地理情報と内容情報の近接性を考慮した地理情報検索

Introducing Word Proximity into Geographic Information Retrieval Ranking

戸田浩之[*1] 安田宜仁[*1] 松浦由美子[*1] 片岡良治[*1]
Hiroyuki Toda Norihito Yasuda Yumiko Matsuura Ryoji Kataoka

[*1]日本電信電話株式会社, NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

Geographic Information Retrieval (GIR) is a retrieval method using geographic constraint in addtion to conventional keyword query. To improve the accuracy, this paper proposes a GIR ranking method utlizing proximity between content words that match against keyword query and geowords that relate to geographic constraint. We conducted a evaluation using Japanese weblog articles and the result shows effectiveness of the proposed method.

1. はじめに

検索エンジンの利用シーンの多様化およびモバイルデバイスを利用した Web アクセスの普及などにより, 検索エンジンへの問い合わせにおいて特定の位置に関する問い合わせが増加している. 検索エンジンへの全アクセス数の約 20%は特定の場所に関係する問い合わせであると言われており [Zhou 05] 位置に関係する問い合わせに的確に答えることは検索エンジンにとって重要な課題となっている.

この課題に対して, 特定の場所に関する問い合わせに高い精度で検索する事を目的とした地理情報検索技術 (GIR: Geographic Information Retrieval) が研究されている [安田 08, Toda 08, Martins 05, Vaid 05]. GIR では, 通常の検索エンジンのようにキーワードと文書の関係だけで検索するのではなく, 文書中の地名表現を座標や領域に抽象化した上で, クエリで示された位置 (地理クエリ) およびキーワード (内容クエリ) と, 文書との関係を用いて検索を行う. 本技術を利用することで, 「新宿のフレンチレストランが知りたい」という情報要求に対して「歌舞伎町のフレンチレストラン」という記述を含む文書が検索可能となるなど, 従来の検索手法にはないメリットがある.

一方, 一般的なテキスト検索のランキング方法として, クエリ中のキーワードの文書中での近接性を考慮する手法が注目を浴びている [Rasolofo 03, Tao 07]. この手法はクエリが複数のキーワードから構成される場合に, それらのキーワードがより近くで出現する文書を優先しようというものである. 様々なテストセットでの評価から, その有効性が確認されており, 検索結果のランキングにおいて重要な要素となっている.

本稿ではこの近接性を GIR のランキングに導入する方法について示す. GIR では上述の通りクエリが地理クエリおよび内容クエリの 2 つの部分から構成されており, 内容クエリと一致するキーワードおよび地理クエリと関連性の高い地名表現が文書中で近接して出現することは適合文書を見つける上で重要な手がかりとなると考えられる. 提案手法では, まず内容クエリと一致するキーワードと地理クエリと関連する地名表現の各ペアについて文書中での距離を特定する. 次に, この距離およびペアを構成するキーワードと地名表現のクエリに対する重要度から各ペアの近接性スコアを算出する. 最後に, 各ペアの近接性スコアから文書の近接性スコアを算出し, 既存のスコアと組み合わせることで文書のスコアを算出する.

連絡先: 戸田浩之, NTT サイバーソリューション研究所, 横須賀市光の丘 1-1, toda.hiroyuki@lab.ntt.co.jp

以下, 2 章では関連研究を示し, 3 章では今回の提案手法の元となる我々の提案する GIR を示す. 4 章で提案手法, 5 章で評価について示し, 6 章でまとめる.

2. 関連研究

テキスト情報検索の分野では, 検索クエリ中のキーワードが文書内で近接して出現していることを考慮して検索結果のランキングを行う手法が提案されている. Rasolofo[Rasolofo 03] らは, クエリ中のキーワードペアの文書中での距離およびそれらキーワードの重要度を元にした近接性のスコアと BM25 とを組み合わせたスコア算出法を提案し, その有効性を示している. TaoTao ら [Tao 07] は, キーワード間の近さを測る指標が検索精度にどのような影響を及ぼすかについて分析している. クエリ中のキーワード全てが含まれる区間の距離やキーワードペアの平均距離, 最少距離などについて比較評価した結果, キーワードペアの最少距離を利用した場合がもっとも高い検索精度を示すことを明らかにしている.

このように近接性の効果は明らかであるが, 我々の知る限り GIR において近接性を考慮した手法は提案されていない.

我々の手法は, 上記に示す Rasolofo および TaoTao の研究を参考に, クエリが内容クエリと地理クエリの 2 つの部分で表現されるという特徴を持つ GIR において近接性を考慮しようというものである.

3. 地名表現の含意する広さを考慮した GIR

以下に提案手法の元とした我々が提案する GIR 法 [安田 08, Toda 08] について示す.

3.1 前処理プロセス

一般に GIR を実現するためには, 内容クエリに基づく検索を行うテキストインデクスと地理クエリに基づく検索を行う地理インデクスを作成する. テキストインデクスは従来の全文検索のインデクスを利用する. 地理インデクスは, テキスト情報から地名表現を抽出するジオパーサと地名表現を緯度経度に変換するジオコーダを利用して文書を解析し生成する [Clough 05]. また, これらの情報に加え, 我々の手法では地理インデクスに地名辞書から取得した地名が含意する広さの情報も含んでいる.

3.2 検索プロセス

GIR ではクエリとして内容クエリ Qc と, 地理クエリ Qg を受け付ける. 前者は従来の全文検索と同様にキーワードで表現

され, 後者はユーザの興味あるエリアの中心点の座標 Qg_p と許容できる距離 Qg_d で表現される. Qg_p は GPS や基地局情報を元にした測位により得られるものである. また Qg_d はユーザの移動手段等によって変更される値であると考えており, 徒歩の場合には 1km としたり, 車で移動している場合には比較的広めに 20km 等と設定するパラメータである.

クエリを受け取ると, 地理クエリを満たす地名表現を少なくともひとつ含む文書を特定し, それら文書の地理スコアと内容スコアを算出する. 内容スコアとしては TF-IDF や BM25 等の一般的な手法を利用する. 地理スコアは以下に示す各地名表現に与えられる地名表現スコアを元に算出する.

3.2.1 地名表現スコア

ある文書中に含まれる i 番目の地名表現 g_i の地名表現スコアは, 地理クエリで指定される場所からの距離と地名表現が含意する広さに基づく逓減項から算出される.

地理クエリ Qg と地名表現 g_i の距離は次式で定義される.

$$d(g_i, Qg) = \begin{cases} d_{inner} & (Qg_p \text{ が } g_i \text{ の示す範囲内}) \\ d_{inner} + d_{edge}(g_i, Qg) & (\text{範囲外}) \end{cases}$$

ここで d_{inner} は定数であり, $d_{edge}(g_i, Qg)$ は g_i の示す領域の外縁からの最短距離である. また, 地名の含意する広さに基づく逓減項は, 地名表現 g_i の広さ $e(g_i)$ に反比例する値とする. 以上より, 地名表現スコアは以下の式で与えられる.

$$S_{geo}(g_i, Qg) = \frac{1}{d(g_i, Qg)} \cdot \frac{1}{e(g_i)} \quad (1)$$

3.2.2 文書のスコア

文書 D の地理スコアは地名表現スコアの和で与えられる. 算出式を以下に示す.

$$Sg(D, Qg) = \sum_{g_i \in G(D)} S_{geo}(g_i, Qg)$$

ここで $G(D)$ は, 文書 D 内に出現する地名表現の集合を示す.

そして, 最終的な文書のスコアはこの地理スコアと内容スコアの積で表現される.

$$S(D, Qc, Qg) = Sg(D, Qg) \cdot Sc(D, Qc) \quad (2)$$

内容スコアおよび地理スコアの両方のスコアが高いことが重要であると考え, 最終的なスコアはそれらの積で算出している.

4. 提案手法

ここでは, GIR において近接性を考慮した検索結果ランキングを行う手法を示す.

4.1 手法概要

近接性を考慮した情報検索では, クエリ q に対する文書 d のスコアは一般に以下の式により定義される [Rasolofo 03, Tao 07].

$$Score(d, q) = SQ(d, q) + SP(d, q)$$

$SQ(q, d)$ は BM25 や TF-IDF 等の通常のスコアであり, $SP(q, d)$ が近接性に関するスコアを示す項である.

これを GIR に適用するため, $SQ(q, d)$ および $SP(q, d)$ の項を以下に示すよう変更する. $SQ(q, d)$ は (2) 式で定義されるスコア $S(D, Qc, Qg)$ で置き換える. $SP(q, d)$ に相当する近接項は, 文書 D 内に出現するキーワード $k_i (\in K(Qc, D))$ と地名表現 $g_j (\in G(D))$ の文書内での距離に基づくものとして定義する. ここで, $K(Qc, D)$ は文書 D 中に出現する内容クエリ Qc に含まれるキーワードの集合を示し, $G(D)$ は文書 D 中に出現する地名表現の集合を示す. 一般的に近接性の考慮と言った場合にはクエリ中の複数のキーワード間の近接性を意味しており, GIR においても内容クエリが複数キーワードで表される場合もあるため, この近接性を考慮することも意味があるが, 今回はこの複数キーワード間の近接性については扱わない.

近接項の算出手順を以下に示す.

1. 文書 D 中のすべての「キーワード $k_i (\in K(Qc, D))$, 地名表現 $g_j (\in G(D))$」ペア間の距離 $\delta(k_i, g_j)$ を特定
2. すべての「キーワード, 地名表現」のペアについて, 上記で求めた距離 $\delta(k_i, g_j)$ およびキーワード k_i の重要性, 地名表現 g_j の重要性を元に, クエリ (Qc, Qg) に関連性が高く, より近接しているペアに高いスコアを与える重み付き近接性スコア $\eta(Qc, Qg, k_i, g_j)$ を算出
3. 上記で求めた「キーワード, 地名表現」ペアの重み付き近接性スコア $\eta(Qc, Qg, k_i, g_j)$ を元に, 文書 D の近接項 $S_{prox}(D, Qc, Qg)$ を算出.

最終的に, クエリ (Qc, Qg) に対する文書 D のスコアは以下の式で定義される[*1].

$$Score(D, Qc, Qg) = S(D, Qc, Qg) + S_{prox}(D, Qc, Qg) \quad (3)$$

以下では, 上記で示した近接項の算出手順の各ステップ毎に詳細を示す.

4.2 キーワードと地名表現間の距離の定義

タイトルや本文, 章や節などの構造を持たない文書を考えた場合には, キーワード k_i と地名表現 g_j の距離 $\delta(k_i, g_j)$ は, その間の文字数や形態素数等の単純な距離として定義できる. しかし, 検索対象となる文書には Web 文書のようにタイトルと本文などの構造が存在する文書も多く存在する. そこで, 本稿ではタイトルと本文という構造を持つ文書を扱うこととする.

このようなタイトルと本文を持つ文書において, キーワード k_i と地名表現 g_j の組み合わせが出現するパターンは次に示す 4 つが考えられる. 両方がタイトルに出現する場合 (TT), 両方が本文に出現する場合 (BB), キーワードがタイトル, 地名表現が本文にある場合 (TB), その逆の場合 (BT) である.

このうちキーワードと地名表現の両方が同じ構造中に出現する場合 (TT, BB) は単純な距離として $\delta(k_i, g_j)$ を算出できる. つまり, 以下のように定義する.

$$\delta(k_i, g_j) = |pos(D, k_i) - pos(D, g_j)|$$

ここで, $pos(D, x)$ は文書 D 中での文字列 x が出現する位置を表す関数である.

一方, キーワードと地名表現が分かれて出現する場合 (TB,BT), 上記のように直接 $\delta(k_i, g_j)$ を算出できず, 文書構造の意味を考慮する必要がある. ここでタイトルと本文の関係を考えると, タイトルは本文全体に関係する情報であり, タイトルは本文のいずれの場所とも一定の近接性を持つと考えられる. そこで, 本研究では TB,BT の場合の距離を以下のように定義する.

$$\delta(k_i, g_j) = l$$

*1 実際のスコア計算では, $S(D, Qc, Qg)$ および $S_{prox}(D, Qc, Qg)$ には検索結果中でのそれぞれの最大値で正規化した値を利用する.

l はタイトルと本文の距離を示すパラメータである.

5. 章の評価では $\delta(k_i, g_j)$ の単位を文字数とし, l を変化させ実験する.

以下では, 上記 $\delta(k_i, g_j)$ に基づく近接性について述べる.

4.3 重み付き近接性スコアの定義

近接性を考慮するとは, 単純に考えれば文書中でキーワードと地名表現が近接している文書に対して, 付加的な重要度を与えるものである. しかし, ここで重要な事は, その近接が起こることによって, 文書の適合度合が高まりそうな場合に高いスコアを与えることであり, キーワードと地名表現が近接していたとしても文書の適合度合が高まらないと考えられる場合には高いスコアは与えるべきではない.

この点について Rasolofo[Rasolofo 03] らは, 近接性スコアを算出する際に, キーワード間の距離だけではなく, キーワードの重要度も考慮し, クエリ中で重要なキーワードが近接している場合により高い値を与えている.

GIR においては, 内容クエリにおけるキーワードの重要性および地理クエリに対する地名表現の関連性がそれぞれ高く, かつそれらが近接する場合に高い近接性スコアを算出すべきであると考えられる. そこで, クエリ (Qg, Qc) において, ある文書中でのキーワード k_i と地名表現 g_j の重み付き近接性スコアを以下のように定義する.

$$\eta(Qc, Qg, k_i, g_j) = \pi(k_i, g_j) \cdot S_{key}(k_i, Qc) \cdot S_{geo}(g_j, Qg)$$

$S_{geo}(g_j, Qg)$ は地理クエリ Qg に対する地名表現 g_j の関連度を示し (1) 式で定義される. また, $S_{key}(k_i, Qc)$ はキーワード k_i の重要度を示し, IDF 値等を利用する. また, $\pi(k_i, g_j)$ は距離 $\delta(k_i, g_j)$ を元にした近接性スコアである. 本研究では近接スコア $\pi(k_i, g_j)$ の算出には以下の関数を利用する.

$$\pi(k_i, g_j) = log(\alpha + exp(-\delta(k_i, g_j)/\beta))$$

この関数は Taotao ら [Tao 07] が利用したものを元に, パラメータ β を追加したものである.

式中の α および β は, 近接性スコア $\pi(k_i, g_j)$ の減衰をコントロールするパラメータであり, α は減衰時のスコアの最小値を, β は減衰の程度をコントロールする. α および β を変化させた場合の, $\pi(k_i, g_j)$ の変化を図 1 に示す.

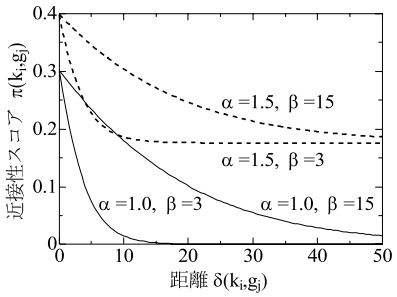

図 1: 近接スコアの例

4.4 近接性スコアを利用した近接項の算出法

(3) 式に示した文書 D の近接項 $S_{prox}(D, Qc, Qg)$ は上記に示したキーワードと地名表現のペアに対する重み付き近接性スコアを元に算出する. 近接性スコアの利用法には様々な方法が考えられるが, ここでは [Tao 07] において高い検索精度を示した, 文書中で最大の近接性スコアをとるペアのスコアのみを利用する方法とした. そこで, クエリ (Qg, Qc) に対する, ある文書 D の近接項を以下のように定義する.

$$S_{prox}(D, Qc, Qg) = \max_{g_j \in G(D), k_i \in K(Qc, D)} \eta(Qc, Qg, k_i, g_j)$$

表 1: 実験に用いたクエリの詳細

Query Type		
地理クエリ: Qg	中心点: Qg_p	札幌駅, 東京駅, 新宿駅, 栄, 梅田, 三宮, 八丁堀, 天神
	許容距離: Qg_d	1 km
内容クエリ: Qc		カレー, ラーメン, 焼肉, ハンバーガー, 紅茶, うどん コーヒー, ケーキ

5. 評価

本評価では 3. 章に示した手法をベースライン (BASELINE) とし, 4. 章に示した近接性を考慮した場合の精度変化の分析およびパラメータ変化による精度変化の傾向分析を行う.

5.1 評価リソース

評価では BM25 ランキングを行う全文検索エンジンと [平野 08] に基づくジオパーサおよびジオコーダを利用した. また, 地名辞書として, 街区レベル位置参照情報[*2]を利用し, 地名の含意する広さを取得した. 検索対象のコレクションには goo ブログ[*3]から収集したブログ記事を利用した. このブログ記事はブログ開設時にブログの主要なトピックとして「食べ歩き」もしくは「地域情報」を選んだブログ著者によって書かれたものである. 総文書数は約 30 万件で, 18,565 人の著者に書かれた記事で構成される. 表 1 に示す地理クエリと内容クエリを組み合わせた 64 のクエリで評価を行った. これらのクエリは携帯端末を片手に周辺にある店やサービスを探すユーザを想定している. 適合性判定データを作成するため, それぞれのクエリ毎に検索結果のプールを作成した. それぞれのプールは評価で利用した全ての実験条件の検索結果上位 50 件に出現した文書から構成される. プールに含まれる文書の総数は 2958 である. 適合性判定は 3 人の被験者の評価により作成した. 評価基準等の適合性判定データ作成の詳細は [安田 08] に示す.

5.2 パラメータ設定

本手法で可変なパラメータは 2 種類あり, ひとつは 4.2 節で示した TB,BT の場合のキーワードと地名表現の距離を規定するパラメータ l であり, もうひとつは 4.3 節で示した近接性の減衰に関するパラメータ α, β である.

l については $l = 1, 10$ と TB,BT の場合に近接性を考慮しない条件 ($l = \infty$ に相当) の 3 条件で, α, β は図 1 に示した範囲 ($\alpha = 1.0$〜$1.5, \beta = 3$〜15) で変化させ, 実験を行う.

5.3 結果と考察

図 2 に TB,BT の場合に近接性を考慮しない条件, 図 3 , 4 に l を 1,10 と変化させた場合の実験結果を示す. 縦軸は MAP(Mean Average Precision) を示す. 横軸は α を示し, 各 β 値毎にグラフを描画している.

l を変化させた場合の結果から, 全体的に $l = 1$ とした場合の MAP が高く, l を 10 に設定した場合の精度は TB,BT の場合に近接性を考慮しない条件と大きな変化が無いことがわかる.

*2 http://nlftp.mlit.go.jp/isj/
*3 http://blog.goo.ne.jp/

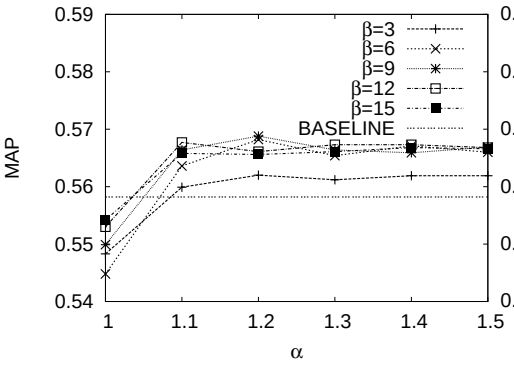

図 2: 検索精度 (TB,BT 時の近接性考慮無し)

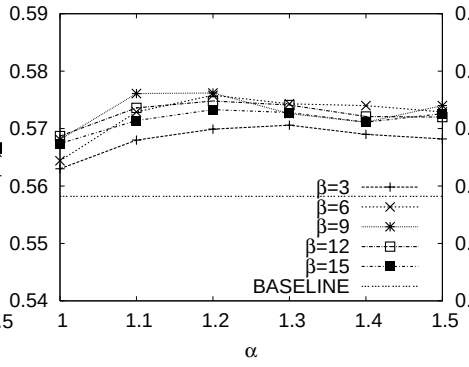

図 3: 検索精度 ($l = 1$)

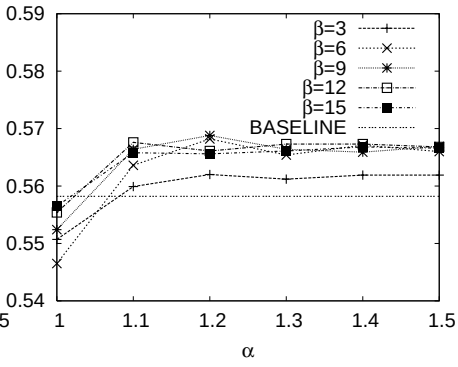

図 4: 検索精度 ($l = 10$)

表 2: 検索精度 (P@X は検索結果上位 X 件の適合率を表す)

手法	MAP	R 精度	P@5	P@10
BASELINE	0.5582	0.5080	0.5688	0.5047
TB,BT の場合に近接性の考慮無し ($\alpha = 1.2, \beta = 9, l = 1$)	0.5688	0.5171	0.5969	0.5031
提案法 ($\alpha = 1.2, \beta = 9, l = 1$)	0.5762**	0.5237*	0.5969	0.5125

*は, BASELINE との間で有意な精度向上が見られた場合を示す. 検定にはウィルコクソン符号付き順位検定 (危険率 5%) を利用. 特に**は $p < 0.01$ を示す.

これは, l が大きすぎると近接性スコアが小さくなりすぎ, スコアに与える影響が少なくなった為であると考えられる. 一方, $l = 1$ は TB,BT の場合に本文内で隣り合っているのと同等の近接性を与えることであり, タイトルの情報が場所に寄らず本文の情報と関連するという事が再認識できる結果となった.

次に α の変化に伴う精度変化を見ると, $\alpha = 1.0$ では精度が低いが, $\alpha = 1.1$〜1.3 辺りでピークの精度を示し, その後変化しないか, なだらかに低下していく傾向が見られる. $\alpha = 1.0$ で精度が低い理由として, このパラメータ設定の場合, キーワードと地名表現間に一定以上距離が開くと近接性スコアがほぼ 0 になる点が考えられる. これにより, 地理クエリとの関連の強い地名表現がキーワードと離れた位置にある文書と, 関連が弱い地名表現がキーワードと近い位置にある文書が存在する場合, 後者に高いスコア付けがなされる可能性があり, いくつかの条件でベースラインを下回る精度となってしまったのではないかと考えられる.

β についてみると, $\alpha = 1.0$ の場合を除けば β を大きくするにつれ精度が向上し, $\beta = 15$ でやや低下するという傾向がある. β を大きくすることで精度が改善していることを考えると, $\beta = 3$ で精度が低いのは近接性スコアの減衰が急激すぎるためと考えられる. また, $\beta = 15$ の場合は逆に近接スコアの減衰が緩やかすぎるため, 精度向上幅が小さくなったと考えられる.

表 2 には近接性を考慮しないベースライン, TB, BT の場合近接性を考慮しない手法 ,4. 章で示したすべての近接性を考慮した提案法の検索精度を示す. この結果から提案法は全ての評価基準でベースラインを上回ること, TB, BT の場合に近接性を考慮しなければベースラインと有意な差がないが, それらを考慮することで有意な精度向上を得られる事がわかった.

6. まとめ

本稿では, GIR において内容クエリおよび地理クエリのそれぞれと関連のある情報の文書内における近接性を考慮し, 検索結果のランキングに利用する方法を提案した. ブログ記事を利用した評価では, 近接性を考慮しない手法と比較して, 精度が向上することを確認した. また, 文の構造を考慮することが有益であるとの知見が得られた. 今後は文の構造をより有効に利用する手法に取り組む予定である.

参考文献

[Clough 05] Clough, P.: Extracting metadata for spatially-aware information retrieval on the internet, in *GIR '05*, pp. 25–30 (2005)

[Martins 05] Martins, B., Silva, M. J., and Andrade, L.: Indexing and ranking in Geo-IR systems, in *GIR '05*, pp. 31–34 (2005)

[Rasolofo 03] Rasolofo, Y. and Savoy, J.: Term Proximity Scoring for Keyword-Based Retrieval Systems, in *ECIR '03*, pp. 207–218 (2003)

[Tao 07] Tao, T. and Zhai, C.: An exploration of proximity measures in information retrieval, in *SIGIR '07*, pp. 295–302 (2007)

[Toda 08] Toda, H., Yasuda, N., Matsuura, Y., and Kataoka, R.: Incorporating place name extents into Geo-IR ranking, in *CIKM '08*, pp. 1489–1490 (2008)

[Vaid 05] Vaid, S., Jones, C. B., Joho, H., and Sanderson, M.: Spatio-textual Indexing for Geographical Search on the Web, in *SSTD '05*, pp. 218–235 (2005)

[Zhou 05] Zhou, Y., Xie, X., Wang, C., Gong, Y., and Ma, W.-Y.: Hybrid index structures for location-based web search, in *CIKM '05*, pp. 155–162 (2005)

[安田 08] 安田 宜仁, 戸田 浩之: 検索位置のごく周辺を対象とした地理情報検索, 人工知能学会論文誌, Vol. 23, No. 5, pp. 364–373 (2008)

[平野 08] 平野 徹, 松尾 義博, 菊井 玄一郎: 地理的距離と有名度を用いた地名の曖昧性解消, 情処全国大会, 3D-7 (2008)