F-018

クエリ分布を考慮した一般化ピボット法の距離定義による特性評価

Evaluation by Distance Definitions of Calculating Generalized Pivots Considering Query Distribution

小林 えり † 斉藤 和巳 † 池田 哲夫 † 青山 一生 ‡ 服部 正嗣 ‡
Eri Kobayashi Kazumi Saito Tetsuo Ikeda Kazuo Aoyama Takashi Hattori

1. はじめに

近年，Web 上には多量のデータが蓄積されており，与えられたクエリに類似するオブジェクトをデータベース等から求める類似検索技術の重要性はますます高まっている．この類似検索は距離空間を対象とすることが多々ある．距離空間はオブジェクト集合とオブジェクト間の非類似度を表す距離関数から成る．距離関数とは距離公理（non-negativity, identity, symmetry, triangle inequality）を満たす関数を指す．また，検索対象である距離空間は高次元であることが多く，その場合，オブジェクト間の距離を求めるためには膨大な計算量を要する．

多量のデータ（オブジェクト集合）の距離空間を高速に検索するためには，距離計算量を削減する必要がある．この要請に応える方法の 1 つとして，距離空間中の不動点であるピボットを利用する方法がある．効果的なピボット集合を選択する手法として Bustos らはインクリメンタル法（逐次選択法）を提案している [Bustos 03]．このインクリメンタル法は，高効率なピボット生成用に定義した目的関数を最大化するピボットをオブジェクトから逐次選択することで，ピボット集合を生成する方法である．

これに対し，距離空間の任意の点をピボットとして求める一般化ピボット法も提案されている [Kimura 09, Kobayashi 14]．一般に，オブジェクト集合の中からピボット集合を選択する場合と比較し，距離空間の任意の点としてピボット集合を求めれば目的関数値の向上が自然に期待できる．

距離計算量削減に対して，より効率的なピボット集合を生成するためには，入力されるクエリに付随する情報を目的関数に反映することが考えられる．即ち，クエリはユーザの嗜好やトレンドなどによる偏りや分布を有すると考えられるため，クエリ分布を考慮した目的関数を設計し，分布を有するオブジェクト部分集合で目的関数を最適化することである．つまり，このようなクエリ分布を学習データとして用いることにより検索のさらなる高速化が期待できる．

これまでにクエリ分布を考慮した一般化ピボット法による類似検索高速化法が提案され [Kobayashi 15]，分布考慮の有効性が確認されている．しかしながら，文献 [Kobayashi 15] では L1 距離定義のみしか評価していなかった．我々は，L1 距離定義で高効率であったこの方法を L2 距離定義に適用したところ，効率が著しく低くなるという驚くべき結果を得た．よって，本稿では L1，L2 距離定義による結果の違いに着目し，それぞれの性能を評価する．

2. 類似検索問題

類似検索問題には様々な問題設定があるが，本稿ではレンジクエリ問題を扱う．レンジクエリ問題は，オブジェクト集合 $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ とクエリ $q_m \in Q = \{q_1, \cdots, q_M\}$ とレンジ r が与えられたとき，$\mathbf{q}_m$ と $\mathbf{x}_n$ の距離 $d(\mathbf{x}_n, \mathbf{q}_m)$ が r 以下となるようなオブジェクト集合を求める問題である．

Bustus らの提案したピボット法（インクリメンタル法）は，オブジェクト間の距離計算回数を削減するために，一部のオブジェクトを選定してピボット集合を求める．最適なピボット集合は $P_B^* = \arg\max_P \mathcal{F}_B(P)$ より求まる．

$$\mathcal{F}_B(P) = \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} D(\boldsymbol{x}_m, \boldsymbol{x}_n; P \subset X) \quad (1)$$

$$D(\boldsymbol{x}_m, \boldsymbol{x}_n; P \subset X) = \max_{1 \le k \le K} |d(\boldsymbol{x}_m - \boldsymbol{p}_k) - d(\boldsymbol{x}_n - \boldsymbol{p}_k)| \quad (2)$$

ただし，$K = |P|$, $\boldsymbol{x}_n, \boldsymbol{x}_m \in X$ である．式 2 の max 関数内は，オブジェクトペア $\{\boldsymbol{x}_m, \boldsymbol{x}_n\}$ の距離に対する，ピボット $\boldsymbol{p}_k$ から各オブジェクトまでの距離を用いて算出した下限値であり，式 2 は $\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_K$ を用いて算出した最大下限値である．ここで留意すべきは，Bustos らのピボット選択法が，検索問題のクエリがデータベースのオブジェクト分布と独立同分布から生成されるという仮定に基づいている，ということである．ここで，式 2 中のオブジェクトの一方はクエリ $\boldsymbol{q}_m$ とみなすことができる．最大下限値が r より大きいオブジェクト集合を $E = \{\mathbf{x}_n \mid D(\mathbf{q}_m, \mathbf{x}_n; P) > r\}$ とすると，E に属すオブジェクト集合に対しては距離計算が不要となるため，類似検索の計算時間の短縮が期待できる．

3. クエリ分布を考慮した一般化ピボット法

クエリ分布がデータベースのオブジェクト分布とは異なると仮定する．この仮定の下，Bustos らの方法を拡張して一般化ピボット法による目的関数を定義すると以下の式で定義する．最適なピボット集合は $P^* = \arg\max_P \mathcal{F}(P, Q)$ より求まる．

$$\mathcal{F}(P, Q) = \sum_{n=1}^{N} \sum_{m=1}^{M} D(\boldsymbol{q}_m, \boldsymbol{x}_n; P \subset \mathcal{X}) \quad (3)$$

$$D(\boldsymbol{q}_m, \boldsymbol{x}_n; P \subset \mathcal{X}) = \max_{1 \le k \le K} |d(\boldsymbol{q}_m - \boldsymbol{p}_k) - d(\boldsymbol{x}_n - \boldsymbol{p}_k)| \quad (4)$$

ただし，$\mathcal{X} = \{x | x \in \boldsymbol{R}^H\}$ であり，$\boldsymbol{R}^H$ は与えられた制約により限定されたユークリッド空間を指す．ここで，ピボット集合を X ではなく，$\mathcal{X}$ の部分集合としている点に留意されたい．Bustos らの手法では，与えられたオブジェクト集合からピボットを逐次選択するのに対し，一般化ピボット法では，R^H の任意の点をピボットとして構築する．

ここでクエリ集合 Q をユーザのクエリ分布を示す学習クエリ集合とすると，式 3 で求まるピボット集合 P は，ユーザの嗜好に対応したピボット集合になると考えられる．ユーザの今後検索するであろうクエリ集合（未知クエリ集合）もまた，学習クエリと同様なクエリ分布を持つと考えられ，よって，先ほど生成したピボット集合を用いれば，未知クエリ集合に対して効果的な枝刈りが見込める．

クエリ分布を考慮した L1，L2 距離定義に基づく一般化ピボット法のアルゴリズムについて，詳しくは文献 [Kobayashi 15, Kimura 09] を参照されたい．

4. 実験評価

4.1. 実験データ

実験データとして，YahooNews の記事データを用いた．各記事を形態素解析して得られた単語頻度ベクトルをオブジェクトベクトルとする．単語頻度ベクトルとは，記事内にその形態素（単語）の出現回数を要素とするベクトルを指す．記事データセットの総オブジェクト数は 324,528，総出現ターム数（ベクトルの次元数）は 91,522 である．YahooNews は“ 国内 ”,“ 経済 ”,“ エンタメ ”,“ 生活 ”,“ 地域 ”,“ サイエンス ”,“ スポーツ ”,“ 世界 ”の 8 つのジャンルに分類され，一つの記事が複数のジャンルを持つことはない．本実験では，これら 8 ジャンル中の一種についてのみ検索を行うユーザを想定し実験評価を行う．

4.2. 実験設定

ピボット性能評価のため，オブジェクト集合を 3 つに分割した．1 つ目は，ピボット生成の目的関数を最適化する際に用いるユーザのクエリ分布を考慮した学習クエリ集合 Q_i，2 つ

†静岡県立大学大学院経営情報イノベーション研究科
‡NTT コミュニケーション科学基礎研究所

ジャンル		国内	経済	エンタメ	地域	生活	サイエンス	スポーツ	世界
地域	L1	59.5	43.4	36.3	73.7	43.8	38.2	46.8	36.8
	L2	4.3E-07	4.3E-07	0.0E+00	1.7E-05	0.0E+00	0.0E+00	0.0E+00	1.5E-05
スポーツ	L1	19.1	18.9	17.9	21.8	19.1	19.0	84.7	20.3
	L2	2.17E-02	2.12E-02	1.92E-02	2.01E-02	2.01E-02	2.21E-02	1.03E-01	2.08E-02

表 1: ジャンルを考慮した枝刈り率による評価 (%表示)

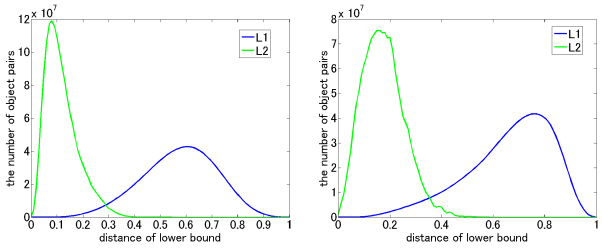

図 1: 下限値 $D(\boldsymbol{q}_m, \boldsymbol{x}_n; P \subset \mathcal{X})$ の分布

目は検索性能評価用クエリ集合 S_j，3 つ目は検索対象データベース（オブジェクト集合）$X_{i,j} \subset X$, $(X_{i,j} = X \setminus \{Q_i \cup S_j\})$ である．ただし，学習・評価クエリ集合は一種類のジャンルの記事のみで構成し，添え字 i, j でそのジャンルを表し，同一ジャンルの学習，評価クエリ集合内での記事の重複はないものとする $(S_i \cap Q_i = \emptyset)$.

本実験では，まず，2 つの異なる距離空間，L1 距離定義と L2 距離定義において，学習クエリ集合 Q_i とデータベース（検索対象オブジェクト集合）$X_{i,j}$ を用いてピボット集合を生成する．次に，あるクエリ分布に基づく評価クエリ集合 S_j を用いて，生成したピボットの類似検索性能（距離計算回数の削減率）を評価し比較する．P_i は学習クエリのジャンルに i を使用した，ジャンル i に特化したピボット集合であることを示す．実験では学習クエリ数 $|Q_i| = 5,000$, 評価クエリ数 $|S_j| = 5,000$ とし，よって検索対象データベースのサイズは $|X_{i,j}| = 314,528$ となる．

4.3. 評価指標

実験では，学習クエリ Q_i，評価クエリ S_j のジャンルを変化させた，8 × 8=64 パターンでのレンジクエリ問題における類似検索性能を評価した．紙面の関係上，最も性能の低い結果と高い結果となった“ 地域 ”,“ スポーツ ”ジャンルの結果のみを掲載する．

評価指標として枝刈り成功率（距離計算回数の削減率）を用いる．式 4 で計算される，ピボット集合 P_i で距離計算を省略することができたオブジェクトの集合を E_i とすると，枝刈り成功率は $|E_i|/|X_{i,j}|$ で表される．この値が大きいほど，クエリ分布を考慮して学習したピボット集合 P_i によって効果的な枝刈りが行えたことになる．

レンジ距離 r は，KNN-Search 法を用いて設定した．本稿では 3 個程度の類似記事が入る距離，$r = 0.5$ を採用した．ここで，オブジェクトペア間の距離は 1.0 に正規化されており，最大距離は 1.0 であることに注意されたい．

4.4. 実験結果

表 1 は，ピボット数 K を $K = 10$ に設定し，L1 距離定義と L2 距離定義との各々の距離空間で検索を実行した場合の枝刈り成功率を示す表である．行はピボット生成に用いる学習クエリのジャンルを，列は評価クエリのジャンルを表している．仮定に従えば，学習クエリと評価クエリのジャンルが同じ $(i = j)$ のときに枝刈り成功率が最も高く，それ以外 $(i \neq j)$ では低くなることが予想される．表 1 より，距離定義に関わらず地域，スポーツともに学習・評価クエリのジャンルが同じ場合 $(i = j)$ が最も値が高く，学習・評価クエリのジャンルが異なる場合 $(i \neq j)$ は枝刈り率が低くなる傾向が見られ，ジャンルが類似検索性能に影響を与え，ジャンルを考慮したほうが類似検索の高速化が期待できると分かる．

距離定義による違いを見てみると，地域，スポーツ共に L1 距離定義での性能が圧倒的に高く，L2 距離定義の性能はほぼ 0%であった．このように距離定義により検索性能に大きな相違があり，L1 距離定義を用いた場合に高速化が期待できる．

4.5. 考察

距離定義による結果の大幅な差について，ピボットによる距離の下限値 $D(\boldsymbol{q}_m, \boldsymbol{x}_n; P \subset \mathcal{X})$ の分布の偏りが原因であると考える．図 1 は距離の下限値の分布を示したグラフであり，横軸に距離の下限値 $(0 \leq D(\boldsymbol{q}_m, \boldsymbol{x}_n; P \subset \mathcal{X}) \leq 1)$ を，縦軸に該当数をとる．なお，図 1 の分布の面積（該当数の総和）は，距離定義，ジャンルに関わらず同一の値であることに注意されたい．青線が L1 定義，緑線が L2 定義での結果を示している．

L2 定義の下限値の分布はそのほとんどが距離 0.1〜0.3 に集中しており，今回設定したレンジ $r = 0.5$ を超える距離の下限値を有するオブジェクトペア（該当数）がほぼ存在しないため，枝刈りがほとんど出来ないことが分かる．一方，L1 定義での分布は最頻値が右側に寄っており，L2 定義よりも枝刈りが期待でき，より多くの距離計算回数が削減できると分かる．ジャンル間の性能差についても，最も性能の高いスポーツは最頻値が 0.8 付近にあり，対し，最も低い地域は最頻値が 0.6 付近にあるため分布の違いが要因の 1 つであると考える．

5. おわりに

本稿では，距離定義に着目し，クエリ分布を考慮した一般化ピボット法の検索高速化性能を評価した．L1,L2 距離定義両者とも，クエリ分布の考慮が高速化に貢献することを確認し，さらに，両者の性能差について，距離の下限値の分布の違いから L1 定義の方が L2 定義よりも高速化が見込めることが分かった．今後は他のデータにて検証を行う．

謝辞 本研究は，総務省 SCOPE(No.142306004)，及び，科学研究費補助金基盤研究 (C)(No. 26330138) の助成を受けた．

参考文献

[Bustos 03] B. Bustos, G. Navarro, and E. Chávez: “ Pivot Selection Techniques for Proximity Searching in Metric Spaces ”, Pattern Recognition Letters, Vol.24, No.14, pp. 2357-2366, (2003).

[Kimura 09] 木村 学, 斉藤 和巳, 上田 修功: “ 効率的な類似検索のためのピボット学習法 ”, 情報処理学会論文誌, Vol.50, No.8, pp.1883-1891, (2009).

[Kobayashi 14] E. Kobayashi, T. Fushimi, K. Saito and T. Ikeda: “ Similarity Search by Generating pivots based on Manhattan distance ”, PRICAI 2014, (2014).

[Kobayashi 15] 小林 えり, 斉藤 和巳, 池田 哲夫, 青山 一生, 服部 正嗣: “ クエリ分布を考慮した類似検索の高速化 ”, 第 29 回人工知能学会全国大会（JSAI2015）, (2015).