3C1-OS14-5

交通ネットワーク上の関連地名の分散パターンの分析

Distribution Pattern Analysis of Associated Geographical Names on Transportation Networks

風間 一洋
Kazuhiro KAZAMA

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

I present a new text mining approach combined with network analysis to quantify the distribution patterns of the associated geographical names. I extract geographical names, which are user's search queries, from search engine query logs and compare the similarities of any pair of geographical names using Jaccard's coefficient. I found that a set of associated geographical names for each geographical name shows a specific spatial distribution pattern on transportation network. I define a measure for quantifying such characteristics and discuss its characteristics and application for information navigation.

1. はじめに

サーチエンジンのクエリは，潜在的に何らかの地域性や局所性を持つことが多い，検索した結果が多すぎたり，複数地域の結果が混在する場合には，さらに地名を追加して検索結果を絞り込む．ただし，最初に検索対象地域が明確に決まっていない場合は，複数の地域又はある程度広い地域から，ユーザ自ら最も適切な地域を指定しなければならず，このような場合の検索支援はほとんどおこなわれていない．

本稿では，地域性を持つ情報探索における関連地域の推薦・提示のために，サーチエンジンのクエリログから地域性をもつクエリを抽出し，地名に関連する地名群が地図上に分散するパターンの違いを分析し，その定量化指標を定義する．さらに，定義した定量化指標の特徴を分析すると共に，どのような場合に適用できるかを議論する．

2. 地理的検索と地名の関連性

2.1 地理的検索と交通ネットワーク

一般に地域情報検索の UI としてディレクトリ型やマップ型が使われている．ディレクトリ型は地域を階層的なカテゴリーに分割し，階層を降りることで情報を絞り込み，別のカテゴリに移動することで別の地域の情報を閲覧する．マップ型は地図に検索結果を表示し，地図の拡大により情報を絞り込み，地図上を移動して別の地域の情報を閲覧する．これらの UI は，情報の地域的関連性は距離に反比例し，それが連続的に変化すると仮定して設計している．

しかし，現実の地域情報探索では，必ずしもこの仮定は成り立たない．まず，都市部では二つの地点の間を直線状に移動することは難しい上に，日常生活では車，バス，電車などを用いて移動することが多い．例えば，日本の東京，大阪などの都市部では鉄道網が発達しており，これらの地域の住民は電車や地下鉄を使って移動することが多く，この場合二つの地点間の移動は鉄道網のトポロジーの制約を強く受ける．地図上で等距離の場所でも実際に移動する場合の距離や移動時間は異なるので，目的地や経路の選択の際にはネットワークトポロジーを考慮する必要がある．

連絡先: 風間 一洋，NTT 未来ねっと研究所，〒 180–8585
東京都武蔵野市緑町 3–9–11，kazama@ingrid.org

さらに，各地域は，例えば，工業地域・商業地域・住宅地域などの用途の違い，その地域に住む・訪れる年齢層の違い，山の手・下町・寺町・武家町などの社会的特性の違い，空港や市場などの施設の有無など，それぞれ異なる特性を持つ．隣接する地域であっても同じ特性を持つとは限らないので，探索する情報によっては，近くの地域よりも，類似した特性の少し離れた地域が適切であることも多い．

つまり，地域情報探索においては，二次元平面上の近接性だけではなく，交通ネットワークの制約や地域特性も考慮した上で探索地域を選ぶことが重要となる．

2.2 地名の関連性と地域選択可能性

実際のサーチエンジンにおける各ユーザの地名の使用状況は，クエリログに記録される．ユーザは自分が探索したい範囲を地名で指定するので，クエリログから得られるユーザが情報検索に使用した地名の集合は，居住地や勤務地などのユーザの行動範囲や，仕事や遊びなどのユーザの行動目的を反映している．これから，多くのユーザが使用した地名の組の間には，何らかの有意な関連性があるとみなすことができる．

ただし，地域情報検索では，検索対象によって地域選択可能性が異なることに注意する必要がある．地域選択可能性は，ある検索対象が存在する地域の程度である．

例えば，コンビニエンスストアや居酒屋を探す場合には，広範囲に多数分散した店舗の中から選ぶことができるので，地域選択可能性が高い．このような場合は，通常は移動コストが小さい現在地から近い店舗を選択する．しかし，モロッコ料理店のようにかなり珍しい場合や，デパートのように多くの集客が見込める地域にしか出店しない場合には，条件に適合した店舗が存在する地域が限られるので，地域選択可能性が低い．このような場合に現在地の近くに検索対象が存在しなければ，交通機関を利用してそれが存在する地域まで移動する必要がある．

地域の観点から考えると，地域選択可能性が高い検索対象しか存在しない地域は，訪れるのが近隣の人々に限られるが，逆に地域選択可能性が低い検索対象が多く存在する地域は，広範囲から多くの人達が訪れるはずである．つまり，地域間の関連性としては，前者は近隣の地域とが強く，後者は遠い地域の関連性が強くなると推測される．

サーチエンジンのクエリログから得られる地名の関連性を分析する際には，このような検索対象の選択可能性が高い場合と低い場合が混在し，地名の関連性に影響を与えていることを

考慮する必要がある．そこで，本稿では，まず実際にクエリログを分析して得られた地名の関連性を，交通ネットワーク上の近接性と比較することで，関連性の特性を明らかにする．

3. 関連研究

篠田は，Web 上の経路探索サービスから得られる経路情報を用いて，首都圏の鉄道ネットワークの基本的な特性量の分析をおこなった [篠田 07]．橋本は，福島県郡山市のバスと鉄道の路線を基準メッシュにより分割してネットワークを作成し，連結性行列や有値運行行列を用いたグラフ理論法的手法により，結節地域の構造の分析をおこなった [橋本 89]．Limtanakool らは，西ヨーロッパの都市部における 39 都市の人間の移動と都市の社会人口・経済・交通的近接性・観光の 4 つの属性を比較することで，都市の順位付けをおこなった [Limtanakool 07]．本稿では，首都圏の鉄道ネットワーク上のクエリログで使われた地名と関連する地名群を求め，その分布パターンの特性を分析している．

4. 地名関連度

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，南武線と私鉄で囲まれた地域内の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．さらに，後述するように鉄道ネットワーク上の経路や距離を考慮するために，対象とする駅のペアの間の経路をなるべく寸断しないように，分析対象を環状になっている路線群とその内部に限定した．

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，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を，正規化して一つの駅と見なすことができる．

路線定義ファイルから得られる正規前の全地名集合を V_0，正規化後の全地名集合を V，二つの地名間接続の集合を E，交通ネットワークを $G = (V, E)$ とする．正規化後の地名数は 616 である．

4.2 地名の利用状況の分析

次に，地域情報検索において，対象地域の地名がどのように利用されているかを，クエリログを用いて分析する．

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から，各ユーザごとに V_0 に含まれる地名を抽出し，乗り換え可能駅と表記の揺れを抽出後に正規化した．抽出対象となる地名は，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表記の揺れであり，さらにこれらの地名の末尾に「駅」を付加した形式（例，「吉祥寺駅」）も抽出する．

この結果として得られる地名集合を $V = \{v_0, \ldots, v_{m-1}\}$，クエリにこれらの地名を用いたユーザ集合を $U = \{u_0, \ldots, u_{n-1}\}$ とすると，あるユーザ u_i がクエリに使用した地名集合 V_{u_i} $(0 \leq i \leq n-1)$ と，ある地名 v_i を使用したユーザ集合 U_{v_i} $(0 \leq i \leq m-1)$ が得られる．

4.3 地名関連度の計算

地域情報検索で各ユーザがクエリに使用する地名は，勤務地，遊び場所などのユーザの行動範囲や， 仕事，趣味などのユーザの嗜好を反映する．つまり，同じ地名を使用するユーザは，行動範囲や嗜好に関して何らかの類似性があると考えられる．逆に，地名の組に対して，それらを使用したユーザの集合がどの程度類似しているかを分析すれば，地名の間にどの程度の関連があるかを求めることができる．

そこで，各ユーザが検索に使用した地名集合 V の中の任意の二つの地名 v_i，v_j（ただし，$i \neq j$）の関連度 $R(v_i, v_j)$ として，地名 v_i をクエリに使用したユーザ集合 U_{v_i} と地名 v_j をクエリに使用したユーザ集合 U_{v_j} の類似度を地名関連度とする．類似度としてはいくつかの指標が考えられるが，本稿では Jaccard 係数を用いた．地名 v_i と地名 v_j の Jaccard 係数 $J(U_{v_i}, U_{v_j})$ は，次のように計算できる．

$$J(U_{v_i}, U_{v_j}) = \frac{|U_{v_i} \cap U_{v_j}|}{|U_{v_i} \cup U_{v_j}|} \tag{1}$$

なお，クエリログ中の地名の出現頻度の代わりに，地名を使用したユーザ数に着目した理由は，少数のユーザだけが特定の地名を頻繁に使うことで生じるバイアスを避けるためである．

すべての地域間関連度を求めるために，地名集合 V のすべての二つの地名の組の Jaccard 係数を計算し，地名関連度行列を作成した．この地名関連度行列を用いれば，ある地名と関連する地名を推薦することができる．

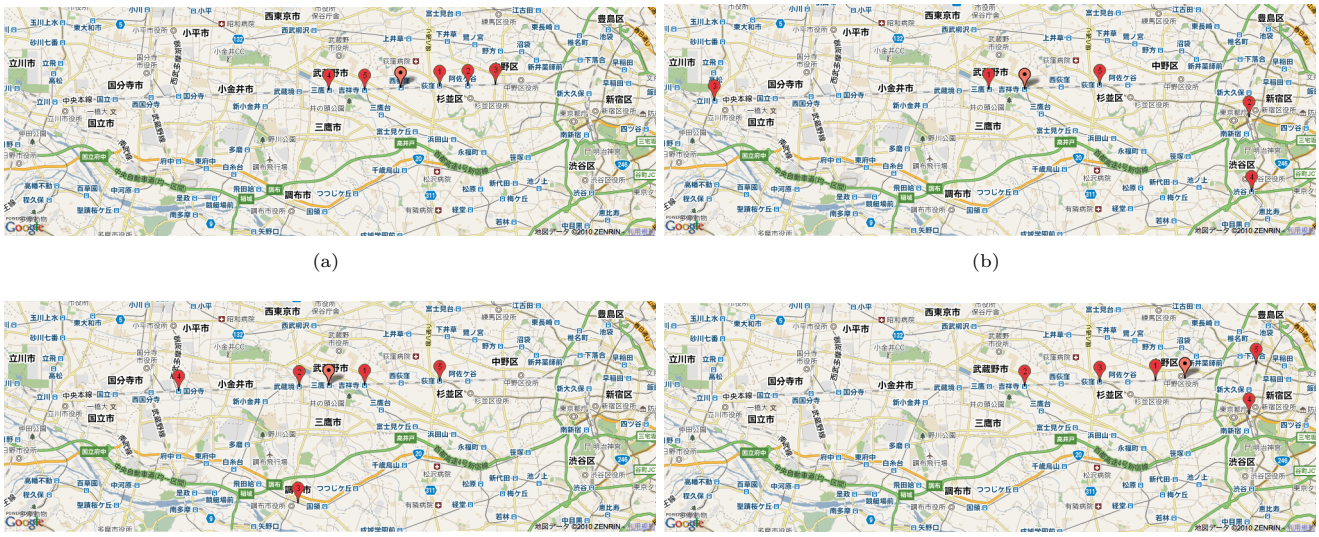
5. 地理的分散パターンと地名関連特性

5.1 関連地名の地理的分散パターン

ある地名との Jaccard 係数が上位の地名群を地図上にプロットした場合の関連地名の地理的分散パターンは，地名によって異なる特徴を示す．例えば，「西荻窪」，「吉祥寺」，「三鷹」，「中野」という地名に対して，地名関連度が大きい順から 5 件の地名を Google Maps API を用いて地図上にプロットした例を，図 1(a)，図 1(b)，図 1(c)，図 1(d) に示す．地図上のマーカのうち，中心部が黒点のマーカーは現在注目している場所を示し，中心部が数字の複数のマーカがそれと関連する地名と順位を表す．

まず，西荻窪の場合には，図 1(a) に示すように交通ネットワーク上で近接する地名との関係が強い．しかし，西荻窪に隣接する吉祥寺の場合には，図 1(b) に示すように離れていても発展している地名との関係が強い．この違いは，西荻窪は近くの狭い範囲から訪れるのに対して，吉祥寺の場合には遠くの広い範囲から訪れるように，その地域が人々を惹き付ける力に違いがあることを示す．また，吉祥寺には地域選択可能性が低い検索対象が多く存在し，それらが関連する地名の検索対象と類似していているからであり，吉祥寺に関しては検索対象の近接性ではなく類似性に基づいた地域情報検索が多くおこなわれていることを示す．

三鷹の場合には，図 1(c) に示すように基本的に西荻窪の場合と同様に近接する地名との関係が強いが，交通ネットワーク上のグラフ距離が遠い「調布」との関連も強い特徴がある．これは，三鷹又は調布の近隣にいるユーザには，少し遠くても他方の地域にも訪問する傾向があるからだと考えられる．もちろ



(a) (b)

(c) (d)

図 1: 関連地名の地理的分散パターンの例

ん，地名間の路線バスによる移動を考慮していないからとも考えられるが，武蔵関の場合にも同様に別の路線の保谷と関係があってもバス路線は存在しない反例がある上に，通常に首都圏の路線バスは鉄道網で分散された地域内の移動が主であり，路線間の乗り換えに使われることはあまりない．

中野の場合には，図 1(d) に示すように，西荻窪のように近接した地名だけと関係しているわけではないが，吉祥寺ほど広範囲ではないという，中間的な地理的分散パターンを示す．

以上のように，地名の地理的分散パターンは，その地名の特性を表していると考えられる．

5.2 グラフ距離

近接性には，地域間のユークリッド距離，平均移動時間，さまざまな定義が考えられる．

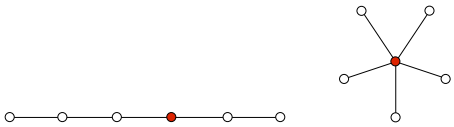
本稿では交通ネットワークの構造の制約を受けるという前提に立ち，地名集合 V の中の任意の二つの地名 v_i，v_j（ただし，$i \neq j$）の近接性を判定するために，交通ネットワーク上のグラフ距離，つまり地名 v_i と地名 v_j を結ぶ最短経路上のエッジ数を用いる．

地名集合 V のすべての二つの地名の組のグラフ距離を計算し，地名距離行列を作成しておく．この地名距離行列を用いれば，到達可能な任意の二つの地名のネットワーク上のグラフ距離を高速に取得できる．

5.3 地名関連特性

関連地名の地理的分散パターンの特徴を定量化する場合に，ある地名と関連する地名の交通ネットワーク上の平均グラフ距離を分析するのは適切ではない．例えば，5 件の地名が一番密に配置された場合のグラフ距離は，図 2(a) に示すように関連地名が直線上に一番密に配置された場合には 1，1，2，2，3 となるが，図 2(b) に示すようにその地名の次数が 5 の場合にはすべて 1 となる．このように平均グラフ距離はネットワークトポロジーに大きく影響されてしまう．

そこでネットワークトポロジーの影響を除外するために，ある地名 v_i における細密配置パターンに着目する．細密配置パターンは，交通ネットワーク上のある地名 v_i に近い順から指定された個数の地名をネットワーク上に配置したものである．ただし，同じ距離の地名が複数存在することもあるので，細密



(a) (b)

図 2: 関連地名の細密配置パターンの例

配置パターンは必ずしも一意ではないが，どの場合であっても地名 v_i からの距離集合は常に等しい．

本稿では，細密配置パターンを基準として，実際の地理的分散パターンとの距離差の加重平均を計算し，これを地名関連特性値とする．すなわち，地名 v_i の Jaccard 係数の大きい順から k 件の関連地名から求められる地名関連特性値 $P(v_i, k)$ は，次のように計算する．

$$P(v_i, k) = \frac{\sum_{j=0}^{k-1} w_j \times (d_{ij} - d'_{ij}))}{\sum_{j=0}^{k-1} w_j} \quad (2)$$

$$where\ w_j = \frac{k-j}{k} \quad (3)$$

d_{ij} は地名 v_i と Jaccard 係数の大きな順から $j+1$ 番目の地名との間のグラフ距離であり，地名関連度行列の列を降順にソートした時の $j+1$ 番目のグラフ距離である．d'_{ij} は地名 v_i に近い順から $j+1$ 番目の地名のグラフ距離であり，地名距離行列の地名 v_i の列を昇順にソートした時の $j+1$ 番目のグラフ距離である．

ここでは特に関連性が強い関連地名だけを対象とするために，上位 k 件に制限している．また，w_j は Jaccard 係数が大きい地名の特性をより考慮するための重みである．一般に順位が低くなるにつれてグラフ距離は大きくなるが，低順位で一気に距離差が増大する場合があるので，高順位から距離差が大きい場合を特に高く評価するために重みを用いている．

表 1: 地名使用ユーザ数の上位 10 件の地名関連特性値

順位	地名	ユーザ数	$P(v_i, k)$
1	東京	612,571	3.87
2	新宿	258,023	4.2
3	渋谷	179,202	3.67
4	池袋	170,725	4.2
5	銀座	161,230	3
6	羽田空港国内線ターミナル	112,907	10.27
7	上野	111,797	4.47
8	秋葉原	109,459	3
9	川崎	105,696	4.33
10	品川	93,307	3.13

6. 評価

$k = 5$ として，すべての地名 v_i に対して地名関連度 $P(v_i, k)$ を求めた．ただし，地名関連度値には，地名を使用するユーザ数が極端に少ない場合には遠方の使用頻度が高い地名との関連性が，またその単語が地名だけでなく人名のような他の目的にも用いられる場合には本来は使用頻度が低いと思われる地名であっても使用頻度が高くなるために，極端に高い値を示す問題があるので，前者はその使用したユーザ数が 500 人以上に限定し，後者はブラックリストを用いて，除去した．この結果，対象とする地名数は 516，地名関連特性値の平均は 1.27 となった．

表 1 に，地名使用ユーザ数の上位 10 件の地名関連特性値を示す．地名使用頻度が高い地名は都心にある山手線周辺部に集中しており，それらの地名関連特性値は高いことがわかる．

さらに，今回対象とした地域内の地名関連特性値を可視化して，図 3 に示す．各地名の座標は緯度・経度から求めたものであるが，簡略化のために経路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場合は，同一の地名として扱っている．

大きな緑のノードは，地名関連特性値が 2 以上の地名である．例えば，山手線周辺部にかなり密に分布し，郊外部はそれよりかなり密度は低くなるものの，葛西臨海公園，多摩センター，調布，府中，吉祥寺，浅草，板橋，自由が丘，舞浜，新百合丘などの主要駅に，広く分布していることがわかる．

7. 考察

地名関連特性値が高い地名は，以下のような地域であると推測できる．まず，駅周辺に飲食街，デパート，役所などが多く発達している地域である．次に，複数の路線の乗り換えが可能だったり，急行が止まる駅の地域である．最後に，羽田空港，浅草のように，飛行場や浅草寺など，他には存在しない重要な施設が存在する地域である．逆に，地名関連特性値が低い地名は，上記の特徴を持たないオフィス街や住宅地であることが多い．

以上の結果から，地名関連特性は，以下のような目的に適用できると思われる．まず，地名関連特性値が高い地名は一般に集客性が高い地域であるので，クエリログを分析してそのような地域を自動発見することができる．次に，グラフ間距離が近い地名だけでなく，グラフ間距離は遠くても地名関連特性値が高い地名を同時に推薦することで，情報探索における有効なショートカットを提供できる．さらに，地名関連特性値が高い地名を中心にして階層型分類をおこなうと同時にショートカットを提供することで，ディレクトリ型の地域情報探索の探索効率を向上できると考えられる．

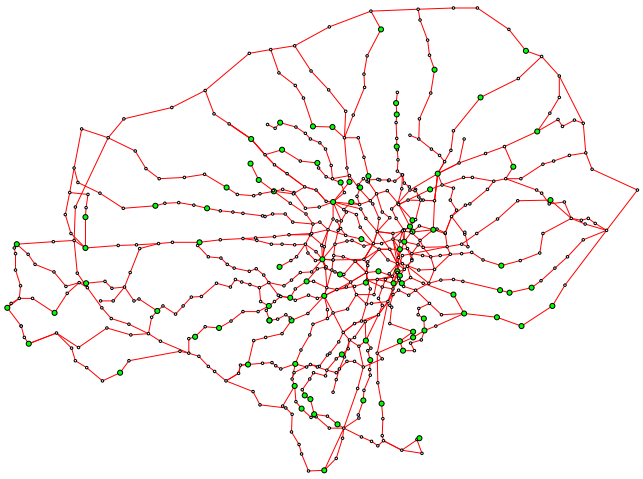


図 3: 地名関連特性値の可視化

8. おわりに

本稿では，クエリログから地名の関連性を抽出し，関連する地名群を地図上にプロットした場合の分散パターンが地名により異なることを示し，その特性を定量化する指標として地名関連特性値を提案した．

地域情報探索では単に検索対象を近接性だけで扱うことが多かったが，特に地域選択性が低い重要または珍しい検索対象の存在と，そのような検索対象への移動に用いる交通ネットワークの制約は，近接性を超えた地域間の関連性を考慮した情報推薦の必要性を示している．

今後の課題は，以下の二つである．現在は地名の判定は登録された地名辞書との一致だけで判断しているためにノイズが発生しており，ユーザがクエリに用いた地名集合との地理的分散や，同時に用いた検索語などの情報を用いて，ある単語が地名として使用されたのかどうかを判定する必要がある．また，現在は単に同一ユーザが使用したかどうかで地名の関連性を判定しているだけなので，さらに得られた関係性がどのような意味を持っているかを調べる必要がある．

参考文献

[橋本 89] 橋本 雄一：公共交通ネットワークからみた郡山市の結節構造, 地域調査報告, Vol. 11, pp. 27–39 (1989)

[Limtanakool 07] Limtanakool, N., Schwanen, T., and Dijst, M.: Ranking Functional Urban Regions : a Comparison of Interaction and Node Attribute Data, in *Cities*, Vol. 24, pp. 26–42, Elsevier (2007)

[篠田 07] 篠田 孝祐：Web を利用した空間ネットワークの分析, 第 21 回人工知能学会全国大会 3G7-1 (2007)