2I1-4

クエリ分布を考慮した類似検索の高速化

Similality Search Acceleration Considering Query Distribution

小林 えり[*1] 斉藤 和巳[*1] 池田 哲夫[*1] 青山 一生[*2] 服部 正嗣[*2]
Eri Kobayashi Kazumi Saito Tetsuo Ikeda Kazuo Aoyama Takashi Hattori

[*1]静岡県立大学経営情報イノベーション研究科
Graduate School of Management and Information of Innovation, University of Shizuoka

[*2]NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

We address a similarity search problem which utilizes a query distribution prepared according to user preference. We deal with this problem as a range query problem, and solve it by a pivot-based approach. In order to improve the search efficiency, we formalize the problem, using a newly introduced pivot generation scheme. In our experimental evaluation using a news article database, we confirmed that our method reasonably improved the search efficiency, and found that the level of improvement significantly varied depending on the genres of the query distributions.

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，与えられたクエリから類似したオブジェクトを検索する類似検索研究の重要性はますます高まっている．類似検索とは，クエリと類似したオブジェクトをデータベースなどの中から検出する問題を指す．オブジェクト間の類似度は距離関数から求まり距離関数は，非負性，対称性，および三角不等式の性質を満たす．データの多くは高次元で表現されるが，高次元空間に存在するオブジェクト間の距離を求めるのには大量の計算が必要となる．そのため，類似検索ではこの計算量を削減し，検索を高速化するために一部のオブジェクトをピボット集合として選定して利用する方法が提案されている．

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を提案している [Bustos 03]．このインクリメンタル法では，より良いピボット集合の指標として目的関数を定義し，目的関数を最大化するようなピボットを逐次追加することでピボット集合を得る．これに対し，オブジェクト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計算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，オブジェクト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，文献 [Kobayashi 14] において，実データを用いた実験により，目的関数，ピボット集合構築時間，レンジクエリによる類似検索性能の観点で一般化法は従来法よりも優れた性能を示すことが確認されている．

一般に，検索クエリ集合にはユーザの嗜好，トレンドなどによって何らかの偏り・分布が存在すると考えられる．例えば，ある記事データベースにおいて，株に興味のあるユーザは日経平均・東証等の株価に関する記事は勿論のこと，企業の新作発表や経営動向に関する記事もよく閲覧するため，ユーザのクエリ分布はデータベースのオブジェクト分布よりも経済色の強いオブジェクト分布になると考えられる．そこで，ユーザのクエリ分布を有する学習データを用いてピボットを構築すれば，各ユーザに対応したピボットが生成され，さらなる類似検索の高速化が期待できる．

本研究ではクエリ分布を考慮した一般化ピボット法の類似検索性能への貢献度を検証する．本稿の構成は以下となる．まず，類似検索問題について説明する．次に，マンハッタン距離に基づく一般化ピボット法について述べる．次いで，記事データを用いて，クエリ分布を考慮した際の一般化法の性能評価を報告する．最後に，本研究をまとめ今後の課題について述べる．

連絡先: 小林 えり，静岡県立大学院経営情報イノベーション学科，静岡県静岡市駿河区谷田 52-1，054-264-5436，rili0906@gmail.com

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 [Zezula 06, Samet 06]．例えば，与えられた空間に対する解の性質，即ち厳密解か近似解かの設定や問い合わせ方法，クエリから k 番目までの近似解にあるオブジェクトを検出する K-NN クエリ，またはクエリからある一定の距離以内のオブジェクトを検出するレンジクエリかの設定などがある．本稿ではクエリから一定のレンジ内にあるオブジェクトを検出する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 $\mathbf{q}_m$ （$Q = \{\mathbf{q}_1, \cdots, \mathbf{q}_M\}$） 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．この問題を解くのに要する計算時間を短縮するために，本稿ではピボット法を用いる．

ピボット法は，オブジェクト間の距離計算回数を削減し検索を高速化させるために，一部のオブジェクトを選定してピボット集合を求める．例えば，Bustos らの提案した局所最適選択法（local optimum selection）や逐次選択法（incremental selection）は，式 2 の目的関数を最大化するピボット集合 P_B^* を選択する．

$$P_B^* = \arg\max_P \mathcal{F}_B(P) \quad (1)$$

$$\mathcal{F}_B(P) = \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} D(\boldsymbol{x}_m, \boldsymbol{x}_n; P \subset X) \quad (2)$$

$$D(\boldsymbol{x}_m, \boldsymbol{x}_n; P \subset X) = \max_{1 \le k \le K} |d(\boldsymbol{x}_m - \boldsymbol{p}_k) - d(\boldsymbol{x}_n - \boldsymbol{p}_k)| \quad (3)$$

ただし，$K=|P|$, $\boldsymbol{x}_n, \boldsymbol{x}_m \in X$ である．[*1] 式 3 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界値を意味する．従って，式 3 は $\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_K$ を用いて算出した最大下界値である．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．即ち，式 3 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．レンジクエリ問題と距離の最大下界値との関係を明確にするために，最大下界値が r より大きいオブジェクト集合を $E=\{\mathbf{x}_n \mid D(\mathbf{q}_m, \mathbf{x}_n; P) > r\}$ と定義する．明らかに，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

3.1 クエリ分布を考慮した目的関数

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次のようになる．

$$P^* = \arg\max_P \mathcal{F}(P) \tag{4}$$

$$\mathcal{F}(P) = \sum_{n=1}^{N}\sum_{m=1}^{M} D(\boldsymbol{q}_m, \boldsymbol{x}_n; P \subset \mathcal{X}) \tag{5}$$

$$D(\boldsymbol{q}_m, \boldsymbol{x}_n; P \subset \mathcal{X}) = \max_{1\le k\le K} |d(\boldsymbol{q}_m - \boldsymbol{p}_k) - d(\boldsymbol{x}_n - \boldsymbol{p}_k)| \tag{6}$$

ただし，$\mathcal{X}=\{x | x \in \boldsymbol{R}^H\}$ であり，$\boldsymbol{R}^H$ は H 次元のユークリッド空間で表現されるオブジェクト空間を指す．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の中からピボットを逐次選択するのに対し，一般化ピボット計算法では，オブジェクト空間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 5 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．

文献 [Kimura 09, Kobayashi 14] より，一般化ピボット法はユークリッド，マンハッタンの 2 つの距離定義に対応しており，本稿ではマンハッタン距離定義に基づく一般化ピボット法（L1PGM 法）を採用する．

3.2 マンハッタン距離を用いた場合の目的関数及び解法

いま，任意のオブジェクト $\mathbf{x}_n$ とクエリ $\mathbf{q}_m$ のペア $\{\mathbf{x}_n, \mathbf{q}_m\}$ に対し，ピボット $\mathbf{p}_k$ で距離の下界値が最大化されるペアからなる集合を以下で定義する．

$$S_k(P) = \{\{\mathbf{x}_n, \mathbf{q}_m\} \mid k = \arg\max_{1\le k\le K} |d(\mathbf{x}_n, \mathbf{p}_k) - d(\mathbf{q}_m, \mathbf{p}_k)|\} \tag{7}$$

さらに，オブジェクト $\mathbf{x}_n$ に着目したとき，$S_k(P)$ において，$\mathbf{x}_n$ とペアで出現するクエリの集合 $S_{k,n}(P)$ を次のように定義する．

$$S_{k,n}(P) = \{\mathbf{q}_m \mid \{\mathbf{x}_n, \mathbf{q}_m\} \in S_k(P)\} \tag{8}$$

このとき，$|d(\mathbf{x}_n, \mathbf{p}_k) - d(\mathbf{q}_m, \mathbf{p}_k)|$ の絶対値を距離の大小で外すとき，距離 $d(\mathbf{x}_n, \mathbf{p}_k)$ がプラス符号で現れる回数 $c^+_{n,k}(P)$ と，マイナス符号での回数 $c^-_{n,k}(P)$ を相殺した係数 $c_{n,k}(P)$ は以下となる．

$$\begin{aligned} c_{n,k}(P) &= c^+_{n,k}(P) - c^-_{n,k}(P), \\ c^+_{n,k}(P) &= |\{\mathbf{q}_m \in S_{k,n}(P) \mid d(\mathbf{x}_n, \mathbf{p}_k) > d(\mathbf{q}_m, \mathbf{p}_k)\}|, \\ c^-_{n,k}(P) &= |\{\mathbf{q}_m \in S_{k,n}(P) \mid d(\mathbf{x}_n, \mathbf{p}_k) < d(\mathbf{q}_m, \mathbf{p}_k)\}|. \end{aligned} \tag{9}$$

また，式 8，9 に関して，$\mathbf{x}_n$ ではなく $\mathbf{q}_m$ に着目すれば，距離 $d(\mathbf{q}_m, \mathbf{p}_k)$ がプラス符号で現れる回数とマイナス符号での回数を相殺した係数 $c_{N+m,k}(P)$ も求めることができる．$\mathbf{x}_1$ に対する係数 $c_{1,k}(P)$ と $\mathbf{q}_1$ に対する係数 を区別するため，ここでは $\mathbf{q}_m$ に着目した際の係数を $c_{N,k}(P)$ より後で数えるとする．

ここで，ピボット候補集合を $\bar{P}(\bar{P}=\{\bar{\mathbf{p}}_1 \cdots \bar{\mathbf{p}}_K\})$ とし，さらにオブジェクト集合 X とクエリ集合 Q を結合した集合を $Y=\{y_1, \cdots, y_{N+M}\}$ とすると，式 9 を用いて以下の補助目的関数を定義する．

$$\begin{aligned} \mathcal{F}(P|\bar{P}) &= \sum_{k=1}^{K}(\sum_{n=1}^{N} c_{n,k}(\bar{P}) d(\mathbf{x}_n, \mathbf{p}_k) + \sum_{m=1}^{M} c_{N+m,k}(\bar{P}) d(\mathbf{q}_m, \mathbf{p}_k)) \\ &= \sum_{k=1}^{K}\sum_{l=1}^{N+M} c_{l,k}(\bar{P}) d(\mathbf{y}_l, \mathbf{p}_k) \end{aligned} \tag{10}$$

一般化ピボット法は式 10 で表される補助関数を最大化するアルゴリズムを反復することで最適なピボット集合を構築する．更新前のピボット集合 $\bar{P}$ より得られた係数 $c_{n,k}(\bar{P})$ を引数に，新たに目的関数を最大化するピボット集合を反復することで求める．更新後のピボット集合を $\hat{P}$ とすると，最大化問題は $\hat{P} = \arg\max_P F(P|\bar{P})$ を満たすピボット集合を求める問題として定式化される．よって，以下の関係が成立する．

$$\mathcal{F}(\bar{P}) = \mathcal{F}(\bar{P}|\bar{P}) \le \mathcal{F}(\hat{P}|\bar{P}) \le \mathcal{F}(\hat{P}|\hat{P}) = \mathcal{F}(\hat{P}) \tag{11}$$

式 11 より，目的関数 $\mathcal{F}(P)$ は反復によって目的関数値の向上が保証されていることが分かる．L1PGM 法のアルゴリズムについて，詳しくは文献 [Kobayashi 14] 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）が 3 回出現すれば，その形態素に対する次元の値を“ 3 ”とし，出現しなければ“ 0 ”とするようなベクトルのことをさす．つまり，類似度の高い記事同士は似たような単語で構成され，かつ単語の出現頻度も似た傾向を取ることを意味する．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，

*1 実際には，Bustos らは全てのオブジェクトペアの代わりに sampled pairs を用いて，式 2 の $F_B(P)$ を求めている．

	国内	経済	エンタメ	生活	地域	サイエンス	スポーツ	世界
国内	41.00	22.20	14.18	14.57	25.59	12.78	13.02	17.23
経済	17.76	39.82	12.89	17.04	14.81	19.10	13.19	16.15
エンタメ	12.48	14.49	27.35	10.09	10.46	11.09	9.19	10.76
生活	12.14	16.08	9.63	30.23	12.50	10.64	12.43	10.18
地域	15.93	14.23	9.45	10.16	28.46	9.73	13.67	10.13
サイエンス	14.71	30.49	16.25	18.93	13.95	42.94	13.76	11.64
スポーツ	17.62	20.06	15.79	16.36	18.96	13.55	53.72	16.56
世界	16.86	17.57	10.84	10.64	10.78	10.07	12.43	33.46
全ジャンル	38.00	36.90	27.79	29.40	29.97	38.69	43.55	30.96
最大値と無考慮の差分	3.00	2.93	-0.44	0.83	-1.51	4.25	10.17	2.50

表 1: ジャンルを考慮した枝刈り率による評価

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

本実験では 3 種のオブジェクト集合を用いて実験を行う．1 つ目はユーザのクエリ分布を考慮した学習クエリ集合 Q_i，2 つ目はユーザの今後の検索クエリ分布を考慮して生成した評価クエリ集合 S_j，3 つ目は学習クエリ集合 Q_i と評価クエリ集合 S_j を除く，検索対象である総記事データベース $\{X_{ij} \subset X\}$ $(X_{i,j} = X - Q_i - S_j)$ である．

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す．例えば Q_i，S_i はジャンル i に該当する記事のみを含む学習クエリ集合，評価クエリ集合を表す．学習クエリ集合 Q_i と同一のジャンル記事から評価クエリ集合 S_j を構成する場合，共通の記事がないように $S_i \cap Q_i = \emptyset$ となる記事を選択した．

本実験では，まず，学習クエリ集合 Q_i とデータベース $X_{i,j}$ を用いてクエリ分布を考慮したピボット集合 P_i を生成し，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，生成したピボットが類似検索性能にどのような影響を与えるのかを評価する．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．実験では検索クエリ，評価クエリ数はともに 5000 とする．

4.3 評価指標

実験では，学習クエリに関して，8 つの単独ジャンルの他に，全てのジャンルの記事をランダムに含む記事データ集合を学習クエリ集合とするパターンを加えた．よって，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，$8 \times (8+1)=72$ パターンでのレンジクエリ問題における類似検索性能の観点で評価する．

評価指標として枝刈り成功率を用いる．式 6 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/N 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．

4.4 レンジ距離の設定

レンジクエリ問題はレンジ距離 r 内に入るオブジェクトを探索する問題であり，妥当なレンジ距離を設定する必要がある．ここで，設定するレンジ距離 r に関して，KNN-Search 法を用いて設定する．図 1 は 8 ジャンルそれぞれの 5000 個の学習クエリ $\mathbf{q}$ とデータベース X 内のオブジェクトによる k 近傍オブジェクトの距離の平均値をプロットした図である．縦軸に k 近傍オブジェクトとの距離の平均値，横軸に k をとり，グラフの色がジャンルを表す．ここで，オブジェクトベクトル長は 1 に正規化されており，オブジェクトが単体上に存在するため，オブジェクトペア間の最大距離は 2.0 であることに注意されたい．

図 1 より，ランク k にもよるが，“ 国内 ”,“ スポーツ ”,“ 経済 ”,“ サイエンス ”,“ 世界 ”,“ 生活 ”,“ エンタメ ”,“ 地域 ” の順に近傍オブジェクトとの距離が大きくなっていることが分かり，“ 地域 ”,“ エンタメ ”,“ 生活 ”ジャンルは他のジャンルに比べて記事同士の距離が大きいことが予測される．また，黒い点線で示す $1.3 < r < 1.5$ の範囲において，クエリと類似する（距離の近い）オブジェクト 5〜10 個の発見が期待できると分かる．ジャンルによって近傍オブジェクトとの距離は異なるが，本実験では，統一してクエリとの距離が 1.3 以下のオブジェクトを類似オブジェクトとし，レンジ距離は $r = 1.3$ を採用する．

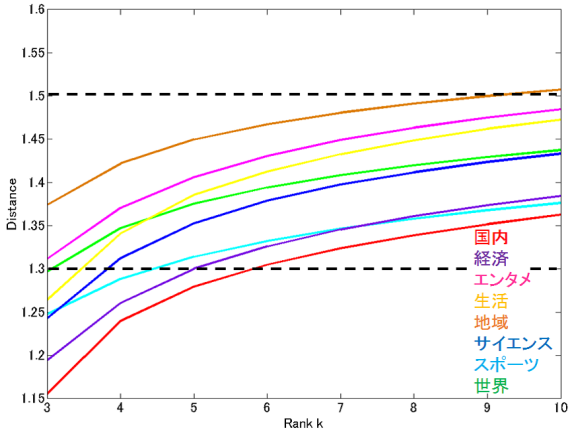

図 1: クエリに対する KNN 距離の平均値

4.5 実験結果

表 1 に 72 パターン各々のピボット数 $K = 10$ での枝刈り成功率を百分率で示す．行にはピボットの学習クエリのジャンルを，列には評価クエリのジャンルを用いる．例えば，1 行 1 列目の値はピボットの学習クエリは“ 国内 ”，評価クエリも“ 国内 ”ジャンルとした場合での枝刈り率を，1 行 2 列目の値はピボットの学習クエリは“ 国内 ”，評価クエリは“ 経済 ”ジャン

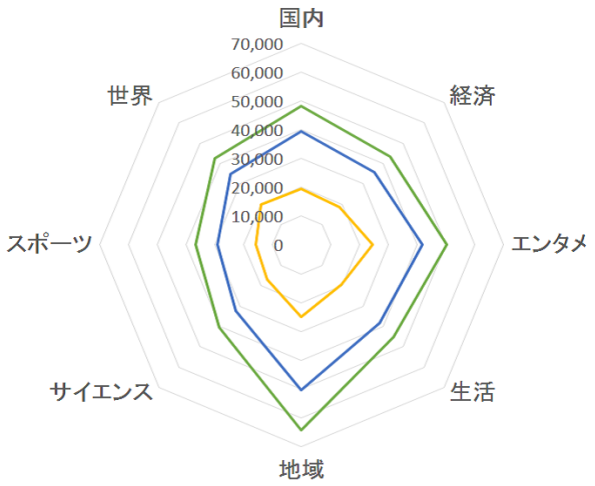

図 2: 各ジャンルの出現単語種類数

ルに限定した場合での枝刈り率を示す．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，全ジャンルを除く，各行のどのジャンルでも学習・評価クエリのジャンルが同じときが最も値が高いことが分かる．最高で 50%，最低で 27%の成功率を示しており，少なくとも 4 分の 1 は枝刈りが期待できる．また，学習・評価クエリのジャンルが異なる場合を見てみると，各行の最高値と比較して大半が 10〜30%近く枝刈り率が低く，ジャンルが類似検索性能に影響を与えていることが確認できる．

学習クエリのジャンル（ピボットのジャンル）を限定せず，全てのジャンルを含む学習クエリで生成したピボット（以後全ジャンルピボットと呼ぶ）での評価を見てみると，“ エンタメ ”,“ 地域 ”ジャンル以外は全て学習と評価クエリのジャンルが同じ場合が最も枝刈り率が高いことが確認でき，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．

4.6 考察

“ エンタメ ”,“ 地域 ”ジャンルに関し，全ジャンルピボットを用いた場合よりも枝刈り成功率が低下した原因を考察する．単語頻度ベクトルの分布を考えた場合，似た単語構成をしたベクトル同士が近傍に位置する．ジャンルにはそのジャンル特有の単語構成があり，そのジャンルでよく出現する単語，つまり特徴単語（特徴次元）があると考えられる．よって，クエリ分布を学習データに与えることでジャンルごとの特徴次元を学習し，各ジャンルに対応したピボットが生成できるようになる．また，特徴次元により，高次元空間上ではあるジャンルがよく見られる分布領域（以下，ジャンル領域とする）が存在すると考えられる．例えば“ 国内 ”ジャンルの記事は第 1〜100 次元まで値が非常に大きく，その他の次元での値がほぼ 0 に近い場合，“ 国内 ”のジャンル領域は第 1〜100 次元空間となる．また，生成されたピボットは対応するジャンル領域付近に位置するようなベクトルとなり，自身と異なるジャンルの記事を効率的に枝刈りできるようになると期待される．

図 2 は各ジャンルの出現単語種類数をレーダーチャートで示した結果であり，外側から 1 以上，2 以上，10 以上の記事に出現する各ジャンルごとの単語の種類数を示す．図 2 より，“ 地域 ”,“ エンタメ ”ジャンルは全てのパターンで出現単語種類数が最も多いことがわかる．出現単語数が多いということは記事間の距離が開きにくくなり，かつ，自身のジャンル領域が他のジャンルの領域の一部を含んでしまうことにつながる．つまり,“ 地域 ”,“ エンタメ ”のジャンル領域には他のジャンルも含まれ，ピボットがジャンルの特有の特徴を学習し辛かったことが性能低下の原因だと考える．

クエリ分布を考慮したピボットは，そのクエリ分布が明確な特徴を有する場合に効果的に作用する．例えばユーザの検索履歴をもとにしたクエリ分布を用いる場合，多様なジャンルの記事を検索するユーザより，ある一つのジャンルや特定分野の記事を検索するユーザのほうが，学習により高性能なピボットを得ることができ，検索の高速化が見込める．

5. おわりに

本研究では，検索クエリにはユーザの嗜好，トレンドなどによって何らかの偏り・分布が存在すると仮定し，過去の検索情報を学習データとしてピボットを生成する，クエリ分布に基づく一般化ピボット法による類似検索の高速化の評価を検証した．クエリ分布を学習データとして生成したピボットは各ユーザのクエリ分布に対応するため，今後のユーザの検索クエリに対して効果的な類似検索の高速化が期待できると考えられる．記事のジャンルをクエリ分布と設定しピボットを生成し，その性能評価したところ，ジャンルが類似検索性能に影響を与えていることが確認でき，検索の高速化を期待できると確認できた．また，生成されたピボットは与えたクエリ分布が特徴的であるほど高い類似検索性能が期待できることが分かった．今後は異なるデータでの検証，実際のユーザの閲覧履歴などのジャンル以外をクエリ分布とした際の評価を行う．

謝辞 本研究は，総務省 SCOPE(No.142306004)，平成 26 年度ふじのくに地域・大学コンソーシアム学術研究，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 Lettes, 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.8，1883-1891(2009).

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

[Zezula 06] P. Zezula, G. Amato, V. Dohnal, and M. Batko: “ Similarity search: The metric space approach ”, Springer, (2006).

[Samet 06] H. Samet: “ Foundations of multidimensional and metric data structure ”, Morgan Kaufman, (2006).