

## 地域コンテキストを考慮した 動的な特徴空間に基づく地理情報例示検索

加藤 誠<sup>†1</sup> 大島 裕明<sup>†1</sup>  
小山 聡<sup>†2</sup> 田中 克己<sup>†1</sup>

本論文では、ユーザがあまり知らない地域での例示による地理情報検索を提案する。飲食店などの地理オブジェクトはキーワードや属性を指定することで検索されることが多い。しかしながら、これらの手法は検索対象地域に対してあまり知識を持っていないユーザにとって利用することが困難な場合がある。地理情報例示検索では、よく知っている地域の例を与えることであまり知らない地域の地理オブジェクトを検索することが可能であり、検索対象地域に関して事前知識を必要としないことが大きな利点となる。選択した例によって動的に決定される特徴空間において、与えられた例に最も近い地理オブジェクトが検索システムの結果として出力される。我々は選択されたオブジェクトと選択されなかったオブジェクトの差異を増幅することによって、頑健に動的な特徴空間を求める方法を提案した。提案手法の有効性を示すために、我々は日本の Web 飲食店紹介サービスのデータを用いて飲食店検索システムを実装し、複数の地域間で実験を行った。既存の特徴空間決定手法と比較することにより、提案手法である差異増幅手法の有効性を確かめた。

### Content-based Geographic Information Retrieval Based on Dynamic Feature Space Using Regional Context

MAKOTO P. KATO,<sup>†1</sup> HIROAKI OHSHIMA,<sup>†1</sup>  
SATOSHI OYAMA<sup>†2</sup> and KATSUMI TANAKA<sup>†1</sup>

We propose a query-by-example geographic object search method for users that do not know well about the place they are in. Geographic objects, such as restaurants, are often retrieved using an attribute-based or keyword query. These methods, however, are difficult to use for users that have little knowledge on the place where they want to search. The proposed query-by-example method allows users to query by selecting examples in a familiar place for retrieving objects in an unfamiliar place. The closest objects to an input are returned based on a distance metric, which is dynamically determined by the

selected examples. One of the challenges is to predict an effective distance metric, which varies for individuals. Our proposed method is used to robustly estimate the distance metric by amplifying the difference between selected and non-selected examples. We developed a restaurant search using data obtained from a Japanese restaurant Web guide to evaluate our method. Experimental results showed the effectiveness of our method, and that the performance exceeded a previously proposed method.

#### 1. はじめに

##### 1.1 背景

飲食店やホテル、ランドマークなど、大量の地理情報が Web 上で利用可能になってきている。これらの情報は、周辺に住む人たちにとってはもちろん、旅行先や転居先などで周辺情報を得るのに役立つ。近年では、地理情報を扱った Web サービスが多く登場しており、飲食店情報サイトであるぐるなび<sup>\*1</sup>や食べログ<sup>\*2</sup>、ホテル情報サイトの Booking.com<sup>\*3</sup>などが例としてあげられる。また、いくつかの大手検索サイトも、これら商業的 Web サイト内の情報や、Web ページに記載のある情報などを集約し、地図インタフェースを介して地理情報を提供している。現実世界の地理情報が Web 上にアップロードされるほど、検索対象が狭い地域内であってもユーザが目的の情報を得ることは難しくなる。そのため、情報が爆発的に増大する昨今においては、地理情報に対する情報検索技術はその重要度を増していると考えられる。

一般的に、ホテルや飲食店などの地理オブジェクトはキーワードクエリや属性指定を行うことで検索される。しかし、キーワードクエリはユーザの複雑な要望を表現するには向かず、属性指定をするような検索では入力に時間がかかる。検索意図を単純なキーワードで表現できる場合であったり、十分に時間があつたりする場合には問題ないが、外出先などで飲食店を探すようなタスクの場合にはこれらの検索方法は適切ではない。また、両者ともに検

<sup>†1</sup> 京都大学大学院情報学研究所社会情報学専攻

Department of Social Informatics, Graduate School of Informatics, Kyoto University

<sup>†2</sup> 北海道大学大学院情報科学研究科複合情報学専攻

Division of Synergetic Information Science, Graduate School of Information Science and Technology, Hokkaido University

\*1 <http://www.gnavi.co.jp/>

\*2 <http://tabelog.com/>

\*3 <http://www.booking.com/>

索意図の明確化をユーザに求める。地理情報検索に限ったことではないが、条件を指定し条件に合致した検索結果を得る、演繹的な検索方法は、曖昧で言語化が難しい検索意図を持ったユーザにとってはクエリを入力しにくい。たとえば、「そこそこの値段で地域の特産物が食べられるような京都駅周辺の日本食料理店」という検索意図は、明示的に条件を指定するような検索には不向きであると考えられる。特に、その場所を初めて訪れた人間には、京都で「そこそこの値段」とはどの程度か、京都の「特産物」とは何か、といった知識を持っていないことが想定される。検索対象に対して知識を持っていない場合、演繹的な検索方法はより困難になると考えられる。

## 1.2 提案手法

我々は地理オブジェクトを検索する方法として、帰納的な検索方法、例示検索 (Query-by-example) を採用した<sup>1)</sup>。例示検索はデータベースへの問合せ方法<sup>21)</sup>、また、マルチメディア検索で用いられる検索方法 (Content-based retrieval<sup>6),18),20)</sup> の1つであり、さかんに研究が行われてきた分野である。例示による検索は、条件を指定する検索とは対称的に、いくつか適合する例をあげてそれに共通する要素に基づいて検索する、すなわち、帰納的な検索であると考えられる。帰納的な検索方法の利点として、明示的に検索条件を指定する必要がなく、自分が知っている情報の中で適合だと思われる例を選択すればよいということがあげられる。实例はユーザの言語化できない検索意図を伝え、複数の例を選択しクエリとすることによって複雑な検索意図を入力できる。

例示して検索するという考えは、自然に地理オブジェクト検索へと適用できる。図1に、我々が作成した地理情報例示検索のインタフェースを示す。このインタフェースには2つの地図が表示されており、左がユーザのよく知っている地域や地元の地図、右が地理オブジェクトを検索したい地域の地図になっている。それぞれ、ソースマップ、ターゲットマップと呼ばれる。ユーザは、どの地域が地図に表示されるか、また、ソースマップ内でどの地理オブジェクトが自分の検索意図に適合しているかを選ぶことができる。提案手法では、選ばれた例との類似性に基づいて、ターゲットマップ中の地理オブジェクトがランキングされ出力される。

提案する地理情報例示検索では、ユーザは一種のアナログを用いることができる。アナログを検索に組み込むことで、たとえユーザが検索したい対象について知識を持っていなかったとしても、もし自分がよく知っている領域内で適合した情報を選ぶのだとしたら、と想像することでクエリを生成することができる。冒頭で、よく知らない地域の情報を検索するクエリを生成することは困難である、と述べたが、自分がよく知る地域の例を選択することは



図1 地理情報例示検索のインタフェース。ユーザはソースマップ(左)上で検索意図と適合する地理オブジェクト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、ソースマップで選ばれた例をクエリとして、ターゲットマップ(右)中の地理オブジェクトが検索される。検索結果は地図の下方に表示される

Fig. 1 Query-by-example interface for geographic object search. Users can click and select objects relevant to their search intentions on source map (left). When search button is clicked, objects on target map (right) are ranked and returned based on those selected as query in source map. Search results are shown below maps.

容易であろう。また、当然のことながら、よく知らない地域内で適合した例を見つけ出しクエリにすることは、問題の設定上、現実的ではない。知らない地域の情報を、知っている地域の情報を例示して検索するという考えが、この提案手法の要である。

## 1.3 問題点

入力された例に基づいて地理オブジェクトをランキングするために、オブジェクト間の類似度を定義する必要がある。しかし、提案手法の類似度計算には2つの問題がある：

- (1) 類似性尺度はユーザ、コンテキストによって異なる。また、
- (2) 異なる分野におけるオブジェクト間の類似度計算方法、である。

第 1 に、心理学の研究でも指摘されるように、類似性尺度はユーザ、さらには、そのコンテキストによって変化する<sup>19)</sup>。別のいい方をすれば、類似性を判断する基準は人によっても、また、状況によっても変わりうるということである。たとえば、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まったく違っても考えられる。さらに、同じユーザであっても状況によってこの判断基準は異なる。寒波の中、体を温めたい場合には、辛い料理を出す店、韓国料理店と中華料理店の類似度は高くなるかもしれない。提案手法にとって、この動的な類似性尺度は考慮しなければならない問題の 1 つである。

第 2 に、異なる地域間、すなわち、知っている地域と知らない地域内のオブジェクト間の類似度計算が問題となる。まったく異なる集合に属する 2 つのオブジェクトの類似度を直接的に計算することは不適切である。たとえば、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 つの領域の間には、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、共通点を数え上げて類似度を計算するような手法では人間の感覚に合致しない。

以降、本論文では、ここであげた 2 つの問題のうち、前者にのみ着目して議論を続ける。ユーザが与えた例から類似性尺度を推定する手法の 1 つに Ishikawa らの MindReader<sup>12)</sup>がある。しかしながら、MindReader を我々の問題に適用した場合、過学習の問題から十分な推定精度が得られないことが予想される。検索のために入力されるキーワードなどのように、検索時にユーザから入力される情報は一般に多くない。特に Content-based search や例示検索の場合には入力方法が難しいことから、多くの入力をユーザから期待することはできない。ましてや、外出時などに検索を行うことが多い地理情報検索においては、十分な例は入力されないと考えられる。そのため、距離尺度推定手法は少数の入力であっても高い精度を達成しなくてはならない、すなわち、推定手法は「頑健」でなくてはならない。

我々は MindReader を拡張することによって、選択された例と選択されなかった例の情報をを用い、両者の差異を増幅させる、頑健な類似性尺度推定手法を提案する。提案手法によって推定された距離尺度がいかに適切であるかを、我々は人手で構築されたテストセットを用いた実験により評価した。少数の入力からであっても既存手法である MindReader よりも優れた精度を得られたことから、我々は提案手法の頑健性とその有効性を示した。

## 2. 関連研究

### 2.1 地理情報検索

Markowitz ら<sup>8),16)</sup> は、Web 検索の結果を入力された地域に関連あるページに限定したり、その地域に関連する順にソートしたりすることで、地理情報検索エンジンを作り上げた。Lieberman ら<sup>15)</sup> が開発した STEWARD も指定した地域について書かれている Web 文書を取得するもので、言及している地域の可視化も行っている。Hiramoto ら<sup>11)</sup> は電子地図上でのズームやパンなどの地図操作から、自動的にユーザの意図を読み取り、それに応じた Web ページを取得する方法を提案している。地理オブジェクトに対する効率的な索引付けや検索技術は情報検索のコミュニティの中で多く研究がなされてきた<sup>14)</sup>。また、情報検索における基本的な手法は地理情報検索においても適用されている。質問拡張や適合フィードバックなどは、たとえば GeoCLEF 2006<sup>4)</sup> などにおいて用いられ、特に地理情報に関連する固有名詞に着目して質問拡張や適合フィードバックを行うことで検索精度を向上させている。また、質問拡張は地理オントロジによっても行われ、クエリ中に含まれる地名をより広域に解釈したり、位置情報の表現変換（「清水寺の近く」というクエリ表現を「京都市内」と解釈する）を行ったりすることで検索精度の向上を行っている<sup>5),13)</sup>。しかしながら、大量の地理オブジェクトから目的の情報を得るためには、空間的なクエリやキーワードクエリを超えた直感的な検索方法が必要であると我々は考えている。

### 2.2 Content-based Retrieval

例示検索 (content-based retrieval) は画像や音楽、地理情報検索など様々な情報検索の分野で研究されてきた<sup>6),7),18),20)</sup>。Multimedia Analysis and Retrieval System (MARS)<sup>17)</sup> は Rui らによって開発された画像検索システムで、適合フィードバックの機構を備えている。MARS は選択された正例の各次元の分散を計算することで、ユーザごとに異なる類似度を考慮して適合フィードバックを行っている。Ishikawa らは類似度計算をするときにどの属性が重要か、どの属性間の相関が重要かを推測する理論的な手法 (MindReader<sup>12)</sup>) を提案している。Ashwin ら<sup>2)</sup> は MARS や MindReader のようにユーザの類似性尺度推定を行う際に正例だけをを用いると、不適合なデータに近いものも得られ続けてしまうという問題を指摘している。そこで彼らは適した検索領域を決定するために、ユーザから与えられた負例を用いる方法を提案している。その検索境界は属性空間を適合領域と不適合領域に分離することによって決定される。我々の手法と Ishikawa ら、Ashwin らの手法との違いは 4.2 節で詳しく議論する。

### 3. モデル

この章では地理オブジェクトの例示検索について述べる．まず，地理オブジェクトのデータ表現を定義し，次に情報検索の枠組みに沿って，提案する検索方法について説明する．

#### 3.1 データ表現

地理オブジェクトは位置情報を含め，いくつかの属性を備えている．また，それぞれの属性値は  $k$  次元空間の点として表現される．

地理オブジェクトのスキーマは関係データベースと同じように属性集合によって定義される： $R = \{a_1, a_2, \dots, a_M, \text{pos}\}$ ．ここで  $a_i$  は名前やカテゴリなどの属性である．pos は緯度と経度で表現される位置情報であり，これこそが一般的なオブジェクトと一線を画す特徴である．地理オブジェクトのスキーマ  $R$  は飲食店やホテル，ランドマークなどの各地理オブジェクトクラスに対して事前に定義される．

地理オブジェクトスキーマ  $R$  に対するオブジェクト  $o$  は  $N$  次元空間の点で表現される： $o = (o_{a_1}, o_{a_2}, \dots, o_{a_M}, o_{\text{pos}})$ ．ここで  $N = \sum_{i=1}^M n_i + n_{\text{pos}}$  であり， $n_i$  はベクトル  $o_{a_i}$  の次元数， $o_{a_i}$  は属性  $a_i$  の属性値である．ベクトル  $o_{a_i}$  は飲食店の予算などのスカラー値を 1 次元ベクトルで表したり，飲食店の紹介文などを tf-idf ベクトルなどで表現できる． $o_{\text{pos}}$  は緯度と経度で表される位置ベクトルであり，その次元数は  $n_{\text{pos}} = 2$  である．

#### 3.2 検索モデル

地理オブジェクトの例示検索ではよく知っている地域内で例を選択することで，詳しくない地域の地理オブジェクトを取得し，類似度を元にランキングする．ユーザは知っている地域，知らない地域の指定も行い，それぞれソースドメイン  $O_s$ ，ターゲットドメイン  $O_t$  と呼ばれる．

この 2 つのドメインはすべてのオブジェクト集合  $O$  の部分集合である：

$$O_s \subset O, O_t \subset O. \quad (1)$$

2 つのドメインは一般的に互いに疎である： $O_s \cap O_t = \phi$ ．これは，一般的に知っている分野と知らない分野が疎であることに由来する．また，我々はユーザが  $O_s$  内の要素すべてを知っていることを仮定できる．これは，ユーザそれぞれが  $O_s$  を設定することができるためである（必要に応じてユーザは知っている飲食店だけを選択して  $O_s$  とすることができる）．

ターゲットドメイン  $O_t$  のオブジェクトを検索するために，ユーザはソースドメイン  $O_s$  の部分集合をクエリとして選択する．したがって，地理オブジェクトの例示検索での全クエ

リ集合  $Q$ ，および，検索対象データ  $D$  は以下のように定義される．

$$Q = \mathfrak{P}(O_s), D = O_t. \quad (2)$$

ここで， $\mathfrak{P}(x)$  は  $x$  の冪集合である．

情報検索の定義に従って，全クエリ集合  $Q$  と検索対象データ  $D$ ，そして，ランキング関数 Rank が定義される必要がある．Rank はクエリ  $Q$  と検索対象データ  $D$  から実数  $\mathbb{R}$  への関数である： $\text{Rank} : Q \times D \rightarrow \mathbb{R}$ ．

この関数は与えられたクエリに対して得られる結果，そしてそのランキングを決定する．地理オブジェクトの例示検索では，ターゲットドメインのオブジェクトは与えられたクエリ  $Q_i \subset Q$  との類似度によってランキングされる．よって，ここで検索対象データ  $d_j \in D$  の  $\text{Rank}(Q_i, d_j)$  はその類似度として与えられるべきである．類似度は距離の逆であると考えられることも可能で，その方が我々の手法と MindReader を比較しやすい．したがって，ここではランキング関数を以下のように定義する：

$$\text{Rank}(Q_i, d_j) = \exp(-\text{dist}(Q_i, d_j)). \quad (3)$$

項  $\text{dist}(Q_i, d_j)$  は選択されたオブジェクト集合  $Q_i$  と検索対象データ  $d_j$  との距離である． $\exp(-x)$  は単に距離の逆をとっているだけであり，この形式は距離の逆数を  $[0, 1]$  に正規化している．

### 4. 類似性尺度の推定

前章では地理オブジェクトの例示検索をモデル化した．次にユーザによって変化する距離尺度を推定する方法について議論する（本論文では，類似性尺度と距離尺度の推定は等価である）．ここでは，MindReader における距離尺度推定手法を説明した後に，その手法の限界と我々の提案する検索に直接適用すべきでない理由を示す．そして，地理オブジェクト例示検索の特徴をふまえて，距離尺度推定手法を適用する方法を提案する．本章で我々が用いる記号を表 1 に列挙する．

#### 4.1 MindReader の距離尺度推定手法

一般的に，2 点間のユークリッド距離は以下のように定義される： $\text{dist}(x, y) = \sqrt{(x-y)^T(x-y)}$ ．またこれは， $(x-y)^T I(x-y)$  と等価である（平方根は簡単のため省略している）．ここで現れる単位行列を対角行列に置き換えると，重み付きユークリッド距離となる： $(x-y)^T D(x-y)$ ．重み付きユークリッド距離では，距離計算上，重要な次元は大きく重み付けされ，重要でない次元は小さな重みが与えられる．たとえば，あるユーザにとっては値段属性が重要で，他のユーザにとってはジャンル属性が距離計算上は重要で

表 1 本章で用いる記号  
Table 1 Symbol usage.

記号	定義
$O$	全オブジェクト集合.
$O_s$	ソースドメイン ( $O$ の部分集合).
$O_t$	ターゲットドメイン ( $O$ の部分集合).
$Q$	全クエリ集合 ( $O_s$ の全部分集合).
$Q_i$	クエリ, $Q$ の要素 (オブジェクト集合).
$D$	検索対象データ (= $O_t$ ).
$o$	オブジェクト, $O$ の要素.
$N$	オブジェクト $o$ の次元数.
$q_k$	$Q_i$ の要素 (オブジェクト).
$g_k$	$q_k$ に対するスコア (0/1, もしくは, 段階的な値).
$d_j$	$D$ の要素.
Rank()	ランキング関数 ( $Q \times D \rightarrow \mathbb{R}$ ).
$m$	理想的なクエリベクトル ( $N$ 次元ベクトル).
$W$	距離関数を決定する $N \times N$ 行列.
dist()	距離関数 ( $\mathfrak{P}(O) \times O \rightarrow \mathbb{R}$ ).

あるということが考慮できる. さらに, 対角行列の代わりに対称行列  $W$  を用いることで, この距離関数は各次元間の相関関係を考慮することができるようになる. たとえば, 値段属性とジャンル属性の両方の次元が重要であることを表現できる. この対称行列は距離関数を決定し, あるユーザの類似性尺度推定はこの行列  $W$  を推定することに帰着できる.

MindReader ではクエリ  $Q_i$  とデータ  $d_j$  との距離関数は以下のように定義される:

$$\text{dist}(Q_i, d_j) = (d_j - m)^T W (d_j - m). \tag{4}$$

ただし, ユーザがある理想的なクエリベクトル  $m$  と対称行列  $W$  に対応する距離関数を心の中に持っていることを仮定している. そして, 距離尺度の推定は与えられた例集合  $Q_i$  から  $m$  と  $W$  を推定することであるといえ換えられる. 図 2 に MindReader における距離尺度推定問題を示す. 黒い丸は与えられた, 選択された例であり, 問題は白丸で表現された適合データだけを取得するような, 理想的な境界線とその中心を発見することである.

距離尺度推定の基本的なアイデアは, 選ばれた例集合  $Q_i$  と理想的なクエリ  $m$  との距離を最小化するような距離関数を発見することである. もしユーザが理想的なクエリ, 理想的な距離尺度を想定していたならば, その距離尺度において, そのユーザが選択した例と理想的なクエリは非常に近くなるはずである. これらの仮定に従うと, 理想的なクエリ  $m$  と理想的な距離尺度を決定する行列  $W$  は以下の最小化問題を解くことで得られる:

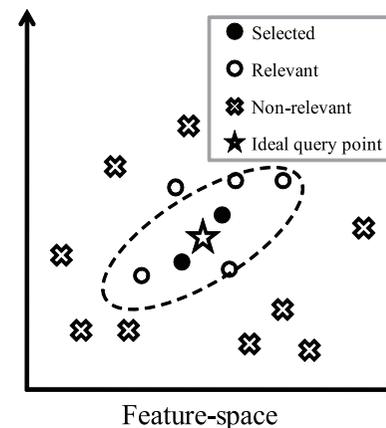


図 2 MindReader における距離尺度推定問題. 黒い丸は与えられた, 選択された例であり, 問題は白丸で表現された適合データだけを取得するような, 理想的な境界線とその中心を発見することである

Fig. 2 Prediction of distance metric from selected examples in MindReader. The problem is to find ideal bound and its center, which are illustrated as dotted line and a star.

$$\min_{m, W} \sum_{q_k \in Q_i} g_k (q_k - m)^T W (q_k - m). \tag{5}$$

ただし, 以下の制約を満たす:

$$\det(W) = 1. \tag{6}$$

ここで,  $\det(W)$  は行列  $W$  の行列式を表す (行列式が 1 であるという制約は本質的ではないが, 行列を一意に決定するために必要である). スカラー値  $g_k$  は, 選択された例がどれほど適合しているかを表すスコアで, ユーザにより入力されるものである. また, そのデフォルト値は 1 である (段階的な値,  $[1, 5]$  などでもよい).

この問題は Ishikawa らにより解析的に解かれた. 理想的なクエリ  $m$  は選択された例の ( $g_k$  によって重み付けされた) 中心と等しくなり, 理想的な距離尺度を決定する行列  $W$  は選択された例の ( $g_k$  によって重み付けされた) 共分散行列の逆行列と等しくなる.

加えて, MARS<sup>17)</sup> における次元への重み付け手法は式 (4) の特別な場合であるとも証明されている. 行列  $W$  を対称行列に限定し, 選択された例へのスコアを  $g_k = 1$  に限定すると, この対角行列の要素は各次元における分散の逆数に比例する:  $w_{ii} \propto \frac{1}{\sigma_i^2}$ . これは, MARS での提案手法と等価である.

## 4.2 提案する距離尺度推定手法

ユーザごとに変わる距離尺度は、Ishikawa らによって理論的に求められた。しかし、MindReader には本質的な問題が 2 つ存在するため、我々の提案する検索には直接適用することができない：

- (1) 少数の例しか与えられなかった場合、MindReader は適切な距離尺度を推定できない。また、
- (2) 画像特徴量などとは異なり、オブジェクトのいくつかの特徴は距離尺度において標準的な重みを持っている。

1 つ目の問題は、少数の例しか与えられなかった場合についてである。1 章で述べたように、極少数の例を入力したユーザが、どの属性を重視しているかを推定することは困難である。しかしながら、被験者によるデータセット作成 (5 章) でも見られるように、我々の検索システムでは多くの入力をユーザから期待することはできない。そのため、距離尺度推定手法は頑健でなくてはならず、1 つの例しか与えられなくても動作するべきである。

2 つ目の問題は、人物や飲食店などといったオブジェクトには常識的な、標準的な距離尺度が存在していることである。たとえば、他の属性がほとんど同じだが、平均して 1,000 円と 10,000 円の料理を出す 2 つの店は似ていないと判断されることが多いだろうが、名前だけが異なっていて他の属性がほとんど同じ店は似ていると判断されるであろう。これらの例に見られるように、我々があるオブジェクトに対して標準的な距離尺度を持っていることは容易に受け入れられるであろう。また、標準的な尺度からかけ離れた距離尺度はユーザにとって受け入れがたいものである。しかしながら、MindReader ではこの標準的な距離尺度を無視しており、特に十分な例が与えられない場合には極端な距離尺度が得られる。

### 4.2.1 差異増幅による距離尺度推定

距離尺度推定における主な問題はユーザから与えられる入力の少なさにある。一般的な例示検索とは異なり、我々の問題では、クエリとして選ばれた例  $Q_i$  のほかに、選ばれなかった例  $\bar{Q}_i = O_s - Q_i$  も存在する。選ばれた例である  $Q_i$  も選ばれなかった例  $\bar{Q}_i$  もユーザの意図を含んでいると考えられ、また、選ばれなかった例を考慮することで我々は地域コンテキストを考慮することができる。地域コンテキストとはすなわち、ソースドメイン  $O_s$  にほかならず、地域コンテキストを考慮するとは、クエリとして選ばれた例  $Q_i$  だけでなく、選ばれなかった例  $\bar{Q}_i$  も考慮することである。たとえば、高い飲食店が建ち並ぶ地域から安い店を選んだとき、安い飲食店が多い地域から同じ店を選んだときを考えると、両者の例示の意味が異なることが想像できる。つまり、前者はユーザが値段に注目していることを暗に

表している。選択された例はつねに正例と考えることができるが、一方で選択されなかった例を負例と見なすことは適当ではない。したがって、我々は Ashwin ら<sup>2)</sup> が提案している正例、負例を利用した距離尺度推定手法をそのまま適用することはできない。

そこで我々は選ばれた例  $Q_i$  と選ばれなかった例  $\bar{Q}_i = O_s - Q_i$  の差異を増幅することで、距離尺度推定を行う手法を提案する。3 章で述べたように、ユーザは  $O_s$  内すべての要素を知っていると仮定する。このとき、ユーザに選ばれた例はそのユーザにとって理想とするクエリ  $m$  に非常に近いと想像できる一方で、選ばれなかった例はそのユーザにとって理想とするクエリから遠かったために選ばれなかったと考えることができる。我々の提案手法の基本的なアイデアは、選択された例  $Q_i$  と理想的なクエリ  $m$  間の距離を最小化し、選ばれなかった例  $\bar{Q}_i$  と理想的なクエリ  $m$  間の距離を最大化することである。選ばれなかった例は理想的なクエリ  $m$  に影響を与えることはない。クエリ  $m$  を選択された例  $Q_i$  の平均に固定したまま、選ばれなかった例により、選択・非選択の例の差異を増幅させることで、理想的な距離尺度を決定する行列  $W$  を変化させる。

以上の議論を含めて、我々は式 (5) の再定義を行った：

$$\min_{\mathbf{W}} \frac{1}{|Q_i|} \sum_{\mathbf{q}_k \in Q_i} g_k (\mathbf{q}_k - \mathbf{m})^T \mathbf{W} (\mathbf{q}_k - \mathbf{m}) - \frac{\alpha}{|\bar{Q}_i|} \sum_{\mathbf{o} \in \bar{Q}_i} h(Q_i, \mathbf{o}) (\mathbf{o} - \mathbf{m})^T \mathbf{W} (\mathbf{o} - \mathbf{m}). \quad (7)$$

ただし、以下の制約を満たす：

$$\|\mathbf{W}\| = 1, w_{ij} \geq 0. \quad (8)$$

ここで、 $w_{ij}$  は行列  $W$ 、 $v_k$  の要素であり、 $m$  はクエリ  $Q_i$  の ( $g_k$  により重み付けされた) 要素の平均である。

$$\mathbf{m} = \frac{1}{N_g} \sum_{\mathbf{q}_k \in Q_i} g_k \mathbf{q}_k. \quad (9)$$

主な変更は、和の範囲、理想的なクエリベクトル、行列  $W$  の制約条件に見られる。第 1 に我々は、選ばれなかった例を含めたすべてのオブジェクトと理想的なクエリとの差異を考慮した。選ばれなかった例には関数  $-ah(Q_i, \mathbf{o})$  で重み付けされ、理想的なクエリとの距離を最大化するようにしている。関数  $h(Q_i, \mathbf{o})$  には様々な種類が考えられ、後に 2 種類の関数を示し実験で比較を行う。第 2 に、理想的なクエリは選択されたクエリの平均に固

定されている．理想的なクエリはもはや最小化問題中の変数ではなくなったが，結果としては MindReader と同様にクエリの平均となっている．最後に，行列  $\mathbf{W}$  が負の値や過度に高い値を持たないように制約条件が少し変更されている（いまだにこの制約は本質的ではない）．

次に考えるべき問題はどの選ばれなかった例を差異増幅に用いるかである．選択された例に比べて選択されなかった例は多いため，すべての例との差異をとるべきでなく，ある特定の例との差異をとるべきかもしれない．どの非選択例との差異を増幅させるかは，関数  $h(Q_i, \mathbf{o})$  で決定することができる．

#### 4.2.2 差異増幅を行う例の選択手法

##### 選択された例と地理的に近い非選択例

差異増幅を行う例を選択するための 1 つ目の基準は，選択された例との地理的な距離である．ここでの仮定は，例選択のときにユーザは地理的に近い例どうしの差異を強調しているということである．ユーザが地図インタフェースを用いて，例集合からクエリとなる例を選択するときには，選んだ例の周囲にある例にも気づいていると考えることができる．もしユーザにとって，その非選択例が選んだ例と特徴空間において近いのならば，同様にその非選択例もクエリとして選択する可能性が高い．つまり，選択された例に地理的に近いが選択されていない例は，ユーザが考える特徴空間においては理想とする例から遠いと考えることができる．そこで，我々は関数  $h(Q_i, \mathbf{o})$  を下記のように設定することによって，選択された例から地理的に近い例だけを差異増幅のために用いた．

$$h(Q_i, \mathbf{o}) = \begin{cases} 1 & \text{GeoDist}(Q_i, \mathbf{o}) < \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

関数 GeoDist はオブジェクト  $\mathbf{o}$  と選択された例  $Q_i$  の平均との地理的な距離を表している．

$$\text{GeoDist}(Q_i, \mathbf{o}) = \frac{\|\mathbf{o}_{\text{pos}} - \mathbf{m}_{\text{pos}}\|}{\text{StdGeo}(Q_i)}, \quad (11)$$

ただし，StdGeo( $Q_i$ ) の値は選択された例間の平均地理的距離であり，これは正規化のために使われている．

$$\text{StdGeo}(Q_i) = \frac{1}{|Q_i|} \sum_{\mathbf{q}_k \in Q_i} \|\mathbf{q}_{k,\text{pos}} - \mathbf{m}_{\text{pos}}\|. \quad (12)$$

##### 選択された例と特徴空間において近い非選択例

差異増幅を行う例を選択するための 2 つ目の基準は，選択された例との特徴空間における距離である．ここでの仮定は，例選択のときにユーザは特徴空間において近い例どうしの差異を強調しているということである．もしユーザがある例を選択し，選択した例に近い（よく似た）別の例を選ばなかったのならば，我々はそのには理由があるのだと想像することができる．すなわち，特にそのユーザにとってはそれらの例は異なっているように感じられたと考えることができる．そこで，関数  $h(Q_i, \mathbf{o})$  は特徴空間において近い例にだけ 1 を与えるように定義した．

$$h(Q_i, \mathbf{o}) = \begin{cases} 1 & \text{FeatureDist}(Q_i, \mathbf{o}) < \beta \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

しかしながら，この時点では，距離尺度を決める過程の途中であるため，まだ特徴空間における距離尺度は決まっていない．再帰を避けるために，ある標準的な距離尺度によって FeatureDist の値を計算する．次項で述べるように，標準的な距離尺度は事前に定義されている．特徴空間における距離である FeatureDist は標準的な距離尺度を表す行列  $\hat{\mathbf{W}}$  を用いて下記のように定義される．

$$\text{FeatureDist}(Q_i, \mathbf{o}) = \frac{(\mathbf{o} - \mathbf{m})^T \hat{\mathbf{W}} (\mathbf{o} - \mathbf{m})}{\text{StdFeature}(Q_i)}, \quad (14)$$

ただし，StdFeature( $Q_i$ ) の値は選択された例間の特徴空間における平均距離であり，これは正規化のために使われている．

$$\text{StdFeature}(Q_i) = \frac{1}{|Q_i|} \sum_{\mathbf{q}_k \in Q_i} (\mathbf{q}_k - \mathbf{m})^T \hat{\mathbf{W}} (\mathbf{q}_k - \mathbf{m}). \quad (15)$$

#### 4.2.3 距離尺度の過学習へのペナルティ

これまではオブジェクトの属性は皆同等に扱われてきた．しかし，オブジェクト間の距離を計算するときに，画像特徴量などとは異なり，いくつかの属性は重要で，それ以外はあまり重要でないことがある．そこで，標準的な距離をあらかじめ決定しておき，推定により得られた，過度な距離尺度にはペナルティを与える．この種のペナルティは機械学習において，パラメータの過学習を防ぐ目的で導入されることがある．

$\hat{\mathbf{W}}$  を標準的な距離尺度を表す行列だとすると，最終的に距離尺度推定問題は以下のように定義される：

$$\min_{\mathbf{W}} \frac{1}{|Q_i|} \sum_{\mathbf{q}_k \in Q_i} g_k(\mathbf{q}_k - \mathbf{m})^T \mathbf{W}(\mathbf{q}_k - \mathbf{m}) - \frac{\alpha}{|Q_i|} \sum_{\mathbf{o} \in Q_i} h(Q_i, \mathbf{o})(\mathbf{o} - \mathbf{m})^T \mathbf{W}(\mathbf{o} - \mathbf{m}) + \frac{\rho}{2} \|\mathbf{W} - \hat{\mathbf{W}}\|^2. \quad (16)$$

ただし、式 (8) の制約を満たす。項  $\|\mathbf{W} - \hat{\mathbf{W}}\|^2$  は、標準的な距離行列  $\hat{\mathbf{W}}$  から大きく離れた  $\mathbf{W}$  へとペナルティを与え、そのペナルティは推定された距離行列をできるだけ標準的な距離行列に近づけようとする。

## 5. 実験

この章ではまず地理情報例示検索の実装について述べ、4章で議論した距離行列を求める方法について説明する。また、実験のテストセット、評価尺度、比較手法について説明した後、実験結果についての考察を述べる。

### 5.1 実装

我々が用いた地理オブジェクトはグルメ情報検索サイト「ぐるなび」から得られたものであり、地図インタフェースは Google Maps API<sup>\*1</sup>を利用して実装した。

ぐるなび Web サービス<sup>\*2</sup>を介して得られた飲食店の数は 46,945 件である。飲食店情報はいくつかの属性を有しているが、その中でも 5 つの属性のみ (名前, カテゴリ名, カテゴリラベル, 紹介文, 予算) を特徴付けに用いた。

名前とカテゴリ名, 紹介文の 3 つの属性値はテキストで構成され、tf-idf によって重み付け、ベクトル化される。全テキスト中で  $j$  番目のテキストの、全語彙中で  $i$  番目の語の数は  $tf_{ij}$  で表され、全語彙中で  $i$  番目の語を含むテキスト数は  $df_i$  によって表す。これにより、各属性は語彙数次元の特徴ベクトルで表現され、その  $i$  番目の要素は  $tf_{ij}idf_i$  となる。ただし、 $idf_i$  は  $idf_i = \log(L/df_i)$  であり、 $L$  は全テキスト数である。最終的に各特徴ベクトルは L2 ノルムが 1 になるように正規化される (この操作により、ベクトル間の距離の最大値は 2 になる)。

しかし、このベクトルは語彙数と等しい次元を持つため、疎かつ距離行列のパラメータも非常に多くなってしまふ。そのため、*latent semantic analysis*<sup>10)</sup> によってベクト

ルの次元数 27,212 は、50 次元へと圧縮されている。また、カテゴリラベルの属性値は複数のラベル、たとえば、{和食, 海鮮, 居酒屋} で構成されている。この属性も同様に tf-idf によりベクトル化され、次元数 158 は 20 次元へと圧縮された。予算属性はその最大距離が 2 になるように正規化された 1 次元ベクトルである。したがって、飲食店情報のスキーマは  $R = \{\text{text, category\_label, budget, pos}\}$  となり、オブジェクトの次元数は  $50 + 20 + 1 + 2 = 73$  となる (ただし、位置を表す 2 次元は距離計算のときには無視される)。

### 5.2 距離行列の推定

制約条件 (式 (8)) の下での最適化問題 (式 (16)) は解析的には解けないが、半正定値計画問題として数値的に解くことは可能である<sup>3)</sup>。しかし、距離行列の次元が大きいため多大な計算時間を要し、検索に対応することができない。そこで、距離行列を対角行列に制限することで最適化問題の次元数を減らす。この制限により、式 (16) は凸計画問題となり、パラメータ数はオブジェクトの次元と等しくなる。

また、標準的な距離行列  $\hat{\mathbf{W}}$  は support vector machine (SVM) 回帰によって得られる。我々はあらかじめ 275 の飲食店ペアに対して、5 段階でそれらの距離を与え訓練データとした。ただし、制約として  $\|\hat{\mathbf{W}}\| = 1$  と  $\hat{w}_{ij} \geq 0$  ( $i, j = 1, 2, \dots, N$ ) を加えた。

### 5.3 テストセット

4 人の被験者によって生成されたテストセットを性能評価に用いた。テストセットはあらかじめ用意した検索意図と、各意図に対するクエリ、適合度付きデータで構成されている。検索意図、ソースマップ (ユーザがクエリを選択する地図)、そして、ターゲットマップ (検索対象データを含む地図) を表 2 に示す。ソースマップとターゲットマップには日本の主要な都市を選択した。また、5 つの検索意図はユーザと地域によって基準が異なりにくい、絶対的な検索意図から 2 つ (検索意図 ID : 1, 2), ユーザと地域によって基準が異なりうる、相対的な検索意図から 2 つ (検索意図 ID : 3, 4), そして、複雑な検索意図 1 つ (検索意図 ID : 5) から構成されている。ここで、複雑な検索意図とは、複数の意図の組合せから構成される意図を意味している。また、特に今回用いた複雑な意図には 2 つの相対的な意図 (値段, 土地固有のもの) を含み、その地域固有の事前知識が必要であるため、他の検索意図に比べ検索を行うのが難しい場合を想定している。

我々が用いたテストセットの作成方法を以下に示す。まず、被験者それぞれに 1 つのソースマップを提示し、5 分間で地図中に表示される飲食店を閲覧し、2 分間で各検索意図に対してクエリとなるような例を選択してもらった。我々はユーザに対してどのように例を選択すべきか明確な選択方針を与えず、ただ検索意図に適合するような例を自然に選択する

\*1 <http://www.google.com/apis/maps/>

\*2 <http://api.gnavi.co.jp/>

表 2 テストセットに含まれる検索意図とソースマップ, ターゲットマップ  
Table 2 Search intentions, source maps, and target maps.

検索意図		
	ID	内容
絶対的	1	1,000 円ぐらいの飲食店
	2	辛い料理を出す飲食店
相対的	3	魚介類を出す飲食店
	4	高級な飲食店
複雑	5	そこそこ高い値段でその土地特有のものを食べられる飲食店

ソースマップ			ターゲットマップ			
ID	地域	オブジェクト数	ID	地域	オブジェクト数	カッパ係数の平均
1	東京	55	1	京都	49	0.94
2	名古屋	55	2	神戸	50	0.86
3	大阪	59				
4	札幌	51				

よう指示した．これによって，我々はそれぞれのソースマップ，それぞれの検索意図に対してクエリ，すなわち，20 種類の入力を得た．時間制限を設けたのは，実際の用途を想定したためである．時間の制約がなければすべての正例を選択し，すべての負例を選択しなければよく，問題が単純化されてしまう．図 3 にそれぞれの検索意図に対して選択された例の統計情報を示す．平均して 3.4 の例が選択されている．また，実験では選択された例のスコアはすべて  $g_k = 1$  としている．

ユーザがすべての適合する例を選択していないことを示すため，また，各地域ごとに検索意図に適合する飲食店数が異なることを示すため，本論文の第 1 著者がすべての地域において各検索意図に対して適合する飲食店数を選択した．このとき選択された飲食店数の分布を図 4 に示す．検索意図 2, 5 に適合する飲食店は元々少なく，ユーザはほとんどの適合例をクエリとして選択している．しかし一方で，検索意図 3, 4 では適合している飲食店の数に対し，クエリとして選ばれた例は少ないことが分かる．このことから，2 分間という現実的な時間の制約を課した場合，適合しているすべての飲食店を地図中から選択しクエリとすることは困難であると考えられる．そのため，差異増幅においては慎重にならなければならない．また，適合しているにもかかわらず選ばれなかった例を考慮することが必要になる．また，検索意図に適合する飲食店数は地域ごとに大きく異なっている．たとえば，ソースマップ 3 (大阪) では検索意図 1 に適合する店が多く，検索意図 4 に適合する店が少ないため，他の地域に比べ飲食店の平均予算が低いと推測される．また，ソースマップ 4 (札幌) では検索

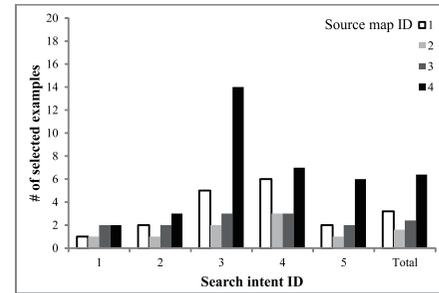


図 3 各ソースマップでクエリとして選択された例の数  
Fig. 3 Number of selected examples as query.

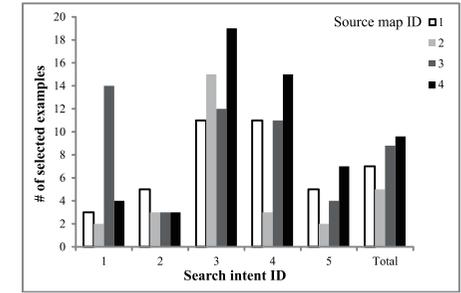


図 4 各ソースマップで各検索意図に適合する飲食店数  
Fig. 4 Number of relevant examples.

意図 3 に合致する飲食店数が多く，地域の特性が際立っている．

次に被験者を 2 人ずつに分け，それぞれのグループに対して 1 つのターゲットマップを提示した．被験者は同様に 5 分間でターゲットマップ中の飲食店を閲覧し，すべての飲食店に対して検索意図に適合する度合いを 5 段階で評価してもらった．複数評価者間で合意が得られている度合いを表す，平均重み付きカッパ係数も表 2 にあわせて記載されている．この係数が 0.80 を超えているため，ほぼ合意がなされていることが分かる．そのため，本実験の検索タスクはユーザによる違いが非常に少なく，実験には少数の被験者数で十分であると考えられる．

以上の被験者による評価によって，20 種類のクエリに対して 2 種類の検索対象データが得られた．各手法の評価は，40 種類の検索結果に対して評価値を算出し，それらの平均をとることによって行った．

#### 5.4 実験設定

我々は提案手法の性能を検証するために，既存の距離尺度推定手法である MindReader との比較を行った．ただし，MindReader の距離行列は提案手法と同様に対角行列に限定している．また，クエリとして 1 つしか例が与えられなかった場合には，単位行列を距離行列として用いている．これは 1 つの例では分散が計算できず，MindReader の手法が適用できないためである．

評価指標には discounted cumulative gain (nDCG) と mean average precision (MAP), そして，第 1 位の適合率 (P@1) を用いた．2 人の評価者によって与えられた評価値の平均を nDCG のための評価値とし，また，MAP と P@1 では 4 以上の平均値を得た場合に適合であるとして評価値を求めた．

5.5 実験結果

5.5.1 差異増幅手法の効果検証

既存手法との比較実験を行う前に、2つの差異増幅手法を比較する。それぞれの手法に対する平均 nDCG を図 5 と図 6 に示す。式 (16) 中のパラメータ  $\rho$  はここでは 1 に固定して他のパラメータを変えて実験している。両図において  $\alpha = 0$  のときが選択されなかった例を考慮しない場合、すなわち、地域コンテキストを考慮しない場合であるため、提案手法は効果的に働いていることが分かる。nDCG の値はそれぞれのパラメータ  $\epsilon$  と  $\beta$  が増加するにつれて向上し、 $\epsilon = 1.0$  と  $\beta = 3.0$  以降ではほぼ一定の値を示している。最大値は  $\alpha = 2.0$  と  $\beta = 3.0$  の特徴空間における近傍を差異増幅に用いた手法で得られてはいるが、もう一方の手法である地理的近傍を用いた手法、そして、他の最大値付近の値との明確な差は見られない。これら 2つの手法を用いた差異増幅手法の比較実験における結論は、選択されなかった例を考慮した場合、すなわち、地域コンテキストを考慮した場合、それらを考慮しないと比べて良い結果が得られるということである。次の実験では、最大値を得た、特徴空間での距離を用いた差異増幅手法を用い、パラメータを  $\alpha = 2.0, \beta = 3.0$  と設定したうえで既存手法との比較を行う。

5.5.2 既存研究との比較

MindReader と提案手法の比較実験の結果を表 3 に示す。ここでは、40 の検索結果から得

られた評価値の平均値を示している。提案手法は MindReader よりも優れた性能を示し、両側  $t$  検定において、nDCG では  $p = 0.0472 < 0.05$ 、そして MAP では  $p = 0.0118 < 0.05$  と、提案手法と MindReader の間には有意な差がある。ただし、P@1 では両側二項検定で有意差を得られなかった ( $p = 0.0758$ )。また、差異の大きさを示すために、提案手法と MindReader 間の Effect size (Cohen's  $d^{(9)}$ ) も示している。特に nDCG と MAP では Cohen's  $d$  において中程度の差異が見られる。提案手法が MindReader の性能を上回ったのは、MindReader では少ない例示によって適切な距離尺度を推定することが難しいためであると考えられる。本実験では各飲食店は約 70 次元のベクトルで表現されており、一方で入力 (例示データ) は約 4 個しか与えられていない。そのため、本実験ではつねに過学習が問題となり、頑健性の高い手法でなければ高い精度を得ることはできない。

また、本論文の第 1 著者によって時間の制約なく網羅的に選択された例、すなわち、図 4 の分布を持った入力を与えた場合に得られた結果を表 4 に示す。提案手法と MindReader の nDCG と MAP、および、MindReader の P@1 に改善が見られ、Cohen's  $d$  の値から各評価指標の差が小さくなっていることが分かる。両側  $t$  検定において、MAP では  $p = 0.0118 < 0.05$  と提案手法と MindReader の間には有意な差があるものの、nDCG では  $p = 0.0844$  と有意差がなくなり、P@1 では MindReader が提案手法を上回っている (ただし、P@1 にも有意差はない、 $p = 0.317$ )。被験者が選択した例数の平均が 3.4 であったのに対し、著者が網

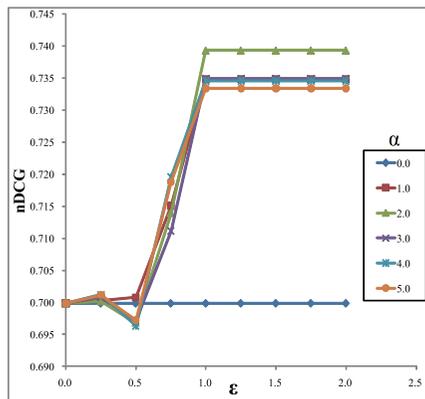


図 5 各  $\epsilon, \alpha$  に対する地理的近傍を用いた差異増幅手法による平均 nDCG

Fig. 5 Average nDCG for each  $\epsilon$  and  $\alpha$ .

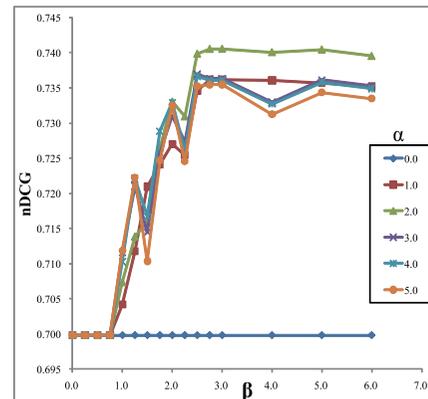


図 6 各  $\beta, \alpha$  に対する特徴空間での距離を用いた差異増幅手法による平均 nDCG

Fig. 6 Average nDCG for each  $\beta$  and  $\alpha$ .

表 3 各手法での平均 nDCG, MAP, P@1. 括弧内の数値は標準偏差を表している

Table 3 Average nDCG, MAP and P@1 for each method. Note that value in a bracket indicates the standard deviation.

	nDCG	MAP	P@1
MindReader	0.675 (0.150)	0.340 (0.196)	0.375 (0.490)
Proposed	<b>0.739</b> (0.134)	<b>0.460</b> (0.217)	<b>0.500</b> (0.506)
Cohen's d	0.456	0.583	0.254

表 4 著者が選択した例を用いた場合の各手法での平均 nDCG, MAP, P@1

Table 4 Average nDCG, MAP and P@1 for each method, where examples selected by the first author were used as a query.

	nDCG	MAP	P@1
MindReader	0.729 (0.135)	0.423 (0.212)	<b>0.525</b> (0.505)
Proposed	<b>0.765</b> (0.122)	<b>0.504</b> (0.208)	0.500 (0.506)
Cohen's d	0.264	0.383	0.049

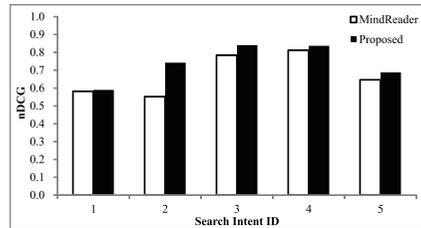


図 7 nDCG の比較  
Fig. 7 Comparison of nDCG.

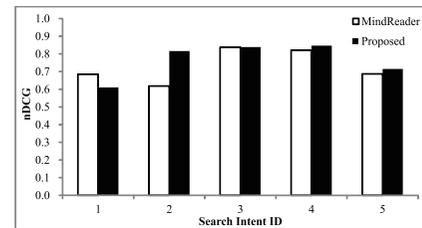


図 8 著者が選択した例を用いた場合の nDCG の比較  
Fig. 8 Comparison of nDCG, where examples selected by the first author were used as a query.

羅的に選択した例の平均数は 7.6 と十分な入力を得られているため、頑健性による提案手法の優位性が薄れたのだと考えられる。ただし、依然として提案手法の nDCG と MAP は MindReader の結果を上回っている。

各検索意図での nDCG を図 7 に、また比較のために本論文の著者によって時間の制約なく網羅的に選択された例を用いたときの nDCG を図 8 にそれぞれ示す。図 7 の結果ではどの検索意図においても提案手法が上回っているが、特に検索意図 2「辛い料理を出す飲食店」では大幅な改善が見られる。これは選択された例が少なかったものの、例を選択する地域においてそもそも辛い料理を提供する飲食店が少なかったため、選択された例と選択されなかった例との差異を強調したときに、辛いという属性(カテゴリ、もしくは、説明文中の語)が際立って強調されたためだと考えられる。この差は図 8 においても変化は見られない。これは図 3 および 4 から読み取れるように、両入力でその選択数に大きな違いがなかったためであると思われる。

被験者の選択例による結果(図 7)においては、検索意図 1 での両手法の nDCG は他の意図と比べ低くなっている。図 3 に見られるように、与えられた例の数がきわめて少なく、そのため、他の検索意図と比べて十分な精度が得られていないと考えられる。特に 1 例しか与えられていなかった場合には、既存手法では推定できず、提案手法でも選ばれなかった例との差異を強調することでしか距離尺度を推定できないため、極端な精度低下を招いていると考えられる。一方で、選択された例が多くなった場合、すなわち、図 8 においては、ややその精度は改善され、かつ、MindReader が提案手法よりも高い nDCG を得ている。

検索意図 3 および 4 (ユーザと地域によって基準が異なりやすい意図)では、地理コンテキストを考慮、すなわち、選ばれなかった例との差異を強調することで、それぞれ「その

地域での海産物は何であるか」、「その地域での高級とは何であるか」を考慮することにより、既存手法より高い精度が得られていると考えられる。ここで述べた効果は 4.2.1 項で説明した、地理コンテキストを考慮する利点に相当する。検索意図 5 は複雑な検索意図であるため、両手法とも精度は低下しているが、この意図においても地域のコンテキストである、「その土地特有のものは何であるか」を距離尺度推定に含めることにより、より良い結果を提案手法は得ていると考えられる。

図 7 と図 8 の nDCG を比較すると分かるように、どの意図においても多くの選択例が与えられた場合に改善が見られる。また、提案手法と MindReader の性能差も選択例の増加にともない減少していることが分かる。このことは、MindReader は少ない例が与えられた場合にはうまく機能しない、という我々の主張とも合致している。

## 6. ま と め

本論文では、知らない場所の地理オブジェクトを検索するために、知っている場所のオブジェクトを例示することによって検索する手法を提案した。また、このような検索時に必要な類似度について、2つの問題を提起した。すなわち、ユーザごとに異なる動的な類似性尺度の問題、異なる分野間の類似度計算の問題である。本論文ではこの中でも、前者の問題、動的な類似性尺度の推定を、選択された例と選択されなかった例の差異を増幅させることによって行った。比較実験で示したように、提案手法は既存の距離尺度推定手法よりも優れ、より頑健であることを示した。

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、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」、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)、文部科学省科学研究費補助金特別研究員奨励費「アナロジーに基づく情報検索に関する研究」(研究代表者: 加藤誠, 課題番号: 22・4687)、文部科学省科学研究費補助金若手研究(B)「オンデマンド利用を目的とする Web からの知識発見に関する研究」(研究代表者: 大島裕明, 課題番号: 21700105)、文部科学省科学研究費補助金若手研究(B)「時間変化するオブジェクト情報の Web からの収集と管理方式の研究」(研究代表者: 小山聡, 課題番号: 21700106)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) Angluin, D. and Smith, C.: Inductive inference: Theory and methods, *ACM Computing Surveys (CSUR)*, Vol.15, No.3, pp.237–269 (1983).
- 2) Ashwin, T.V., Gupta, R. and Ghosal, S.: Adaptable similarity search using non-relevant information, *Proc. VLDB 2002*, pp.47–58 (2002).
- 3) Ben-Tal, A. and Nemirovski, A.: *Lectures on modern convex optimization*, Society for Industrial and Applied Mathematics (2001).
- 4) Bischoff, K., Mandl, T. and Womser-Hacker, C.: Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006, *Evaluation of Multilingual and Multi-modal Information Retrieval*, Vol.4730, pp.946–953 (2010).
- 5) Cardoso, N. and Silva, M.: Query expansion through geographical feature types, *Proc. 4th ACM Workshop on Geographical Information Retrieval*, pp.55–60 (2007).
- 6) Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C. and Slaney, M.: Content-based music information retrieval: Current directions and future challenges, *Proc. IEEE*, Vol.96, No.4, pp.668–696 (2008).
- 7) Chang, N. and Fu, K.: Query-by-pictorial-example, *IEEE Trans. Softw. Eng.*, Vol.6, pp.519–524 (1980).
- 8) Chen, Y., Suel, T. and Markowetz, A.: Efficient query processing in geographic web search engines, *Proc. SIGMOD 2006*, pp.277–288 (2006).
- 9) Cohen, J.: *Statistical power analysis for the behavioral sciences*, Lawrence Erlbaum (1988).
- 10) Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol.41, No.6, pp.391–407 (1990).
- 11) Hiramoto, R. and Sumiya, K.: Web information retrieval based on user operation on digital maps, *Proc. GIS 2006*, pp.99–106 (2006).
- 12) Ishikawa, Y., Subramanya, R. and Faloutsos, C.: MindReader: Querying databases through multiple examples, *Proc. VLDB 1998*, pp.218–227 (1998).
- 13) Jones, C., Alani, H. and Tudhope, D.: Geographical Information Retrieval with Ontologies of Place, *Spatial Information Theory*, Vol.2205, pp.322–335 (2001).
- 14) Larson, R.R.: Geographic Information Retrieval and Spatial Browsing, *GIS and Libraries: Patrons, Maps and Spatial Information*, pp.81–124 (1996).
- 15) Lieberman, M., Samet, H., Sankaranarayanan, J. and Sperling, J.: STEWARD: Architecture of a spatio-textual search engine, *Proc. GIS 2007*, pp.186–193 (2007).
- 16) Markowetz, A., Chen, Y., Suel, T., Long, X. and Seeger, B.: Design and implementation of a geographic search engine, *Proc. WebDB 2005*, pp.19–24 (2005).
- 17) Rui, Y., Huang, T. and Mehrotra, S.: Content-based image retrieval with relevance feedback in MARS, *Proc. IEEE International Conference on Image Processing*, pp.815–818 (1997).
- 18) Smeulders, A., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.12, pp.1349–1380 (2000).
- 19) Tversky, A.: Features of similarity, *Psychological Review*, Vol.84, No.4, pp.327–352 (1977).
- 20) Yoshitaka, A. and Ichikawa, T.: A survey on content-based retrieval for multimedia databases, *IEEE Trans. Knowledge and Data Engineering*, Vol.11, No.1, pp.81–93 (1999).
- 21) Zloof, M.: Query-by-Example: A data base language, *IBM Systems Journal*, Vol.16, No.4, pp.324–343 (1977).

(平成 23 年 4 月 11 日受付)

(平成 23 年 9 月 12 日採録)



加藤 誠 (学生会員)

京都大学大学院情報学研究科博士後期課程在学中, また, 日本学術振興会特別研究員 (DC1)。2008 年京都大学工学部情報学科卒業。2009 年同大学大学院情報学研究科修士課程修了。情報検索の研究に従事。



大島 裕明 (正会員)

京都大学大学院情報学研究科社会情報学専攻助教。2007 年京都大学大学院情報学研究科博士後期課程修了。博士 (情報学)。主に Web, 情報検索, データベースの研究に従事。電子情報通信学会, 日本データベース学会, ACM 各会員。



小山 聡 (正会員)

北海道大学大学院情報科学研究科複合情報学専攻准教授。1994年京都大学工学部数理工学科卒業。1996年京都大学大学院工学研究科数理工学専攻修士課程修了。2002年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了。博士(情報学)。1996~1998年日本電信電話株式会社。2001~2002年日本学術振興会特別研究員(DC2)。2002~2007年京都大学大学院情報学研究科社会情報学専攻助手。2003~2004年スタンフォード大学 Visiting Assistant Professor。2007~2009年京都大学大学院情報学研究科社会情報学専攻助教。2009年北海道大学大学院情報科学研究科複合情報学専攻准教授、現在に至る。機械学習、データマイニング、情報検索等に興味を持つ。2005年度人工知能学会論文賞。2009年度日本データベース学会上林奨励賞。電子情報通信学会、人工知能学会、日本データベース学会、IEEE、ACM、AAAI 各会員。



田中 克己 (正会員)

1974年京都大学工学部情報工学科卒業。1976年同大学大学院修士課程修了。1979年神戸大学教養部助手。1986年同大学工学部助教授。1994年同大学工学部教授(情報知能工学専攻)。1995年同大学大学院自然科学研究科情報メディア科学専攻専任教授。2001年京都大学大学院情報学研究科社会情報学専攻教授、現在に至る。工学博士。主にデータベースとマルチメディア情報システムの研究に従事。人工知能学会、日本ソフトウェア科学会、IEEE Computer Society、ACM 等各会員。