Vol.2011-DBS-152 No.9
Vol.2011-IFAT-103 No.9
2011/8/3

クエリログを用いた地理的検索対象の時空間分析

風　間　一　洋[†1]

地理情報検索において，ユーザの地理的検索対象が時間と共にどのように変動しているかについて調べる．そこで，商用サーチエンジンのクエリログから，東京の首都圏の各地名と共起語を抽出し，頻度の時間的・空間的な変動を分析した．その結果，地理的クエリは一般的なクエリとは異なる時間的変動を示し，現実の人間の行動の影響があること，地名はそれぞれ固有の時間的変動パターンを持ち，共起語集合がその地域の特徴を示すことがわかった．さらに，地図上に特に頻繁に検索されている地名を可視化して．注目されている地名の分散や移動パターンに，現実の人間の行動との関連性があることを示した．

Space-Time Analysis of Geographical Search Targets using Query Logs

Kazuhiro Kazama[†1]

I examined the spatial and time variation of users' geographical search targets in geographical searches. I extracted geographical names of Tokyo metropolitan district and their cooccurrence terms from commercial search engine query logs and analyzed the spatial and time variation of their frequency. I found that the variation of geographical queries was different from one of general queries and it was influenced by human activities in the real world. Each geographical name had its own characteristics and its cooccurrence terms showed regional features. Furthermore, I displayed frequently-searched geographical names on the map. I showed that their distribution and moving patterns were related to real human movement.

†1 日本電信電話 (株) NTT 未来ねっと研究所
NTT Network Innovation Laboratories, Nippon Telegraph and Telephone Corporation

1. はじめに

近年，リモートセンシングや GPS，レーザスキャナなどの空間データ作成のための測定技術の進歩と共に，カーナビゲーションや携帯電話などの測定機能を持つ機器も小型化・普及してきた．このことにより，継続的に蓄積されたデータの時空間分析がさかんにおこなわれるようになった[1)]．このように位置情報を直接追跡できない場合でも，携帯電話の基地局や鉄道の改札などのサービスの利用履歴から位置情報を推定することができる．

本稿では，現実世界の地理的検索対象に対するサーチエンジンのクエリは地域性や局所性を持つことが多いことから，本稿ではサーチエンジンのクエリログを用いる．地理的クエリを分析することで，いつどの地域の地理的検索対象が注目されているかを知ることができる．

さらに，地理情報検索では，直後にユーザの実際の行動を伴うことが多いことから，ユーザの生活・行動パターンの影響を強く受ける傾向がある．例えば，一般にクエリ数は昼間が多く早朝が少ないという周期的変動を持つが，実際の人間の行動のきっかけになることが多い地理的クエリはさらに強く人間の生活・行動パターンの影響を受けるはずであり，各地域に存在する地理的検索対象の違いにより地域ごとに異なる特徴を示すと考えられる．さらに，直後にユーザの行動を伴う地理的クエリだけを抽出できれば，現実世界におけるユーザの移動を推定することもできる．

そこで本稿では，地理的クエリにおける東京の首都圏の各地名とその共起語を抽出し，頻度の時間的・空間的な変動から，地域固有の特徴と現実の人間の行動との関連性を分析する．さらに，地図上に特に頻繁に検索されている地名を可視化して．注目されている地名の分散や移動について調べ，クエリログを用いた時空間分析の可能性について考察する．

2. 地理的検索対象の時空間分析

2.1 地理情報検索とユーザの行動の関連

ある地域に関する情報を探すためには，地理情報検索が用いられる．一口に地理情報検索と言っても，地図検索や地域を階層的なカテゴリーに分割したディレクトリなどさまざまなユーザインタフェースが提供されているが，どれも座標や地名などの形で探索地域を指定する．このように地理情報検索で指定した地域の集合は，そのユーザの居住地や勤務地などの行動範囲や，仕事や遊びなどの行動目的を反映することになる．

また，ユーザが地理情報検索をおこなう時には必ず何らかのきっかけがあり，検索した後

Vol.2011-DBS-152 No.9
Vol.2011-IFAT-103 No.9
2011/8/3

でそれに基づく実際の行動を起こすことも多い．すなわち，地理増俸検索の時刻は，ユーザの生活・行動パターンと密接な関係がある．

2.2 クエリログと時空間分析

サーチエンジンでも，クエリを入力するだけで，地理的検索対象を検索できる．ただし，地理的検索対象を入力するだけでは，ユーザの想定していない地域の情報も検索結果に含まれてしまうことが多いので，目的の探索地域に絞り込むために，例えば「映画　渋谷」のような検索対象を表す単語と地名を組み合わせた形式のクエリがよく用いられる．

ユーザがサーチエンジンの検索に用いたすべてのクエリはクエリログに記録されるので，クエリログを分析すれば，ユーザが用いた地名集合や検索時刻を知ることが可能であり，時間的・空間的な分析を行うことが可能になる．

3. 関 連 研 究

ユーザの使用履歴情報を用いて時空間分析をおこなった事例がいくつか存在する．まず，NTT ドコモのモバイル空間統計では，携帯電話基地局の管理情報を用いて，携帯電話の各時刻における所属エリアと顧客情報を組み合わせることで，地域ごとの人口分布，人工構成，移動に関する統計情報を収集し，街作りや防災計画などの研究目的に提供されている[2),3)]．また，杉山らは，鉄道の改札通過データを用いて駅間の旅客流動を分析した[4)]．これらの研究は，携帯電話基地局や自動改札機などの実世界における位置が固定している機器を用いて，現実の人間の移動を観測している．これに対して，本研究は，検索・注目されている地理的検索対象の時間的・空間的変化を分析するものである．実世界におけるユーザの位置も地名から得られる緯度・経度情報から判定できるが，地理情報検索は必ずしもその後に人間の移動を伴うとは限らないし，移動するまでの遅延も存在することから，これらの研究のように人間の移動を観測することは困難である．

4. 地理的検索対象の時空間データの抽出

4.1 対象地域と地名

本稿では，JR 東日本の京浜東北線，京葉線，武蔵野線，南武線と私鉄の各路線で囲まれた範囲の地域を対象として，駅名を地名として用いた．この地域には，JR 東日本，東武鉄道，西武鉄道，小田急電鉄，京王電鉄，東急電鉄，京成電鉄，新京成電鉄，東京メトロ，東京都交通局，首都圏新都市鉄道，埼玉高速鉄道，東京臨海高速鉄道，東京モノレール，多摩モノレール，ゆりかもめなどの多くの企業のおかげで鉄道網が発達し，首都圏の主な移動手段として活用されている．駅名を地名として用いた理由は，「新宿で待ち合わせ」，「吉祥寺で買物」など，日常生活で駅名で場所を指定する場合が多いからである．

図 1　対象地域の鉄道路線と駅

実際には，環状領域内の駅と接続関係を定義した路線定義ファイルを読み込んで，地名と地名間接続を抽出する．この路線定義ファイルは，「A - B - C」のように各路線の駅とその接続関係をハイフンで接続して表したものである．さらに，「A = B」のように地名の正規化規則を定義することもできる．例えば，仲御徒町と上野広小路のように乗り換えが可能な駅や，「霞ヶ関」と「霞ケ関」，「四ッ谷」と「四ツ谷」のような揺れが存在する駅名の場合には，正規化して一つの駅と見なす．正規化後の地名数は 616 である．

今回の対象地域内の鉄道路線と駅の位置を**図 1** に示す．各駅の座標は緯度・経度から求めたものであるが，簡略化のために駅間は直線で描画している，また，異なる路線でも乗り換え可能な駅は，同一の駅として扱っている．

4.2 地理情報検索クエリの抽出

次に，商用サーチエンジンの 2008 年 2 月から 2010 年 7 月までのクエリログを用いて，地域情報検索クエリを抽出する．クエリログとは，ユーザが検索した日付と時刻，ユーザ ID，検索結果数，入力したクエリを記録したファイルの集合である．

地名を含むすべてのクエリを，地理的クエリとして抽出した．実際には，路線定義ファイルで指定したすべての駅名と正規化規則で定義した表記の揺れと，これらの地名の末尾に「駅」を付加した形式（例，「吉祥寺駅」）を検索語として使用した場合であり，今回の対象地

Vol.2011-DBS-152 No.9
Vol.2011-IFAT-103 No.9
2011/8/3

域以外の地名は対象にしていない．この時点で，クエリを，検索対象の地域を示す地名と，検索対象を示す共起語に分類する．ただし，今回は対象地域が限られるので,「大阪」などの一部の単語は地名として認識されない．

クエリログは複数のファイルに分割されているために，4 コア・2CPU の Mac Pro 上で並列に抽出処理をおこなった．この前処理をおこなうことで，以後で処理するデータサイズを小さくし，高速化できる．

4.3 時空間データの抽出

時空間データは以下のように抽出する．

まず，得られた地名・共起語とその検索時刻の組から，地名・共起語に対して 1 時間単位の検索頻度を調べる．1 日ごとに処理すると，地名によっては必ずしも充分な検索頻度が得られないことがあるので，ユーザの検索頻度には一日単位の周期性がある点に着目して，全期間をまとめて処理した．この結果，地名を $p_i(0 \leq i < N)$，地名数を N，時刻を $t_j(0 \leq j < 24)$ として，地名 p_i の時刻 t_j の検索頻度を f_{ij} とする検索頻度行列 F が得られる．

$$F = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1j} \\ f_{21} & f_{22} & \dots & f_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ f_{i1} & f_{i2} & \dots & f_{ij} \end{pmatrix} \tag{1}$$

ただし，地名によって検索頻度は大きく異なるために，そのままでは時間的変動をうまく比較することはできない．地名の各時刻の検索頻度特性を比較できるように全時刻の頻度の総和が 1 になるように値を正規化し，次のような検索頻度率行列 N を求める．n_ij は 0 と 1 の間の値を持ち，地名 p_i の時刻 t の検索頻度が総検索頻度に占める割合を示す．

$$N = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1j} \\ n_{21} & n_{22} & \dots & n_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ n_{i1} & n_{i2} & \dots & n_{ij} \end{pmatrix} \tag{2}$$

$$where\ n_{ij} = \frac{f_{ij}}{\sum_{k=0}^{23} f_{ik}}. \tag{3}$$

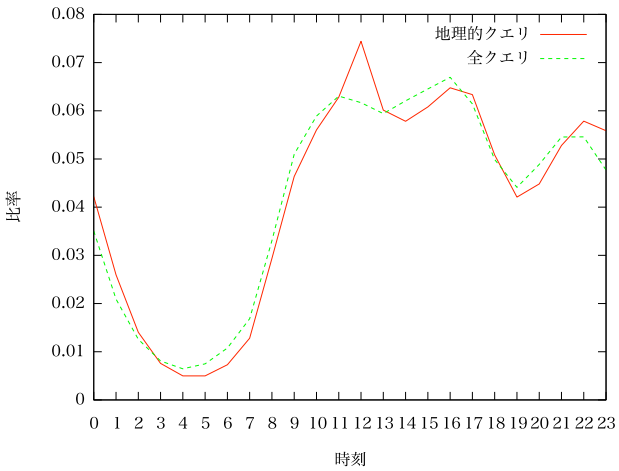


図 2 クエリ頻度の時間変動

5. 時空間データの分析

5.1 地理的クエリ頻度の時間変動

まず最初に，地理的クエリの検索頻度がどのような時間変動を示すかを知るために，今回の対象区域内の地名を含むクエリの 1 時間ごとの検索頻度を調べた．**図 2** に，本手法で求めた地理的クエリと全クエリの検索頻度の時間変動を示す．x 軸は時刻であり，y 軸はその時刻におけるクエリの検索頻度が全体に占める割合である．この結果から，地理的クエリは 12 時台，16 時台，22 時台に検索頻度のピークを持ち，12〜14 時，17〜19 時，22〜3 時の間が地理的クエリの割合が増えていることがわかる．

5.2 共起語順位の時間変動

次に，検索意図の時間的な変動を知るために，各時刻ごとに地理的クエリにおける地名の共起語の検索頻度を計算し，検索頻度が多い方から共起語を 6 語を選び，その順位の時間変動を調べた．この結果を**図 3** に示す.「ランチ」が 11〜13 時,「ラーメン」が 11 時台と 17〜19 時,「居酒屋」が 17〜19 時にピークを持つことから，地理的クエリは昼休みや退社時間などの生活・行動パターンの影響を強く受けていると考えられる．図 2 における昼と夕方の相対的な増加にも，これらの共起語が関係していると思われる．さらに，これらの共起語で

Vol.2011-DBS-152 No.9
Vol.2011-IFAT-103 No.9
2011/8/3

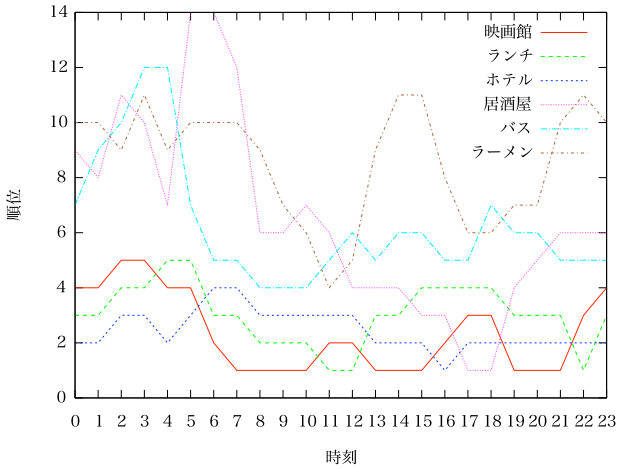


図 3 検索語頻度順位の時間変動

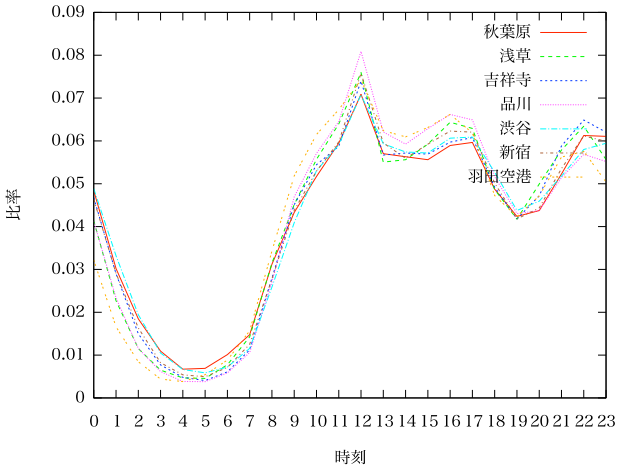


図 4 地名検索頻度の時間変動

は実際の行動との遅延は 1 時間ほどで，地域情報検索のリアルタイム性はかなり高いと考えられる．

これに対して,「映画館」,「バス」などの共起語は，一般的な生活時間帯に安定して順位を保っている．これも人間の生活・行動パターンに関連していると思われるが，遅延時間を推測するのは難しい．

さらに,「ランチ」が 22 時台にピークをもつことや,「ホテル」が実際に宿泊するかなり前の時刻から順位が高いことから，例えば帰宅してからじっくり地理情報を調べるとか，前もって調べてから予約しておくなど，遅延がかなり大きい場合もあると思われる．

なお，図 3 に示した共起語は，検索頻度を上位に絞り込んだことから，必然的に一般的な単語が抽出されている．地名の詳しい特性を明らかにするためには，より特殊な単語に対する詳細な分析が必要である．

5.3 地名検索頻度率の時間変動

次に，各地名を用いたクエリの検索頻度がどのような時間変動を示すかを調べた．**図 4** に，N から得られる各地名の検索頻度率を示す．x 軸が時刻，y 軸が各時刻の検索頻度率である．なお「羽田空港国内線ターミナル」は「羽田空港」と記述しているが，どちらも同様に扱っている．

この図を見ると，地名によって検索頻度率は異なる特性を示すことがわかる．例えば,「品川」は昼間の検索頻度が多く,「秋葉原」は逆に夜間の検索頻度の方が多い傾向を示す．

検索頻度率の時間変動を詳細に分析するために，図 2 に示した地理的クエリの検索頻度から，各時刻 j の検索頻度を全時刻の総和が 1 になるように正規化したベクトルを $\bar{n} = \{\bar{n}_1, \bar{n}_2, \ldots, \bar{n}_{23}\}$ として，各時刻の地名の検索頻度差行列 D を，次のように求める．

$$D = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1j} \\ d_{21} & d_{22} & \ldots & d_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & \ldots & d_{ij} \end{pmatrix} \tag{4}$$

$$where\ d_{ij} = n_{ij} - \bar{n}_j \tag{5}$$

図 5 に，D から得られる各地名の検索頻度差を示す．x 軸が時刻，y 軸が各時刻の全地名クエリとの検索頻度率の差分である．これから，検索頻度が相対的に増える期間が昼間か夜間かという違いだけでなく，相対的な検索頻度差の頂点がどの時刻になるかなどの違いは，地名によって異なることがわかる．

さらに，地名の検索頻度差の頂点時刻の傾向を把握するために，各地名 p_i に対して，最

Vol.2011-DBS-152 No.9
Vol.2011-IFAT-103 No.9
2011/8/3

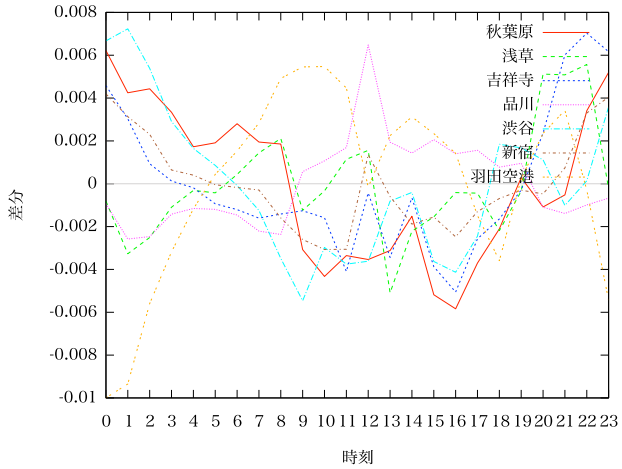


図 5 地名検索頻度差の時間変動

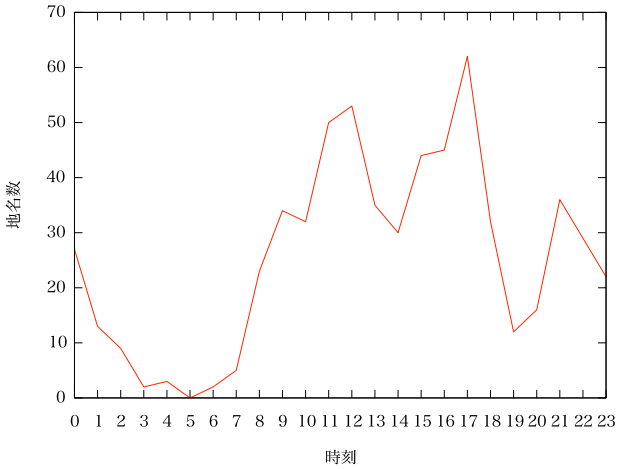


図 6 地名検索頻度差のピーク時刻

大の差分を示す時刻を調べて，そのヒストグラムを**図 6** に示す．これから，12 時付近，17 時付近，21 時付近に地名検索頻度差のピークを持つ地名が多いことがわかる．

5.4 各地名の共起語の分析

各地域の特徴を知るために，地名の共起語を分析する．いくつかの地名に対して得られた共起語の例を**表 1** に示す．左から頻度が高い順に並べているが，公序良俗に反するいくつかの検索語は除去している．

図 3 に示した頻度上位の一般的な共起語に加えて，より特殊で地域特有の共起語が多く見られる．例えば，ヨドバシカメラ，伊勢丹，高島屋などの存在巨大商業施設や，秋葉原は通り魔事件関連，浅草は観光関連，羽田空港は交通関連の検索語が多い．

共起語がその地域の時間的特性に直接関係していることも推測できる．例えば「ランチ」という共起語は吉祥寺，品川，渋谷，新宿で，「カフェ」という共起語は吉祥寺と渋谷で上位に来るが，前者は 12 時台のピークに，後者は 14 時台のピークに関連があり，渋谷では前者のピークが 1 時間遅れて後者と融合した形式になっていると推測される．なお，単に検索回数が多い共起語を抽出しただけなので，例えば品川の検索語の「ブログ」が，漫才師である品川庄司の「品川 blog」を示すように，誤抽出も若干存在する．

表 1 各地名の共起語

地名	検索語
秋葉原	メイド喫茶，通り魔，ヨドバシカメラ，事件，中古，販売，パソコン，フィギュア
浅草	ヨシカミ，大黒家天麩羅，ひょうたん，モンブラン，江戸もんじゃ，神谷バー，観光，グルメ
吉祥寺	東京都民が住んでみたい街，映画館，ランチ，伊勢丹，カフェ，東急，バス，由来
品川	ブログ，ホテル，水族館，ランチ，居酒屋，アトレ，映画館，グルメ
渋谷	映画館，ランチ，ホテル，居酒屋，カフェ，東急，映画，若者
新宿	映画館，ランチ，ホテル，伊勢丹，居酒屋，高速バス，高島屋，ルミネ
羽田空港	駐車場，バス，リムジンバス，みち子，浜焼き鯖寿司，高速バス，時刻表，国際線

5.5 地理的検索対象の時間変動の可視化

ユーザがサーチエンジンで検索する地理的検索対象は，1 日の間に大きく変動する．これを調べるために，地名検索頻度をそのまま用いると特定の地名だけが高く評価されてしまうし，地名検索頻度を正規化した地名検索頻度率を使用しても 1 日の変動の影響が大きく，うまく分析できない．そこで，地名検索頻度率の偏差値を用いて，各時刻にどの地域の地理的検索対象が他よりも注目されているかどうかを調べる．ここで，地名 p_i の時刻 j の偏差値 T_{ij} は，算術平均 μ_j と標準偏差 σ_j から次のように計算される．

Vol.2011-DBS-152 No.9
Vol.2011-IFAT-103 No.9
2011/8/3

$$T_{ij} = \frac{10(d_{ij} - \mu_j)}{\sigma_j} + 50 \tag{6}$$

$$\mu_j = \frac{1}{N}\sum_{i=0}^{N-1} d_{ij} \tag{7}$$

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(d_{ij} - \mu_j)^2} \tag{8}$$

各時刻ごとにすべての地名に対して地名検索頻度率の偏差値を計算し，その値が 55 を超える地名を図 1 で示した路線図上に可視化した．ただし，調査期間中の地名検索頻度が 1,000 を超えない地名については，ユーザから注目されているとは言いがたく，ノイズの原因になりやすいので除外した．この結果を**図??**に示す．

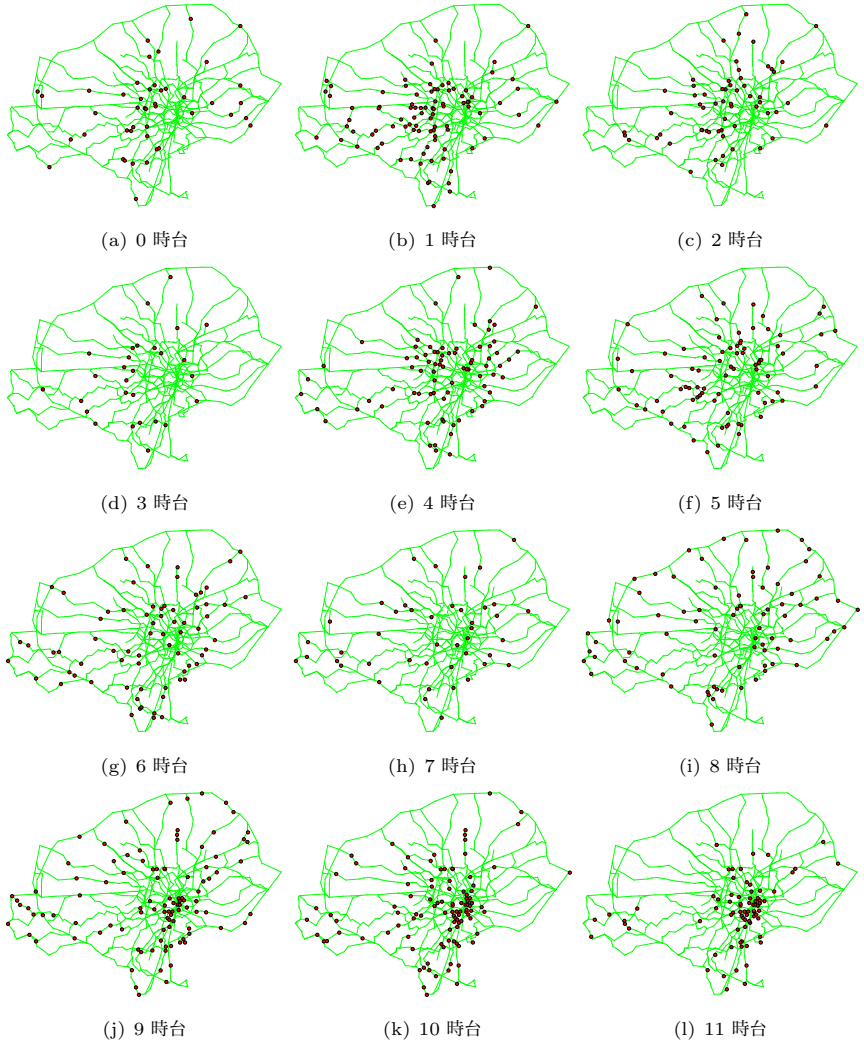
図 6 の各ピーク付近が多くの地域が注目されている時刻ということになるが，地名の分散パターンはそれぞれ異なる．8 時台には対象領域内が広く注目されているが，それが時間の経過と共に東京駅周辺に集中してくる現象が観察できる．また，16 時台は特に山手線の内側が注目されているが，20〜23 時台には山手線の外側から周辺部までに移行している現象が観察できる．

6. 考　　察

本稿の分析から，地理的クエリは一般的なクエリとは異なる時間的変動を示し，人間の行動と関係があること，地名はそれぞれ固有の時間的変動を持ち，地理的クエリで地名と同時に指定される検索語集合は地域固有になることがわかった．さらに可視化結果から，比較的近い位置に存在する地名間には似たような時空間特性があると共に，ユーザの地理的検索対象の存在する地域は，時間と共に移動していくことがわかった．

ただし，クエリログを用いた時空間分析は，GPS のように直接測定対象の位置を計測する手法や，携帯電話の基地局や鉄道の改札のように測定機器の位置が確定している手法とは違い，厳密に測定された位置情報を使う代わりに地名が示す位置を用いるために，解決しなければならない課題がいくつか存在する．

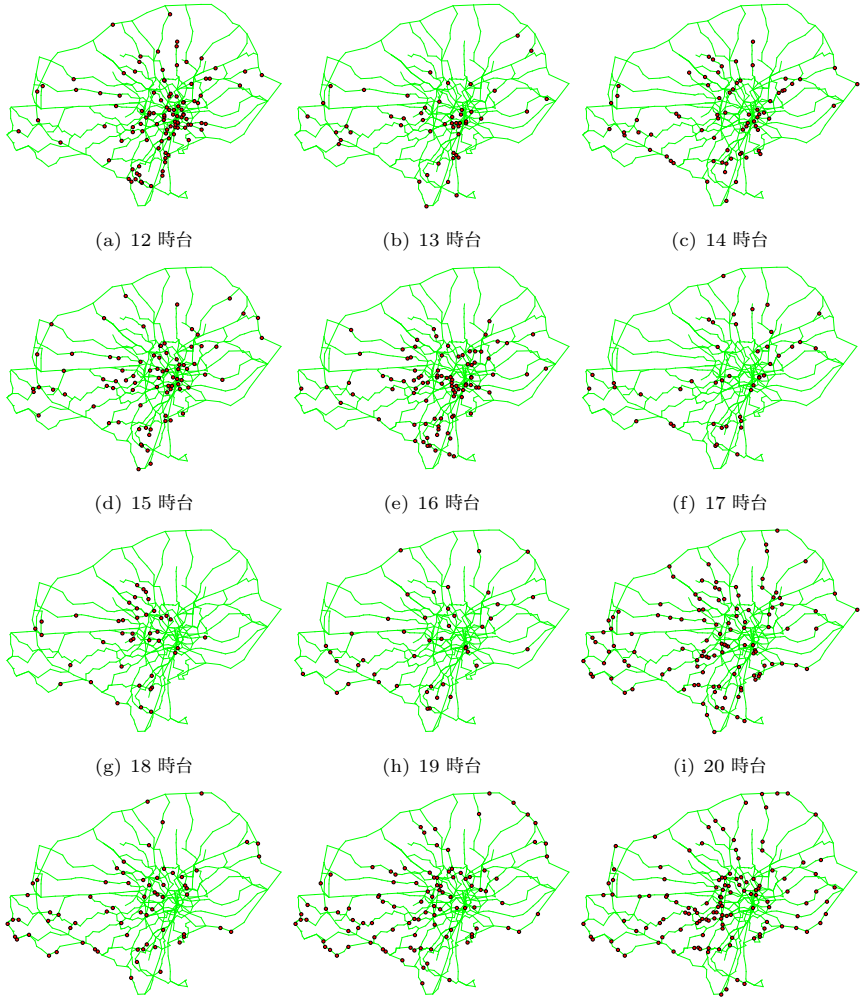
まず，地理情報検索と実世界の人間の行動の間に何らかの関連はあるが，必ずしも密接に結びついているわけではない．例えば，この関連性には，地理的クエリを発行した最中または直後に行動を起こす場合，地理的クエリを発行してから実際に行動するまでのタイムラグがある場合，地理的クエリを発行しても行動を起こさない場合の 3 種類が考えられる．ラン



(a) 0 時台　(b) 1 時台　(c) 2 時台
(d) 3 時台　(e) 4 時台　(f) 5 時台
(g) 6 時台　(h) 7 時台　(i) 8 時台
(j) 9 時台　(k) 10 時台　(l) 11 時台

図 7　地理的検索対象の時間変動 (1)

Vol.2011-DBS-152 No.9
Vol.2011-IFAT-103 No.9
2011/8/3



(a) 12 時台　(b) 13 時台　(c) 14 時台

(d) 15 時台　(e) 16 時台　(f) 17 時台

(g) 18 時台　(h) 19 時台　(i) 20 時台

(j) 21 時台　(k) 22 時台　(l) 23 時台

図 8　地理的検索対象の時間変動 (2)

チ，居酒屋などの共起語の場合は昼休みや退社時間に集中して用いられることから，リアルタイム性がかなり高いと考えられる．しかし，それ以外のリアルタイム性を持たない地理的クエリも相当存在すると推測できる．今回は地理的検索対象が存在する地域の移り変わりという観点から分析した．しかし，さらに実際の人間の移動を推定しようとした場合には，リアルタイム性が高い地域クエリを抽出するなどの処理をおこなう必要がある．

さらに，地名には曖昧性が存在する．例えば，品川で「ブログ」が，大塚で「愛」という検索語が高頻度で使われているように，地名ではなく人名として使われる場合があり，ある単語が地名として使われているのか，それ以外の目的に使われているかの判別が必要である．また,「東京」という地名を東京駅周辺と東京都全域の両方に使うように，示す範囲が必ずしも明確に決定できない．さらに，同一の地名が複数の地域を指すために使われることがある．これらの問題を解決して精度を向上する必要がある．

7. おわりに

本稿では，サーチエンジンのクエリログ中の地名を用いた地理的クエリに着目することでユーザの動向の時空間分析をおこなった．その結果，地理的クエリは一般的なクエリとは異なる時間的変動を示し人間の行動と関係があること，地名はそれぞれ固有の時間的変動を持ち地理的クエリで地名と同時に指定される検索語集合は地域固有になることを示した．さらに，地図上に特に頻繁に検索されている地名を可視化して．注目されている地名の分散や移動パターンを分析した．

今後の課題は，すでに明らかになっている課題に対処すると共に，地名，検索語，時間を多面的に分析することで，人間の流動を推測すると共に，得られた分析結果の原因を容易に探れるようにすることである．

参 考 文 献

1) 相　尚寿，岡部篤行，貞広幸雄，太田守重：時空間解析における基礎概念と解析事例の体系的整理手法，GIS — 理論と応用，Vol.16, No.2, pp.89–98 (2008).
2) NTT ドコモ：モバイル空間統計に関する情報，http://www.nttdocomo.co.jp/corporate/disclosure/mobile_spatial_statistics/ (2010).
3) モバイル社会研究所：社会・産業の発展に寄与する「モバイル空間統計」利活用のあり方に関する報告書，http://www.moba-ken.jp/pdf/research10_01.pdf (2010).
4) 杉山陽一，松原　広，明星秀一，田村一軌，尾崎尚也：改札通過データを用いた旅客流動のリアルタイム推定手法，鉄道総研報告，Vol.23, No.8, pp.11–16 (2009).