタ

本フィルタでは，各地名が実際にビジターがその場所を訪れたという文脈で使用されているかを判定し，訪れていると判定された場合はその地名をルート要素とし，訪れていないと判定された場合はその地名を破棄する．その際に，我々はその地名の使用されている文脈が，その場所でビジターが何らかの行動を行ったと判断される場合に，その地名をルート要素と判断することとした．その判定を行うため，我々は倉島ら[11),12)]の手法を参考とし，「食べる」，「見る」等といった動作動詞と「到着」等といったサ変名詞及び格表現に着目した．また，我々は，格表現の深層格，文節同士の係り受け関係についても考慮した．係り受け解析については，CaboCha[13)] を，動作動詞辞書については日本語語彙大系[14)] を使用した．

「日本語における表層格と深層格の対応関係」[15)] では，表層格を格助詞そのものとし，深層格をチャールズ J・フィルモア[16)] の定めた要件にいくつかの要件を加えたものとして定義し，その対応関係に関する調査を行っている．その際に深層格として「場所」，「場所－始点」，「場所－終点」，「場所－経過」という場所に関する深層格を表しうる格助詞として以下を挙げている

「から」，「へ」，「まで」，「を」，「に」，「で」，「より」，「において」，「に対して」，「にたいして」

そこで，我々は Blog 本文中の地名を含む文に対して係り受け解析を行い，場所を表しうる格助詞が地名と同じ文節に現れ，なおかつその地名が動作動詞かサ変名詞に係っている場合はその地名をルート要素とする．このような文の例としては，「京都駅へ行く」「清水寺に到着」等が挙げられる．これは直接的に動作を表す文である．また，もうひとつのパターンとして，「到着したのは銀閣寺です」や「次に向かった清水寺は」等といった，間接的には動作を表すが，文全体では状態を表す文の抽出も試みる．これは，地名の含まれる文節に場所を表しうる格助詞が含まれていず，動作動詞が地名に係っているパターンである．以上抽出する 2 つのパターンをまとめると，以下のようになる．

**Pattern1:** {place'} ⇒ {verb}
**Pattern2:** {verb} ⇒ {place}
place' : 地名+場所を表しうる格助詞
place: 地名を含む，place' 以外の文節
verb: 動作動詞 OR サ変名詞

ただし，「⇒」は係り受け関係を表す．

また，以上の形で取得できないルート要素として，「金閣寺や銀閣寺へ行った」や「清水寺の塔頭にのぼった」等といった地名の含まれている文節が並立助詞そして並立関係を表す読点，連体助詞「の」を含んでいるパターンが考えられる．これらの並立助詞，読点，連体助詞「の」は「Pattern1」を拡張することにより判定する．「Pattern1」では，地名の含まれる文節の係り先しか見ていなかったが，並立及び連体助詞「の」のケースでは，さらに先の係り先も解析する．さらに先の係り先文節に，場所を表しうる格助詞が存在するならば，その文節が動作動詞に係っているかを判定し，係っていればルート要素として加える．「清水寺の塔頭にのぼった」という例で説明すると，「清水寺の」という文節の先に「塔頭に」という文節があり，格助詞「に」が場所を表す格助詞となりうるので，さらに先を解析し，「のぼる」という動作動詞が得られるため，清水寺はルート要素となる．もしも場所を表す格助詞と動作動詞といずれも含まれていないならば，「Pattern2」に当てはまるかの判定を行う．この文節が係っている文節においてもさらに { 並立助詞，読点，連体助詞「の」} が含まれる場合，すなわち「清水寺の塔頭の先へ」等という文の場合は，「塔頭の」という文節からさらに先の係り先を解析する．もしもその先にも場所を表しうる格助詞が存在しなければ，「Pattern2」に当てはまるかの判定を行う．存在すれば，それが動作動詞に係っているかどうかを判定する．こうして「行ったのは金閣寺と銀閣寺です」という文や「金閣寺と銀閣寺、清水寺へ行った」という文においても地名「金閣寺」をルート要素と判定することが可能となる．このパターンをまとめると，以下のようになる．ただし，(pattern)*は，pattern の任意の繰り返しを表す．

**Pattern1':** {place"} ⇒ ({block} ⇒)* {verb}
place" : 地名+（並立助詞 OR 読点 OR 連体助詞「の」）
block: （並立助詞 OR 読点 OR 連体助詞「の」）を含む文節
verb: 動作動詞 又は サ変名詞

ただし，連体助詞「の」の場合は，「地名 1 +の+地名 2」のような場合が存在する．たとえば，「東本願寺の渉成園を訪れた」という文の場合は，連体助詞「の」は「東本願寺」の中の「渉成園」という空間上の包含関係を表している．この場合，東本願寺と渉成園の両方をルート要素として加えるのではなく，より後方に

出現した「渉成園」だけをルート要素として加える．

以上をまとめると，Blog の本文中の各地名をルート要素に加えるかどうかの判定は，各々の地名を含む文節を以下のアルゴリズムに適用することにより得られる．

**Step1:** 文節が移動を表しうる格助詞を含むか判定
- 含まないなら Step2 へ．
- 含むなら係り先が動作動詞又はサ変名詞を含むか判定
  - 含むなら地名をルート要素へ加える
  - 含まないなら Step4 へ

**Step2:** 文節が並立助詞又は並立関係を表す読点を含むか判定
- 含まないなら Step3 へ
- 含むなら係り先の文節について Step1 へ

**Step3:** 文節が連体助詞「の」を含むか判定
- 含まないなら Step4 へ
- 含むなら係り先の文節について Step1 へ．ただし，係り先の文節にも地名が含まれている場合は，地名を破棄し終了．

**Step4:** 地名を含む文節に対して動作動詞が係っているか判定
- 係っているなら地名をルート要素へ加える
- 係っていないなら地名を破棄し終了

以上のアルゴリズムを Blog 本文中のすべての出現地名を含む文節に適用することにより，各 Blog エントリのルート要素を求める．

以上の手法を手法 1 とし，予備実験で利用した Blog に対して適用してみた．各 Blog エントリ内のすべての地名に対する地名フィルタの判定がすべて正解ならば正解とすることとし，各 Blog エントリに対して評価を行った．その判定の精度の結果が表 2 の手法 1，平均精度の結果が図 1 中の手法 1 の系列である．フィルタ無しという系列が，地名フィルタをかけずに出現地名すべてをルート要素とした場合のグラフである．ただし，この図では各 Blog エントリに含まれる地名数に応じた平均精度を算出している．ここで，誤った判定が行われている文を考察すると，主に以下の理由に起因していることが分かった．

- 動詞の省略（例：次は清水寺へ）
- 助詞の省略（例：京都駅到着）
- 動作動詞にのっていない動詞（例：四条烏丸で乗り換え）

これらの理由は，Blog というメディアがくだけた日

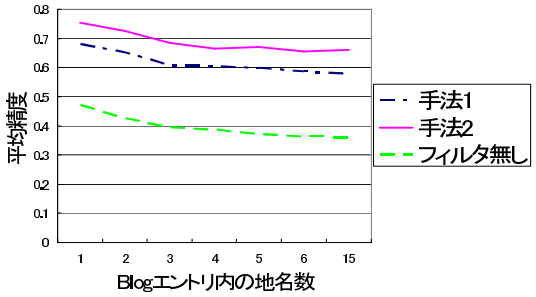


**図 1** 地名フィルタの平均精度

| エントリ内の地名数 | 1 | 2 | 3 | 4 | 5 | 6 | 15 |
|---|---|---|---|---|---|---|---|
| フィルタ無し | 47% | 33% | 11% | 25% | 0% | 0% | 0% |
| 手法 1 | 68% | 59% | 22% | 50% | 50% | 0% | 0% |
| 手法 2 | 74% | 67% | 33% | 25% | 75% | 0% | 100% |

**表 2** 地名フィルタ精度

本語で書かれているという理由による．そこで，我々はそれぞれに対して対策を講じた．まず，「動詞の省略」に対しては，「から」「まで」「へ」という深層格において特に「場所－始点」及び「場所－終点」を表すことが多い格助詞に対しては，その文節が出現した時点で Blog の書き手が実際にその場所を訪れたと判断し，ルート要素として加えることとした．また，助詞の省略については，地名が現れた文節において動作動詞もしくはサ変名詞が現れた時点でその地名をルート要素と判定する．次に動作動詞について述べる．動作動詞辞書には，各動詞が状態を表すか動作を表すかが記されているが，辞書に動詞自体が掲載されていない場合は，動作動詞と判定することとした．この手法を手法 2 とすると，その判定の精度の結果が表 2 の手法 2，その平均精度の結果が図 1 中の手法 2 の系列である．このように，くだけた文章に対する対策を行うことで精度を改善することが可能となった．

以上の手法で誤った判定を行う例として，以下の例が挙げられる．

- 否定が含まれているもの
  （例：銀閣寺には行かずに）
- 時制が未来
  （例：修学旅行では、金閣寺へ行きます）
- 主語が Blog の書き手ではない
  （例：花子は銀閣寺へ行ったらしいが）
- 主語が Blog の書き手ではない
  （例：花子は銀閣寺へ行ったらしいが）
- 助詞「も」が助詞「を」の意味を表す
  （例：金閣寺も行きました）
- 動作動詞も助詞「へ」「から」「まで」も存在しない

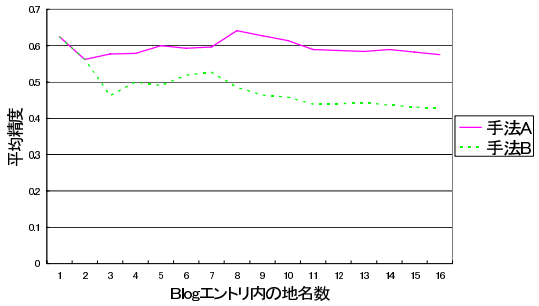


**図 2** 経路生成の平均精度

| 経路 | 出現頻度 |
|---|---|
| 下鴨神社→上賀茂神社 | 24 |
| 嵐山→渡月橋 | 24 |
| 京都御所→下鴨神社 | 17 |
| 阪急→河原町 | 17 |
| 京都御所→上賀茂神社 | 16 |
| 金閣寺→銀閣寺 | 15 |
| 金閣寺→清水寺 | 14 |
| 京都御所→下鴨神社→上賀茂神社 | 14 |
| 三条→四条 | 16 |
| 清水寺→三十三間堂 | 13 |

**表 3** 抽出した代表的経路

（例：次は、銀閣寺です）

- タイトルがビジターの訪れた地名である

#### 3.2.2 シーケンシャルパターン生成

3.2.1 節の地名フィルタにより取得した各 Blog エントリのルート要素に順序付けを行い，地名によるシーケンシャルパターンを生成する．その際に利用する情報が Blog 本文中の地名の出現順序である．Blog の書き手は，自分の行動を日記風に書くことが多いため，その本文中の地名の出現順序が訪れた場所の順序を反映している場合が多い．したがって，我々は Blog 本文中の地名の出現順序を利用する．その際に問題となるのが，複数の地名が入り組んで抽出される場合である．Blog 本文中のルート要素をその出現順序で並べた場合に { 清水寺→金閣寺→銀閣寺 } となっている場合は良いが，実際には { 清水寺→金閣寺→清水寺→銀閣寺 } といった例では清水寺を 2 度訪れたとは考えにくい．特に，Blog の書き手が一度行った場所を回想している場合や，実際にその場所に着く前に向かう目的地について述べている場合にこのようなパターンが出現する．そこで，同じ地名が 2 回以上出現した場合は，最も後の方の順序を，実際に訪れたと定めるルート作成法を考える．このルート作成手法を手法 A，ルート要素を単に出現順序順に並べたルート作成手法を手法 B とし，この 2 つを最初の予備実験で用いたデータに適用して比較してみたところ，まったく同じルートがすべてのエントリに対して得られた．これは，各エントリの有する地名数が少なく，入り組んだ地名の出現パターンがほとんど見られなかったためである．そこで，今度は対象データを，ルート要素を 4 つ以上含む Blog エントリとし，手法 A と手法 B を比較した．その結果を図 2 に示す．このように，精度において手法 A のほうが手法 B よりも高い精度が得られていることが分かる．また，手法 A では抽出できないが手法 B なら抽出できるような，同じ場所を二回訪れている Blog エントリもいくつか存在した．その際に二度以上訪れられる場所というのは,「京都駅」や「四条」といった交通の要所であることが多かった．

### 3.3 代表的な行動経路の抽出

3.2.2 節の手法により抽出した各 Blog エントリのシーケンシャルパターンに対して，最低アイテム数を 2，最低サポート値を 2 とした PrefixSpan を適用することにより頻出する地名のシーケンシャルパターンを抽出し，それらを代表的なビジターの行動経路とする．

実際に抽出されたすべての経路の内で最も頻度が高かった 10 個の経路を表 3 に示す．対象データは，2005 年 5 月 22 日〜6 月 13 日の期間に収集した 7960 件の Blog エントリである．ただし，その内 2 つ以上の京都の地名を含むもの，すなわち経路情報を有するものは，1126 件である．「京都御所→下賀茂神社→上賀茂神社」という経路に類する経路が比較的多く検出されているのは，葵祭りが行われた影響である．

### 3.4 コンテキスト抽出

本節では，各経路におけるコンテキストの抽出手法について述べる．本論文では，経路のコンテキストをその経路を通った人が共通に持つテーマであると考え，それをキーワードの形で抽出する．そのため，指定した経路を含む Blog エントリをまず抽出する．その上で各 Blog エントリをひとつの文書と考え，ベクトルの各次元を一般名詞の有無による {0,1} で表した特徴ベクトル $V_i$ を各エントリ $E_i$ に対して作成する．その上で，経路 $r$ に対する特徴ベクトル $V(r)$ を以下のように定める．ただし，$n$ は経路 $r$ を含む Blog エントリの総数である．

$$V(r) = \frac{\sum_{i=1}^{n} V_i}{n} \quad (1)$$

このようにして求めた特徴ベクトル $V(r)$ の上位 $m$ 個を経路 $r$ におけるコンテキストとする．こうして，その経路を含む Blog の中で幅広く使用されている単語が取得でき，これをその経路のコンテキストとする．また，それぞれのコンテキストの特徴ベクトル $V(r)$ における値を，その経路におけるコンテキスト，すな

| 経路 | 出現頻度 | コンテキスト（文脈度） |
|---|---|---|
| 下鴨神社→上賀茂神社 | 24 | 葵祭 (0.8), 祭り (0.5) |
| 嵐山→渡月橋 | 24 | 人 (0.4), 寺 (0.3) |
| 阪急→河原町 | 17 | 人 (0.5), 電車 (0.3), 店 (0.3) |
| 金閣寺→銀閣寺 | 15 | 修学旅行 (0.6), 班 (0.5) |
| 金閣寺→清水寺 | 14 | 修学旅行 (0.9), 班 (0.6) |
| 三条→四条 | 16 | 人 (0.5), 靴 (0.4) |
| 詩仙堂→曼殊院 | 5 | 人 (0.6), 庭 (0.6), 枯山水 (0.4) |
| 大徳寺→今宮神社 | 2 | 餅 (0.5), 白味噌 (0.5), 手 (0.5) |
| 二条城→神泉苑 | 2 | 外国 (0.5), 一般 (0.5), 庭園 (0.5) |

**表 4**　抽出したコンテキスト

わち文脈の強さを表すとして以降,「文脈度」とする.

表 4 に実際に抽出したコンテキストを提示する．上方に提示しているのがメジャーな経路におけるコンテキストである．このような経路の場合はコンテキストとしてふさわしいものがとれていると思われるが，その結果は予想の出来るもので，そのコンテキストをユーザが閲覧したからといって新たな発見があるとは考えにくい．一方で，下方に提示しているマイナーな経路におけるコンテキストは，例えば「大徳寺→今宮神社」という経路における「餅」と「白味噌」や,「詩仙堂→曼殊院」という経路における「庭」と「枯山水」等はユーザが閲覧して新たな発見があるコンテキストであると考えられる．けれどもこういったコンテキストは，現時点では他のノイズに埋もれてしまっている状態であり，今後は，こういったコンテキストを特に抽出することを試みたい．また，コンテキストの中には，経路のコンテキストというよりは，経路を構成する地名特有のコンテキストと考えられるものも抽出されているため，経路のコンテキストと地名のコンテキストの切り分けが必要である．

## 4. システム概要

### 4.1 インタフェース

システムのインタフェースとしては地図インタフェースを採用する．このようにして，ユーザはシステムにより抽出された経路を地図上で視覚化された形で閲覧できる．経路はルート要素間のラインで表現され，その経路の出現頻度が大きくなるほどそのラインが太くなるように地図上で描画される．また，地図上の各ルート要素にそのルートを含む Blog やその Blog で取り上げられている画像を貼り付けることも考える．このようにして，ユーザは経路から閲覧する Blog を絞り込むことが出来る．

### 4.2 システム操作

#### 4.2.1 表示エリアの切り替え

地図の表示エリアを切り替えることで，表示する経路に対しても切り替えを行う．表示する候補となる経路は，構成するルート要素がすべて表示エリア内に存


**図 3**　システムイメージ


**図 4**　経路の地図マッピング

在する経路であるとする．表示候補の経路集合 $R$ は，以下のように表される．ただし，$E(r)$ は経路 $r$ のルート要素の集合を，$e\ in\ box$ はルート要素 $e$ が表示エリア $box$ 内に存在することを示す．

$$R = \{r \mid \forall e\ in\ box \wedge e \in E(r)\} \qquad (2)$$

このような経路集合 $R$ の内，その出現頻度が高いものから順に $p$ 個を地図上にマッピングする．

実際に経路を地図上にマッピングしたものが図 4 である．左の図が広域図で，抽出された全経路の中で最も出現頻度の高い経路が 10 個マッピングされている．右の図が詳細図で，清水寺周辺のエリアを表示している．表示されている経路は，経路全体が表示エリアに

含まれているものの中で最も出現頻度が高い 10 個の経路であり,「清水寺→祇園」という経路や「清水寺→地主神社」という経路が地図上にマッピングされている.

#### 4.2.2 経路の検索

図 3 のシステムイメージにもある通り，ユーザはブラウザ上のフォームから経路を検索することも出来る．経路を構成するルート要素，ルート抽出の対象となる Blog の投稿期間，そして．コンテキストによる検索を可能とする．ルート要素と期間を指定した検索の検索結果は，それらの条件を満たす経路の中でその出現頻度が高い順に $p$ 個であり，それらを地図上にマッピングし，表示する．一方で，コンテキストによる検索は，文脈度による検索となる．期間やルート要素の条件を満たす Blog の中で指定したコンテキストの文脈度が高い経路を高い順に $p$ 個マッピングし，表示する．

## 5. まとめと今後の課題

我々は，本論文で Blog からユーザの代表的な行動経路とその文脈を抽出し，それらを地理上にマッピングすることにより集約して提示するシステムを提案した．ユーザの代表的な行動経路については，個々の Blog エントリから，地名が指す場所におけるビジターの行動に着目することで各ユーザの行動経路を抽出し，抽出した経路に対して，評価を行った．さらに，抽出した経路の中で代表的なものを PrefixSpan を用いて抽出を行い，その経路のコンテキストを表すキーワードを抽出する手法について述べた．そして，最後にシステムのインタフェースとそれに対する問い合わせ手法について述べた．今後は，より詳細なコンテキストの抽出，システムの実装と全国版の作成，そして，より精度の高いユーザの行動経路抽出に対して取り組んでいく．また，システムの応用例として，Web ブラウザを通してユーザがシステムを利用するだけでなく，ユーザが実際に実空間においてカーナビ等を通してシステムを利用することも考えていきたい．システムの応用例として，例えば，ユーザの実空間での移動履歴に基づいた目的地推薦等が考えられる．筆者らは，BlogCarRadio システムという，実空間において地域 Blog をラジオのように，音声により視聴するシステムを提案している[17] が，それに対する応用などについても考えていきたい．

## 謝　　辞

本研究の一部は，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー：田中克己，平成 14〜18 年度）、文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己），文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，および，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者：田中克己，A01-00-02, 課題番号：18049041）によるものです．ここに記して謝意を表すものとします．

## 参 考 文 献

1) おすすめ京都散策コース
`http://www.kyotokanko.com/osusume.html`.
2) R. Agrawal and R. Srikant: “Fast algorithms for mining association rules”, Proc. 20th Int. Conf. Very Large Data Bases, VLDB (Eds. by J.B. Bocca, M. Jarke and C. Zaniolo), Morgan Kaufmann, pp. 487–499 (1994).
3) R. Agrawal and R. Srikant: “Mining sequential patterns”, Eleventh International Conference on Data Engineering (Eds.by P.S. Yu and A.S.P. Chen), Taipei, Taiwan, IEEE Computer Society Press, pp. 3–14 (1995).
4) R.Srikant and R.Agrawal: “Mining sequential patterns: Generalizations and performance improvements”, Proc. 5th Int. Conf. Extending Database Technology, EDBT (Eds. by P.M.G. Apers, M. Bouzeghoub and G. Gardarin), Vol. 1057, Springer-Verlag, pp. 3–17 (1996).
5) M.J. Zaki: “SPADE: An efficient algorithm for mining frequent sequences”, Machine Learning, **42**, 1/2, pp. 31–60 (2001).
6) J.Ayres, J.Flannick, J.Gehrke and T.Yiu: “Sequential pattern mining using a bitmap representation”, Proc. of SIGKDD ’02, pp. 429–435 (2002).
7) J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q.Chen, U.Dayal and M.-C. Hsu: “PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth”, pp. 215–226.
8) 上松大輝, 沼晃介, 徳永徹郎, 大向一輝, 武田英明：“場 log：weblog 環境における位置情報利用の提案”, 第 6 回人工知能学会セマンティック Web とオントロジー研究会 (2004).
9) maplog
`http://maplog.jp/`.
10) M.Hurst: “Gis and the blogosphere”, WWW2005, 2nd Annual Workshop on the Blogging Ecosystem:Aggregation, Analysis and Dynamics (2005).

11) T.Kurashima, T.Tezuka and K.Tanaka: “Blog map of experience: Extracting and geographically mapping visitor experiences from urban blogs”, Proceedings 6th Web Information Systems Engineering(WISE2005), pp. 496–503 (2005).
12) 倉島健, 手塚太郎, 田中克己 : “街 blog からの体験抽出とその空間的提示手法の提案”, 第 16 回 データ工学ワークショップ (DEWS2005) (2005).
13) CaboCha `http://chasen.org/~taku/software/cabocha/`.
14) 日本語語彙大系 `http://www.ntt-tec.jp/technology/C404.html`.
15) “日本語における表層格と深層格の対応関係”, 三省堂.
16) “格文法の原理—言語の意味と構造—”, 三省堂.
17) H.Kori, T.Tezuka and K.Tanaka: “Ranking of regional blogs by suitability for sonification”, Proceedings The 2nd International Special Workshop on Databases for Next-Generation Researchers (SWOD2006)(in conjunction with ICDE2006) (2006).