地物間の距離を考慮した動的な類似性尺度に基づく地理情報例示検索

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社会情報学専攻 〒 606-8501 京都府京都市左京区吉田本町
†† 北海道大学大学院情報科学研究科複合情報学専攻 〒 060-0814 札幌市北区北 14 条西 9 丁目
E-mail: †{kato,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp, ††oyama@ist.hokudai.ac.jp

あらまし　本論文では，例示による地理情報検索を提案する．提案手法では，検索対象の地理情報は与えられた例との類似性に基づいてランキングされる．本論文で提案する類似性尺度は，入力された正例，また，暗に与えられた負例によって変化し，ユーザごとに異なる類似性尺度を少ない入力からでも推測することが可能である．

キーワード　地理情報検索，類似性尺度

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 [1] や食べログ [2]，ホテル情報サイトの Booking.com [3] などが例として挙げられる．また，いくつかの大手検索サイトも [4]〜[6]，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，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**地理エンティティ**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はにこれらの検索方法は適切ではない．また，両者ともに検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づらい．例えば，「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，その場所に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

我々は地理エンティティ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 [7]．例示検索はデータベースへの問い合わせ方法 [8]，また，マルチメディア検索で用いられる検索方法 (Content-based retrieval [9]〜[11]) の 1 つであり，盛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挙げてそれに共通する要素似基づいて検索する，すなわち，帰納的な検索であると考えることができる．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挙げられる．実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．

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) 中の地理エンティティが検索される．検索結果は地図の下方に表示される．

例示して検索するという考えは，自然に地理エンティティ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，左がユーザの良く知っている地域や地元の地図，右が地理エンティティ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

プ内でどの地理エンティティ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エンティティ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ジーを用いることができる．我々はこれまでにもアナロジーを検索に持ち込むことを提案している [12]．アナロジー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，もし自分が良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，良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，自分が良く知る地域の例を選択することは容易であろう．また，当然のことながら，良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エンティティをランキングするために，エンティティ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．すなわち，

（1） 類似性尺度はユーザ，コンテキストによって異なる．

（2） 異なる分野におけるエンティティ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，更には，そのコンテキストによって変化する [13]．別の言い方をすれば，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．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．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エンティティ間の類似度計算が問題となる．全く異なる集合に属する 2 つのエンティティ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，ここで挙げた 2 つの問題の内，前者にのみ着目して議論を続ける．我々は Ishikawa らが提案した MindReader [14] の枠組みに従い，そのモデルを地理エンティティ間の類似性尺度推定に応用し，地理エンティティ間の距離を考慮することでより頑健な推定手法を提案する．最終的に，人手で構築されたテストセットを用いた比較実験によって，本論文で提案する推定手法の有効性を示した．

2. 関 連 研 究

2.1 地理情報検索

Markowetz ら [15], [16] 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 [17] 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 と Sumiya [18] 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．地理エンティティ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 [19]．しかしながら，大量の地理エンティティ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 [9]〜[11], [20]．Multimedia Analysis and Retrieval System (MARS) [21] 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 [14]) を提案している．Ashwin ら [22] 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用ると，不適合なデータに近いものも得られ続けてしま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

この節では地理エンティティの例示検索について述べる．まず，地理エンティティ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

地理エンティティは位置情報を含め，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．

地理エンティティ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：$\boldsymbol{R} = \{a_1, a_2, \ldots, a_M, \mathrm{pos}\}$．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エンティティと一線を画す特徴である．地理エンティティのスキーマ $\boldsymbol{R}$

は飲食店やホテル，ランドマークなどの各地理エンティティクラスに対して事前に定義される．

地理エンティティスキーマ $\boldsymbol{R}$ に対するエンティティ $\mathbf{e}$ は N 次元空間の点で表現される：$\mathbf{e} = (\mathbf{e}_{a_1}, \mathbf{e}_{a_2}, \ldots, \mathbf{e}_{a_M}, \mathbf{e}_{\mathrm{pos}})$．ここで $N = \sum_{i=1}^{M} n_i + n_{\mathrm{pos}}$ であり，n_i はベクトル $\mathbf{e}_{a_i}$ の次元数，$\mathbf{e}_{a_i}$ は属性 a_i の属性値である．ベクトル $\mathbf{e}_{a_i}$ は飲食店の予算などのスカラー値や，飲食店の紹介文などを tf-idf ベクトルなどで表現できる．$\mathbf{e}_{\mathrm{pos}}$ は緯度と経度で表わされる位置ベクトルであり，その次元数は $n_{\mathrm{pos}} = 2$ である．

3.2 検索モデル

地理エンティティの例示検索では良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エンティティを取得し，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 $\boldsymbol{E}_s$，ターゲットドメイン $\boldsymbol{E}_t$ と呼ばれる．

この 2 つのドメインは全てのエンティティ集合 $\boldsymbol{E}$ の部分集合である：

$$\boldsymbol{E}_s \subset \boldsymbol{E},\ \boldsymbol{E}_t \subset \boldsymbol{E}. \tag{1}$$

2 つのドメインは一般的に互いに疎である：$\boldsymbol{E}_s \cap \boldsymbol{E}_t = \phi$．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．

ターゲットドメイン $\boldsymbol{E}_t$ のエンティティを検索するために，ユーザはソースドメイン $\boldsymbol{E}_s$ の部分集合をクエリとして選択する．従って，地理エンティティの例示検索での全クエリ集合 $\boldsymbol{Q}$，及び，検索対象データ $\boldsymbol{D}$ は以下のように定義される．

$$\boldsymbol{Q} = \mathfrak{P}(\boldsymbol{E}_s),\ \boldsymbol{D} = \boldsymbol{E}_t. \tag{2}$$

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 $\boldsymbol{Q}$ と検索対象データ $\boldsymbol{D}$，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 $\boldsymbol{Q}$ と検索対象データ $\boldsymbol{D}$ から実数 $\mathbb{R}$ への関数である：$\mathrm{Rank} : \boldsymbol{Q} \times \boldsymbol{D} \to \mathbb{R}$.

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．地理エンティティの例示検索では，ターゲットドメインのエンティティは与えられたクエリ $\boldsymbol{Q}_i \subset \boldsymbol{Q}$ 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 $\mathbf{d}_j \in \boldsymbol{D}$ の $\mathrm{Rank}(\boldsymbol{Q}_i, \mathbf{d}_j)$ 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．従って，ここではランキング関数を以下のように定義する：

$$\mathrm{Rank}(\boldsymbol{Q}_i, \mathbf{d}_j) = \exp\left(-\mathrm{dist}(\boldsymbol{Q}_i, \mathbf{d}_j)\right). \tag{3}$$

項 $\mathrm{dist}(\boldsymbol{Q}_i, \mathbf{d}_j)$ は選択されたエンティティ集合 $\boldsymbol{Q}_i$ と検索対象データ $\mathbf{d}_j$ との距離である．$\exp(-x)$ は単に距離の逆を取っているだけであり，この形式は距離の逆を $[0, 1]$ に正規化している．

4. 類似性尺度の推定

前節では地理エンティティ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るべきでない理由を示す．そして，地理エンティティ例示検索の特徴を踏まえて，距離尺度推定手法を適用させる方法を提案する．本節で我々が用いる記号を表 1 に列挙する．

表 1　本節で用いる記号.

記号	定義
$\boldsymbol{E}$	全エンティティ集合.
$\boldsymbol{E}_s$	ソースドメイン ($\boldsymbol{E}$ の部分集合．)
$\boldsymbol{E}_t$	ターゲットドメイン ($\boldsymbol{E}$ の部分集合．)
$\boldsymbol{Q}$	全クエリ集合 ($\boldsymbol{E}_s$ の全部分集合．)
$\boldsymbol{Q}_i$	クエリ，$\boldsymbol{Q}$ の要素．(エンティティ集合．)
$\boldsymbol{D}$	検索対象データ ($= \boldsymbol{E}_t$．)
$\mathbf{e}$	エンティティ，$\boldsymbol{E}$ の要素.
N	エンティティ $\mathbf{e}$ の次元数.
$\mathbf{q}_k$	$\boldsymbol{Q}_i$ の要素 (エンティティ．)
g_k	$\mathbf{q}_k$ に対するスコア (0/1，もしくは，段階的な値．)
$\mathbf{d}_j$	$\boldsymbol{D}$ の要素.
Rank()	ランキング関数 ($\boldsymbol{Q} \times \boldsymbol{D} \to \mathbb{R}$．)
$\mathbf{m}$	理想的なクエリベクトル (N 次元ベクトル．)
$\mathbf{W}$	距離関数を決定する $N \times N$ 行列.
dist()	距離関数 ($\mathfrak{P}(\boldsymbol{E}) \times \boldsymbol{E} \to \mathbb{R}$．)

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：$\mathrm{dist}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}$．またこれは，$(\mathbf{x}-\mathbf{y})^T\mathbf{I}(\mathbf{x}-\mathbf{y})$ 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：$(\mathbf{x}-\mathbf{y})^T\mathbf{D}(\mathbf{x}-\mathbf{y})$．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．例えば，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あるということが考慮できる．更に，対角行列の代わりに対称行列 $\mathbf{W}$ 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．例えば，値段属性とジャンル属性の両方の次元が重要であるということを表現できる．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 $\mathbf{W}$ を推定することに帰着できる．

MindReader ではクエリ $\boldsymbol{Q}_i$ とデータ $\mathbf{d}_j$ との距離関数は以下のように定義される：

$$\mathrm{dist}(\boldsymbol{Q}_i, \mathbf{d}_j) = (\mathbf{d}_j - \mathbf{m})^T\mathbf{W}(\mathbf{d}_j - \mathbf{m}). \tag{4}$$

ただし，ユーザがある**理想的なクエリベクトル** $\mathbf{m}$ と対称行列 $\mathbf{W}$ 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 $\boldsymbol{Q}_i$ から $\mathbf{m}$ と $\mathbf{W}$ を推定することであると言い換えられる．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 $\boldsymbol{Q}_i$ と理想的なクエリ $\mathbf{m}$ 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．この基本的なアイデアから導かれる仮定に従うと，理想的なクエリ $\mathbf{m}$ と理想的な距離尺度を決定する行列 $\mathbf{W}$ は以下の最小化問題を解くことで得られる：

$$\min_{\mathbf{m},\mathbf{W}} \sum_{\mathbf{q}_k \in \boldsymbol{Q}_i} g_k (\mathbf{q}_k - \mathbf{m})^T \mathbf{W} (\mathbf{q}_k - \mathbf{m}). \quad (5)$$

ただし，以下の制約を満たす：

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

ここで，$\det(\mathbf{W})$ は行列 $\mathbf{W}$ 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，$[1, 5]$ などでも良い．)

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 $\mathbf{m}$ 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 $\mathbf{W}$ 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 [21] 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 $\mathbf{W}$ 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．2 つの問題とはすなわち，

（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．

（2） 画像特徴量などとは異なり，エンティティ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，

である．

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．極少数の例を入力したユーザが，どの属性が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，検索タスクのために入力されるキーワードなど，検索時に入力される情報は実に少ない．特に content-based search，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だろう．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．

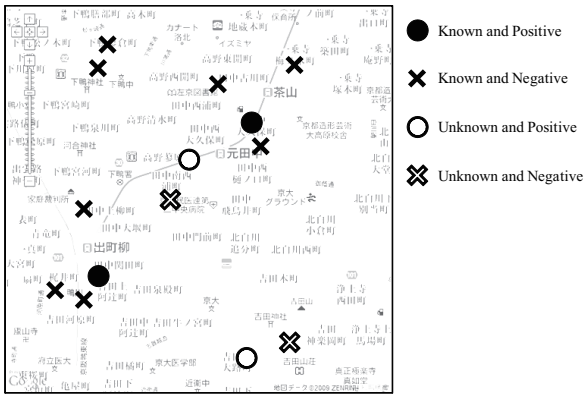

図 2　非明示的な負例.

2 つ目の問題は，人物や飲食店などといったエンティティには常識的な，標準的な距離尺度が存在していることである．例えば，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．我々があるエンティティ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 非明示的な負例

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．そこで，非明示的な負例を仮定する．非明示的な負例とはすなわち，明示的には負例として与えられないが，ユーザの行動からおそらく不適合であるだろうと推測される負例である．選択され与えられた例，すなわち，正例だけでなく，さらなる入力，非明示的な負例によって距離尺度推定はより頑健になる．

地理エンティティの例示検索では，ユーザはエンティティを検索するために地図内で正例だけを選択する．選択されなかった例は 2 種類に分類される．すなわち，ユーザが知らなかった例，そして，知っている負例である．もちろん，ユーザは知っていてかつ適合した例だけを選択し，他の例は選択されない．ユーザが知らない例をユーザが思い描く距離尺度の推定に使うことはできない．ユーザ自身が知らない例は，ユーザについて何の情報も語らないためである．非明示的な負例とはすなわち，ユーザが知っていたにも関わらず選択しなかった負例のことである．

特に地理エンティティ検索においては，非明示的な負例は推定可能である．図 2 にその例を示す．黒い丸はクエリとして選択された例で，それらは明らかに既知かつ適合した例である．白い丸と白いバツ印はユーザが知らない例であり，この議論では無視できる．目的は，選択されなかった，既知かつ不適合な例である黒いバツ印を推測することである．そのために，我々は地理情報の例示検索における正例選択について，単純な仮定を立てる．

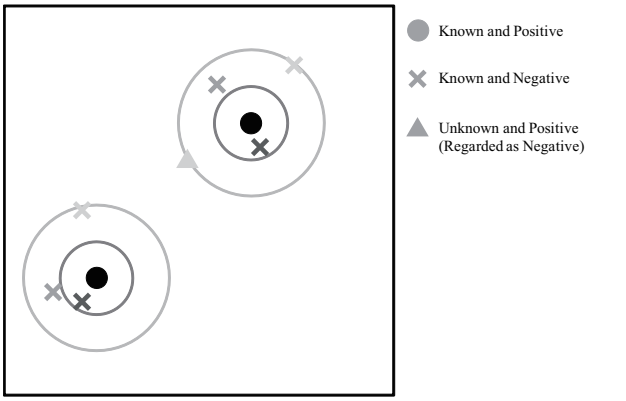

図 3　非明示的な負例の推定．濃い色は高い確率を，薄い色は低い確率を表している．

［**仮定** 1］ 選択された正例と地理的に近い例は，それが知られていたにも関わらず意図的に選択されなかった．すなわち，それらは非明示的な負例である．

図 2 中の選択された正例の周りには，選択されなかった正例がほとんど存在しないことが期待される．これは知っていてかつ適合しているような例はユーザに選択されているはずであるし，既知である例 (この場合は，選択された正例) の周辺のエンティティはまた既知であることが期待されるためである．この仮定は入力インタフェースに強く依存し，また，地理的な要素にも深い関係がある．一般に，良く知っている地域において，ある一定の範囲内で 1 つのエンティティしか知らないことは希であり，その範囲でいくつかのエンティティを知っていることが期待される．加えて，選択した例の近くにある適合したエンティティは，その例が選択された際に認識されているため，おそらく見過ごされないだろう．この仮定は一見，地理エンティティの例示検索に特化したものであるように見えるが，非明示的なフィードバックは情報検索の分野で広く用いられており評価されているものである [23], [24].

次に我々は選択されなかった例がどれくらいの確率で非明示的な負例であるかを推定する必要がある．直感的には，選択された例に近ければ高い確率で，既知かつ選択されなかった例，すなわち，非明示的な負例であることが期待できる．そこで，2 次元混合ガウス分布を既知かつ選択されなかった例の推定に用いる．選択された正例集合 $\boldsymbol{Q}_i$ が与えられたとき，あるエンティティ $\mathbf{e}$ が非明示的な負例である確率 $P(F|\mathbf{e}, \boldsymbol{Q}_i)$ は以下のように定義される：

$$P(F|\mathbf{e}, \boldsymbol{Q}_i) = \frac{P(F, \mathbf{e}|\boldsymbol{Q}_i)}{P(\mathbf{e}|\boldsymbol{Q}_i)} \tag{7}$$

$$P(F, \mathbf{e}|\boldsymbol{Q}_i) = \sum_{\mathbf{q}_k \in \boldsymbol{Q}_i} \frac{g_k}{N_g} \mathcal{N}(\mathbf{e}_{\mathrm{pos}}|\mathbf{q}_{k,\mathrm{pos}}, \boldsymbol{\Sigma}). \tag{8}$$

ここで $N_g = \sum_{\mathbf{q}_k \in \boldsymbol{Q}_i} g_k$ であり，これは選択された例 $\mathbf{q}_k$ に対するスコア g_k の和である．2 次元ベクトル $\mathbf{e}_{\mathrm{pos}}$ と $\mathbf{q}_{k,\mathrm{pos}}$ は緯度・経度を含む位置情報である．確率 $P(F, \mathbf{e}|\boldsymbol{Q}_i)$ は分散 $\boldsymbol{\Sigma}$ を持ち，その平均が選択された例の位置であるようなガウス分布を混合したものである．そして，それぞれのガウス分布は選択された例に対するスコアによって重み付けされている．また，$P(\mathbf{e}|\boldsymbol{Q}_i)$ が一様分布であると仮定すると以下のように簡略化できる：$P(F|\mathbf{e}, \boldsymbol{Q}_i) \propto P(F, \mathbf{e}|\boldsymbol{Q}_i)$．図 3 に非明示的な負例である確率を示す．濃い色は高い確率を表す．この例のように，未知かつ適合しているエンティティも非明示的な負例として扱われることがある．しかし，仮定 1 が理にかなっていれば，誤って判定する確率は低いため，入力の増加によって得られる利益の方が大きいであろう．

推定された非明示的な負例を式 5 へと単純に組み込むことはできない．最も単純な方法は，選択されなかったエンティティ $\mathbf{e}$ に対して，確率 $P(F|\mathbf{e}, \boldsymbol{Q}_i)$ のマイナスの値をスコア g_k として，式 5 の最小化問題を解くことである．しかし残念ながら，スコア g_k が負の場合，MindReader の手法では妥当な結果が得られない．うまく働かない例として，1 つの次元しか持たないようなデータセットを考える．1.0 近辺のデータを取得したいがために，0.9，1.0, 1.1 を正例として，100 を負例として与えた場合，式 5 に従えば，理想的なクエリベクトル (この例の場合ではスカラーであるが) を計算するとその値は遙かに 1 よりも小さくなる．なぜなら，理想的なクエリベクトルは，与えられた例をそのスコア g_k によって重み付けをしたものの平均となるためである．さらに，スコア g_k が負のとき，距離関数は負の値すら取り得る．

2.2 節で述べたように，Ashwin ら [22] は適合フィードバック中での負例を用いて検索領域を推定する手法を提案している．しかしながら，我々の提案手法や content-based retrieval では，負例を明示的に与えない．そのため，明確に正例と負例を分離するような手法は今回用いることができない．

以上の議論を含めて，我々は式 5 に非明示的な負例を取り入れて再定義を行った：

$$\min_{\mathbf{W}} \sum_{\mathbf{e}_k \in \boldsymbol{E}_s} v_k (\mathbf{e}_k - \mathbf{m})^T \mathbf{W} (\mathbf{e}_k - \mathbf{m}). \tag{9}$$

ただし，以下の制約を満たす：

$$\|\mathbf{W}\| = 1,\ w_{ij} \geqq 0. \tag{10}$$

ここで，w_{ij} は行列 $\mathbf{W}$, v_k の要素である．また，v_k は，選択された例に対してはその例に付与されたスコアで，選択されなかった例に対してはパラメータ α によって重み付けされた $-P(F|\mathbf{e}_k, \boldsymbol{Q}_i)$ である．

$$v_k = \begin{cases} g_k & \mathbf{e}_k \in \boldsymbol{Q}_i \\ -\alpha P(F|\mathbf{e}_k, \boldsymbol{Q}_i) & \text{otherwise.} \end{cases} \tag{11}$$

(12)

そして，$\mathbf{m}$ はクエリ $\boldsymbol{Q}_i$ の (g_k により重み付けされた) 要素の平均である．

$$\mathbf{m} = \frac{1}{N_g} \sum_{\mathbf{q}_k \in \boldsymbol{Q}_i} g_k \mathbf{q}_k. \tag{13}$$

主な変更は，和の範囲，理想的なクエリベクトル，行列 $\mathbf{W}$ の制約条件に見られる．第 1 に我々は，選択されなかった例を含めた全てのエンティティと理想的なクエリとの距離を最小化

した．選択されなかった例は $-\alpha P(F|\mathbf{e}_k, \boldsymbol{Q}_i)$ で重み付けされている．すなわち，選択された例と理想的なクエリは類似するように，選択されなかった例と理想的なクエリは類似しないように，距離尺度は決定される．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．(未だにこの制約は本質的ではない．)

4.2.2 距離尺度の過学習へのペナルティ

これまではエンティティの属性は皆同等に扱われてきた．しかし，エンティティ間の距離を計算するときに，画像特徴量とは異なり，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：

$$\min_{\mathbf{W}} \sum_{\mathbf{e}_k \in \boldsymbol{E}_s} v_k (\mathbf{e}_k - \mathbf{m})^T \mathbf{W} (\mathbf{e}_k - \mathbf{m}) + \frac{\rho}{2} \|\mathbf{W} - \hat{\mathbf{W}}\|^2 . \quad (14)$$

ただし，式 10 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

この節ではまず地理情報例示検索の実装について述べ，4 節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，ベースライン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

我々が用いた地理エンティティはグルメ情報検索サイト「ぐるなび」[1] から得られたものであり，地図インタフェースは Google Maps API (注1) を利用して実装した．

ぐるなび Web サービス (注2) 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 などの手法によってベクトル化される．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う．そのため，*latent semantic analysis* [25] 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，例えば，{ 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 次元へと圧縮した．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値である．従って，飲食店情報のスキーマは $\boldsymbol{R} = \{\text{text}, \text{category\_label}, \text{budget}, \text{pos}\}$ となり，エンティティの次元数は $50 + 20 + 1 + 2 = 73$ (ただし，位置を表す 2 次元は距離計算の時には無視される．)

(注1)：http://www.google.com/apis/maps/

(注2)：http://api.gnavi.co.jp/

5.2 距離行列の推定

制約条件 (式 10) の下での最適化問題 (式 14) 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 [26]．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 14 は凸計画問題となり，パラメータ数はエンティティ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．

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．

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，被験者らそれぞれに 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．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全ての正例を選択し，全ての負例を選択しなければよく，非明示的な負例が存在しないことになる．図 4 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全て $g_k = 1$ としている．

次に被験者を 2 人ずつにわけ，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，全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，平均カッパ係数も表 2 に併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わかる．

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．

5.4 ベースライン手法

我々は比較のために，2 種類のベースラインを用意した．

（1） **Standard**: 距離行列を標準的な距離行列 ($\hat{\mathbf{W}}$) に固定した手法．

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．

検索意図	
ID	内容
1	1,000 円ぐらいの飲食店
2	辛い料理を出す飲食店
3	魚介類を出す飲食店
4	高級な飲食店
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		
ID	地域	エンティティ数
1	東京	55
2	名古屋	55
3	大阪	59
4	札幌	51

ターゲットマップ			
ID	地域	エンティティ数	カッパ係数の平均
1	京都	49	0.94
2	神戸	50	0.86

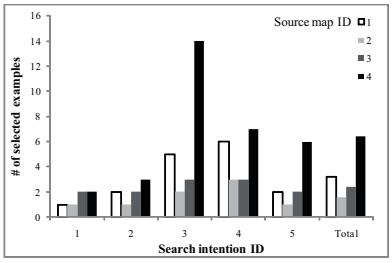

図 4　各検索意図，各ソースマップでクエリとして選択された例の数．

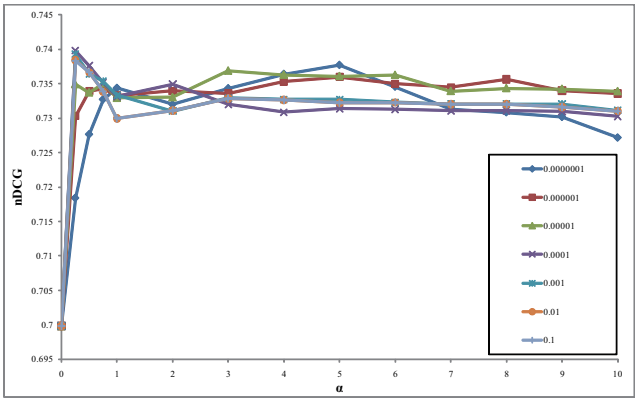

図 5　各 σ，α に対する nDCG.

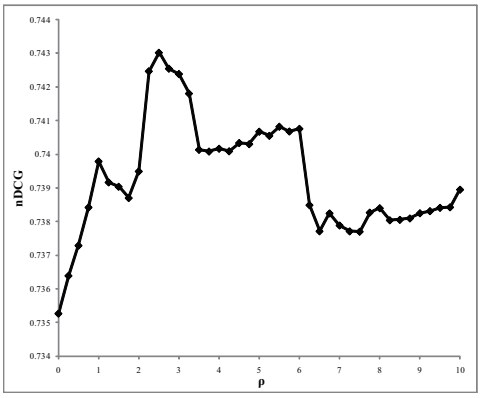

図 6　各 ρ に対する nDCG.

（2） **MindReader**.
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．

5.5 実験結果

ベースライン手法との比較実験を行う前に，提案手法中に現れた 3 つのパラメータについて最適な値を決定する．3 つのパラメータとはすなわち，2 次元ガウス分布の共分散行列 $\mathbf{\Sigma}$，非明示的な負例に対する重み α，そして，標準的な距離行列によるペナルティの重み ρ である．特に，$\mathbf{\Sigma}$ と α については関係が強いため，同時に最適な値を決める．また，$\mathbf{\Sigma}$ を簡略化して $\sigma\mathbf{I}$ とする．これは非明示的な負例である確率は楕円形でなく，円形に広がると考えられるためである．さらに，式 12 中の $v_k = -\alpha P(F|\mathbf{e}_k, \boldsymbol{Q}_i)$ の代わりに，$v_k = -2\pi\sigma|\boldsymbol{Q}_i|\alpha P(F, \mathbf{e}_k|\boldsymbol{Q}_i)$ を用いる．これは，$2\pi\sigma P(F, \mathbf{e}_k|\boldsymbol{Q}_i)$ の部分が最大で 1 になるようにし ($\boldsymbol{Q}_i = \{\mathbf{e}_k\}$ のとき)，また，これに選択された例の数 $|\boldsymbol{Q}_i|$ をかけることで，$\alpha = 1$ のときに，式 14 中で選択された例と同等になるように調整するためである．

それぞれのパラメータに対する normalized discounted cumulative gain (nDCG) を図 5 と図 6 に示す．図 5 では $\alpha = 0$ のときが非明示的な負例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わかる．nDCG が極大値を取るのは，$\sigma = 10^{-4}$ と $\alpha = 0.25$ のときであり，式 14 の最適化問題では選択された例よりも同等以下に扱うべきであることがわかる．$\sigma = 10^{-4}$ では，平均からおおよそ 1km 程度離れた時に値が半分になるようなガウス分布が得られる．また，図 6 では $\rho = 0$ のときが標準的な距離行列によるペナルティがない場合，$\rho = \infty$ のときが標準的な距離行列に一致する．そのため，$\rho = 2.5$ で極大値を取ることは，ペナルティ項が有効に働いていることを意味する．

ベースラインとの比較実験の結果を図 7 (nDCG) と図 8 (mean average precision (MAP)) に示す．被験者 2 人の評価値の平均が 4 以上の場合，適合であると判断し MAP を求めている．また，提案手法のパラメータは以下のように設定している：$\sigma = 10^{-4}$，$\alpha = 0.25$，$\rho = 2.5$．我々の提案手法は nDCG と MAP 両方において総合的に最も良い結果を得ている．一方で MindReader は最も悪い結果となった．これは入力として与えられた例の数が，推定するパラメータ数に比べ非常に小さかったことが原因であると考えられる．特にあまり例示がされなかった検索意図 1 についてはその影響が顕著である．提案手法，MindReader ともに固定した標準的な距離 Standard よりも劣っている．一方で多くの例が与えられた検索意図 3,4 に対し，提案手法は適切に距離尺度を推定していると考えられる．

興味深い結果が検索意図 4，5 で見られる．検索意図 4 は絶対的な検索意図でなく，ソースマップの種類によって変わり得る，相対的な検索意図となっている．提案手法でこの相対性を捉えられたのは，選択された例 (予算が高い) とその周辺の選択

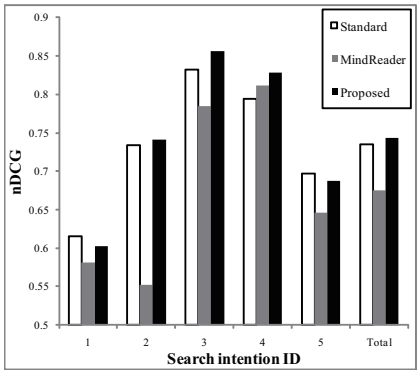

図 7　nDCG の比較.

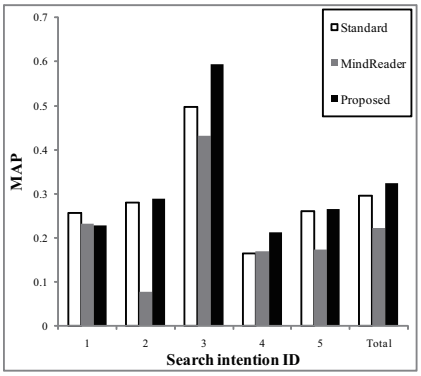

図 8　MAP の比較.

されなかった例 (予算が低い) に 2 分した場合，予算を重要視すればこの 2 つの集合を適切に 2 分できる (式 14 を最適化できる) と推定できたためである．一方で検索意図 5 は今回の提案手法ではうまくいかないような例であり，標準的な距離による手法 Standard とあまり結果が変わらない．その土地の特産物，といったような性質は地域ごとに異なる．そのため，冒頭で述べた「異なる分野におけるエンティティ間の類似度計算方法」が必要になると考えられる．

6. ま と め

本論文では，知らない場所の地理エンティティを検索するために，知っている場所のエンティティ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，非明示的な負例を用いることによって行った．例示されたエンティティのみならず，選ばれなかった例も利用することによって，より頑健な尺度推定を可能にした．

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（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．

文　献

[1] “ぐるなび”, http://www.gnavi.co.jp/.
[2] “食べログ”, http://tabelog.com/.
[3] “Booking.com”, http://www.booking.com/.
[4] “Google maps”, http://maps.google.com/.
[5] “Yahoo! Local Maps”, http://maps.yahoo.com/.
[6] “Bing maps”, http://www.bing.com/maps/.
[7] D. Angluin and C. Smith: “Inductive inference: Theory and methods”, ACM Computing Surveys (CSUR), **15**, 3, pp. 237–269 (1983).
[8] M. Zloof: “Query-by-Example: A data base language”, IBM Systems Journal, **16**, 4, pp. 324–343 (1977).
[9] A. Yoshitaka and T. Ichikawa: “A survey on content-based retrieval for multimedia databases”, IEEE Transactions on Knowledge and Data Engineering, **11**, 1, pp. 81–93 (1999).
[10] A. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain: “Content-based image retrieval at the end of the early years”, IEEE Transactions on pattern analysis and machine intelligence, **22**, 12, pp. 1349–1380 (2000).
[11] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes and M. Slaney: “Content-based music information retrieval: current directions and future challenges”, Proc. of the IEEE, **96**, 4, pp. 668–696 (2008).
[12] M. P. Kato, H. Ohshima, S. Oyama and K. Tanaka: “Query by analogical example: Relational search using web search engine indices”, Proc. of CIKM 2009, pp. 27–36 (2009).
[13] A. Tversky: “Features of similarity”, Psychological Review, **84**, 4, pp. 327–352 (1977).
[14] Y. Ishikawa, R. Subramanya and C. Faloutsos: “Mindreader: Querying databases through multiple examples”, Proc. of VLDB 1998, pp. 218–227 (1998).
[15] A. Markowetz, Y. Chen, T. Suel, X. Long and B. Seeger: “Design and implementation of a geographic search engine”, Proc. of WebDB 2005, pp. 19–24 (2005).
[16] Y. Chen, T. Suel and A. Markowetz: “Efficient query processing in geographic web search engines”, Proc. of SIGMOD 2006, pp. 277–288 (2006).
[17] M. Lieberman, H. Samet, J. Sankaranarayanan and J. Sperling: “STEWARD: architecture of a spatio-textual search engine”, Proc. of GIS 2007 (2007).
[18] R. Hiramoto and K. Sumiya: “Web information retrieval based on user operation on digital maps”, Proc. of GIS 2006, pp. 99–106 (2006).
[19] R. R. Larson: “Geographic information retrieval and spatial browsing”, GIS and Libraries: Patrons, Maps and Spatial Information, pp. 81–124 (1996).
[20] N. Chang and K. Fu: “Query-by-pictorial-example”, IEEE Transactions on Software Engineering, **6**, pp. 519–524 (1980).
[21] Y. Rui, T. Huang and S. Mehrotra: “Content-based image retrieval with relevance feedback in MARS”, Proc. of IEEE International Conference on Image Processing, Vol. 81, pp. 815–818 (1997).
[22] T. V. Ashwin, R. Gupta and S. Ghosal: “Adaptable similarity search using non-relevant information”, Proc. of VLDB 2002, pp. 47–58 (2002).
[23] D. Kelly and J. Teevan: “Implicit feedback for inferring user preference: a bibliography”, ACM SIGIR Forum, Vol. 37, pp. 18–28 (2003).
[24] R. White, I. Ruthven and J. Jose: “A study of factors affecting the utility of implicit relevance feedback”, Proc. of SIGIR 2005, pp. 35–42 (2005).
[25] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman: “Indexing by latent semantic analysis”, Journal of the American society for information science, **41**, 6, pp. 391–407 (1990).
[26] A. Ben-Tal and A. Nemirovski: “Lectures on modern convex optimization”, Society for Industrial and Applied Mathematics (2001).