アナロジーに基づく地理情報検索

加藤誠† 大島裕明† 小山聡‡ 田中克己†

†京都大学大学院情報学研究科
‡北海道大学大学院情報科学研究科

1 はじめに

ユーザ自身があまり詳しくない分野の情報を，検索によって取得することは困難である．これは，情報を取得するためのクエリを作ることが難しく，また，検索結果を閲覧してもその結果がどのようなものであるか判断しがたいためである．例えば，外出先において飲食店情報を検索する際，その地域での相場や特産物などを知らなければ，具体的な問い合わせを行うことは難しい．また，結果を閲覧するときも同様で，検索で得られた飲食店がその地域においてどのような位置づけの店なのか判断できない．そこで，本論文ではアナロジーに基づく地理情報検索を提案し，特に飲食店やホテルなどを対象にした地理エンティティ検索について述べる．

2 検索モデル

アナロジーに基づく地理情報検索では，ユーザが良く知っている分野をソースドメイン $\boldsymbol{E}_s$，あまり知らない分野をターゲットドメイン $\boldsymbol{E}_t$ と呼ぶ．これらのドメイン $\boldsymbol{E}_s$，$\boldsymbol{E}_t$ は全データ集合 $\boldsymbol{E}$ の部分集合である．ユーザはソースドメインの部分集合をクエリとして選択し，ターゲットドメインの情報を検索することができる．

ソースドメインとターゲットドメインを用いて，全クエリ集合 $\boldsymbol{Q}$，及び，検索対象データ $\boldsymbol{D}$ は以下のように定義される：$\boldsymbol{Q} = \mathfrak{P}(\boldsymbol{E}_s)$, $\boldsymbol{D} = \boldsymbol{E}_t$．ここで，$\mathfrak{P}(x)$ は x の冪集合である．図1に，我々が作成した地理エンティティ検索のインタフェースを示す．

本論文では，直接的なエンティティ間の類似度に基づく検索方法を示した後，アナロジーに基づく対応付けにより検索を行う方法を提案する．

3 直接的な類似度に基づくランキングと動的な距離尺度推定手法

与えられたクエリ $Q \in \boldsymbol{Q}$ とあるデータ $d_j \in \boldsymbol{D}$ のランク関数を負の距離，すなわち，類似度として定義

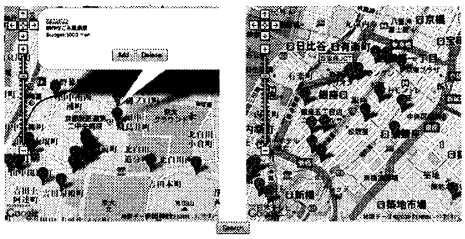

図1: 地理情報例示検索のインタフェース．ユーザはソースドメイン(左)上で検索意図と適合する地理エンティティを選択することができる．検索ボタンがクリックされると，ソースドメインで選ばれた例をクエリとして，ターゲットドメイン(右)中の地理エンティティが検索される．

する：

$$\mathrm{Rank}(Q, d_j) = -\mathrm{dist}(Q, d_j). \tag{1}$$

検索対象データ $\boldsymbol{D}$ はこの値によってランキングされ，ユーザに提示される．しかし，距離関数 dist をどのように決定するかという問題がある．距離関数はユーザによっても，またコンテキストによっても異なる．例えば，2つの飲食店，予算4,000円のフランス料理店と，予算4,000円の日本料理店は似ているとも，似ていないとも判断される場合がある．値段のみに着目すれば，両飲食店は同じと考えられるし，ジャンルに着目すれば全く違うとも考えられる．

我々は，この動的な距離尺度を非明示的な負例によって推定する方法を提案している [1]．動的な距離尺度推定とは，n 次元空間中の2点に対する距離関数を決定する行列 $\mathbf{W}$ を推定することである：$\mathrm{dist}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{W} (\mathbf{x} - \mathbf{y})$.

Ishikawa ら [2] はこの問題に対して，選択された例同士の距離が最小となるような $\mathbf{W}$ を用いている．しかし，与えられた例が少ない場合，$\mathbf{W}$ が求められない場合や極端な値を持つことがある．そこで我々は，非明示的な負例をこの距離尺度推定に用いることで，頑健な距離尺度推定を可能にした．

非明示的な負例とは，選択された例と地理的に近いが，ユーザが選択しなかった例のことである．図2に非明示的な負例の例を示す．我々は2次元混合ガウス

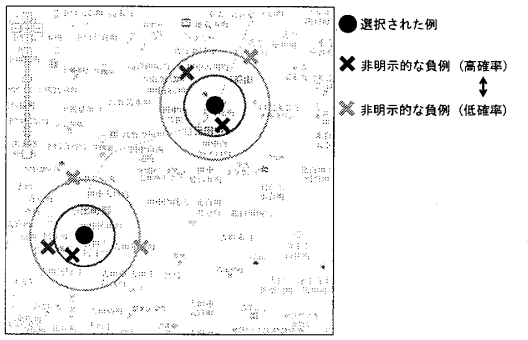

図2: 非明示的な負例の推定．濃い色は高い確率を，薄い色は低い確率を表している．

分布を用いることで，選択された例に近いエンティティを高い確率，遠い例を低い確率で非明示的な負例であると推定している．選択された例同士の距離が近くなるように，また，選択された例とその確率で重み付けされた非明示的な負例が遠くなるように，最適な行列 $\mathbf{W}$ が求められる．

我々は評価のために，20種類のクエリと2種類の検索対象データからなるテストセットを作成した．データには飲食店サイト「ぐるなび[1]」のデータを用い，ソースドメインとターゲットドメインには，東京，大阪，名古屋，京都，札幌，神戸の6都市を選択した．検索対象データは5段階で評価され，評価指標には normalized discounted cumulative gain (nDCG) を用いた．固定した距離尺度と MindReader との比較の結果，nDCG は固定距離尺度が 0.718，MindReader が 0.688，そして，提案手法が 0.735 となった．クエリとして選ばれた例の平均数が 3.4 と少なかったにも関わらず，MindReader よりも提案手法が高い値を得ていることから，非明示的な負例によってより頑健な距離尺度推定が可能になったと考えられる．

4 アナロジーに基づくランキング

前節では，直接的な類似度に基づくランキングについて述べたが，全く異なる集合に属する2つのデータの類似度を直接的に計算することは不適切である．例えば，日本と中国の国内にある飲食店の類似度を計算することは難しい問題である．2つのドメインの間には，相場や一般的な食べ物，ジャンル，距離感覚の違いなど大きな違いがあるため，両ドメインの特徴空間の各次元が同じ意味を持っているとは限らない．

そこで，我々はアナロジーの方法論に基づたランキング手法を提案する．アナロジーでは，関係の類似性が直接的な類似性よりも重視される [3]．そのため，アナロジーに基づくランキングでは，全く異なる集合に属する2つのデータの直接的な類似度を計算するのでなく，その関係の類似性によってランキングを行う．関係の類似性とは，要素間の関係が類似していることを指す．例えば，ある集合内で突出して相場が高い店は，他店よりも“相場が高い”という関係にある．この店と，他の集合において，他店と“相場が高い”という関係にある店は関係の類似性が高い．関係の類似性を用いることで，全く異なる2つのドメインの間を横断した例示検索が可能であると考えている．しかし，2語間の関係の類似度を求める方法は提案されているが，属性とその属性値で表現されるようなエンティティ間の関係を発見する方法，また，関係間の類似度計算手法は提案されていない [4]．そのため，アナロジーに基づくランキングにはこれらの技術を確立する必要があると考えている．

[1] http://www.gnavi.co.jp/

5 まとめと今後の課題

本論文では，アナロジーに基づく地理情報検索手法を提案した．今後は地理エンティティなどのデータに対して，関係を発見する方法，また，関係間の類似度を計算する方法について検討したい．

謝辞

本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，科学研究費補助金（課題番号 18049041，21700105），および，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」（研究代表者：田中克己）によるものです．ここに記して謝意を表します．

参考文献

[1] M. P. Kato, H. Ohshima, S. Oyama and K. Tanaka: “Query by example for geographic entity search with implicit negative feedback”, In Proceedings of ICUIMC 2010, to appear (2010).

[2] Y. Ishikawa, R. Subramanya and C. Faloutsos: “Mindreader: Querying databases through multiple examples”, In Proceedings of VLDB 1998, pp. 218–227 (1998).

[3] K. Holyoak and P. Thagard: “Mental leaps: Analogy in creative thought”, The MIT Press (1996).

[4] P. D. Turney: “Similarity of Semantic Relations”, Computational Linguistics, **32**, 3, pp. 379–416 (2006).