DBSJ Journal, Vol.11, No.1
June 2012

例示検索のための集約点に基づくドメイン適応

Domain Adaptation based on Relative Aggregation Points for Content-based Search

加藤 誠† 大島 裕明‡
田中 克己‡

Makoto P. KATO Hiroaki OHSHIMA
Katsumi TANAKA

本論文では例示検索におけるドメイン適応の必要性を実験的に示し，各ドメインにおける相対的集約点(RAP)を対応させることによりドメイン適応を行う方法を提案する．例示検索において，例を選択するドメインと検索対象となるドメインが著しく異なるとき，ある検索意図の下に選択された例とその意図に適合するデータが表面的には類似しない場合がある．我々はドメインの差異を考慮した例示検索を可能にするため，RAP に基づくドメイン適応を提案する．RAP はあるクラスに属するデータの期待値であり，RAP が異なるドメインにおいて対応していると仮定し，RAP に基づいて新たな特徴空間を構成することによりドメイン適応を行う．我々は飲食店例示検索のためのテストコレクションによって例示検索におけるドメイン適応の必要性を論じ，既存手法との比較によって RAP に基づくドメイン適応の有効性を示した．

We propose a domain adaptation method for content-based retrieval, based on *relative aggregation points* (RAPs). In content-based retrieval, input examples and relevant data are not always similar, particularly when domains where the user selects examples and where the system retrieves data are heterogeneous. In order to bridge the domain gap, a new feature space is constructed based on RAPs that are estimated in each domain. We conducted a test-collection-based experiment to verify the effectiveness of the RAP-based method.

1. はじめに

例示検索は画像や音声，地理情報など検索対象となるデータがキーワードで表現しづらい場合に有効である．たとえば我々が提案する地理情報例示検索[5]において，ユーザは良く知っている地域（ソースドメイン$\mathcal{D}^{(S)}$と呼ぶ）で例を選択することで知らない地域（ターゲットドメイン$\mathcal{D}^{(T)}$と呼ぶ）の地理情報を検索することができる．例示は検索対象地域についてまったく知識がない場合でも可能であり，たとえば「比較的安いその土地独自の料理が食べられる店」を値段やカテゴリを指定することなく，良く知っている地域での例示によって検索することができる．しかし，ソースドメインとターゲットドメインが著しく異なる場合，選択された例と正解データが表面的には類似しない場合がある．たとえば，京都の学生街で比較的安いその土地独自の料理が食べられる店を選んで神戸の飲食店を検索したとする．京都で選ばれた店が「おばんざい」を出す2,500円程度の和食の飲食店だった場合，神戸で「おばんざい」を出す2,500円程度の店は与えられた例との類似度は高いがユーザの意図には適合しない．神戸では「神戸牛」を出す3,000円程度の洋食店がこのユーザの意図に一致すると考えられるが，京都で選ばれた例とは値段帯もカテゴリも異なる．これは両ドメイン（京都と神戸）において平均的な飲食店の値段や独自の料理が異なるためである．この問題はドメイン適応[8]の問題であると考えられるが，例示検索の場合についてはこれまで研究されていない．そのため，本論文の2節ではドメイン適応についての関連研究を述べ，3節にて例示検索のためのドメイン適応を定義し既存の問題との相違点について議論する．

本論文では著しく異なるドメインにおいて例示検索を行う方法を提案する．提案手法では異なるドメインにおいて表現方法は異なるが同じ意味を持つ，特徴空間上の点を対応させることでドメイン間の差異を是正する．たとえば，京都と神戸には最も高い店，平均的な店，もっとも多いカテゴリの店などが存在し，それらはドメインによって異なるが同じ役割を担っていると考えることができる．我々はこれらの点を**相対的集約点**（relative aggregation point, *RAP*）と呼び，各データをこの点との類似度によって特徴付けることにより，ドメインの差によって引き起こされる例示データと適合データの齟齬を是正する．

実験は2つの疑問—(i) 例示検索にドメイン適応は必要であるか，また，(ii) RAPによるドメイン適応は既存手法を有意に上回ることができるか—に答えるように設計された．1つ目の疑問を言い換えれば，ある検索意図に対する適合データは検索対象データもしくは検索地域によって異なるか，とも言える．我々は100クエリ（20検索意図 × 5地域）を含む地理情報例示検索のためのテストコレクションを作成し，異なる地域における例示検索のnormalized discounted cumulative gain (nDCG)が同一地域のそれよりも著しく劣ることを示し，第1の疑問に対して解を与えた．第2の疑問に答えるために，様々なベースライン手法と我々の提案手法を比較し，我々の提案手法が有意に検索性能を改善させることを示した．

2. 関連研究

ドメイン適応[8]は品詞タグ付け[3]や文書分類[2,4,6,9]など様々なタスクにおいて取り組まれてきた問題である．これらのタスクは共通して，あるソースドメイン$\mathcal{D}^{(S)}$で学習したモデルは異なるドメイン（ターゲットドメイン$\mathcal{D}^{(T)}$）では有効に働かないことを問題視している．Blitzerらはstructual correspondence learning (SCL)という手法をドメイン適応のために提案している[3]．この手法では，ソースドメインとターゲットドメインにおいて同様の意味を持つ語を発見し，その語との相関に基づき両ドメインにおいて表面的に異なるが同じ意味を持つ語を推測して両ドメイン間の差異を是

† 学生会員 京都大学大学院情報学研究科・日本学術振興会特別研究員 kato@dl.kuis.kyoto-u.ac.jp
‡ 正会員 京都大学大学院情報学研究科
{ohshima, tanaka}@dl.kuis.kyoto-u.ac.jp

正している．SCLは品詞タグ付け，および，文書分類[2]に用いられており，ドメイン適応における最先端の手法の1つであると考えられる．文書分類に対するドメイン適応は数多く研究されており，共クラスタリングに基づくドメイン適応手法[4]，スペクトラルクラスタリングに基づくドメイン適応[6]，pLSAによるドメイン適応[9]などがある．

NakajimaとTanakaは相対的なクエリの処理方法について提案を行っている[7]．相対的なクエリはある集合Xとその要素xで構成され，ベクトル空間においてはXの重心からxへのベクトルとして表される．彼らの手法との違いは下記の2点である: (i) 例示検索がランク付けされたデータを返すのに対し，相対的クエリ処理では1つのデータを返すことを目的としている．(ii) 例示検索のためのドメイン適応では，ある特徴が別のドメインでは異なる意味を持ちうることを想定するのに対し，相対的クエリ処理では，ある特徴はどのドメインでも同じ意味を持つがその特徴量が相対的である場合を想定している．さらに我々は彼らの手法をベースライン手法の1つとすることで，性能面においても提案手法との差別化を行っている．

我々の手法もこれまで提案されてきたようにソースドメインとターゲットドメイン間で同じ意味を持つものの存在を仮定している．しかし，既存手法はある特徴が共通することを仮定する一方で，我々はどの特徴も同じ意味を持つとは限らないと考えている．そのため，我々は共通する特徴を用いてドメイン適応を行うのでなく，対応する役割（最大，平均, 最小など）を持つ点を用いてドメイン適応を行っている．また，我々が用いるテストコレクションはこれまで利用されてきたものとは大きく異なり，値段などの連続値やカテゴリなどの離散値，テキストなどの高次元特徴を持つ．加えて，分類問題とは異なり，ラベルの分布が大きく偏っているため，既存の提案手法を適応しても良い結果が得られないと考えられる．このことについては実験において詳しく議論する．

3. 例示検索のためのドメイン適応

本節では例示検索のためのドメイン適応の定義を与え，ドメインの違いについても論じる．ドメインは定義域$\mathcal{X}$と周辺確率$P(X)$)によって特徴付けられ，まとめて$\mathcal{D} = (\mathcal{X}, P(X))$と表記される[8]．また，ラベル集合$Y$と，ドメイン$\mathcal{D}$から$P(X)$にしたがって得られたインスタンス集合$X \subset \mathcal{X}$とのペアをドメインデータと呼び，$D = \{(x_1, y_1), (x_2, y_2), \ldots\}$で表す．

例示検索において，ソースドメイン$\mathcal{D}^{(S)}$から得られたインスタンス集合$X^{(S)}$を用いてユーザは例示を行い，クエリはインスタンス集合$X^{(S)}$の部分集合$Q \subset X^{(S)}$，もしくは，$D^{(S)} = \{(x^{(S)}, +1) \mid x^{(S)} \in Q\} \cup \{((x^{(S)}, -1) \mid x^{(S)} \in (X^{(S)} - Q)\}$で表現することができる．

検索対象データはターゲットドメイン$\mathcal{D}^{(T)}$から得られたインスタンス集合$X^{(T)}$であり，システムはランク付けされたインスタンス集合$X^{(T)}$を出力する．ランクはクエリ$D^{(S)}$およびターゲットドメインのインスタンス集合$X^{(T)}$から推定されるランク関数$f: X^{(T)} \to \mathbb{R}$によって与えられる．すなわち，順序$\prec$が$X^{(T)}$上に$x_i^{(T)} \prec x_j^{(T)} \Leftrightarrow f\left(x_i^{(T)}\right) < f\left(x_j^{(T)}\right)$で定義される．また，得られる順序集合$\left(X^{(T)}, \prec\right)$はある検索指標によって評価されるものとする．このとき，例示検索のためのドメイン適応は，この検索指標の値を最大化するランク関数fを推定する問題であると定義できる．

たとえば地理情報例示検索の場合，ユーザはまず京都をソースドメイン，神戸をターゲットドメインとして選択する．このとき，ドメインの定義域$\mathcal{X}^{(S)}$と$\mathcal{X}^{(T)}$は等しく，これは実現可能なすべての飲食店集合である．京都の飲食店集合$X^{(S)} = \left\{x_1^{(S)}, x_2^{(S)}, x_3^{(S)}\right\}$は周辺確率$P\left(X^{(S)}\right)$にしたがってサンプリングされたものであると仮定される．ユーザは$X^{(S)}$からその部分集合$Q = \left\{x_1^{(S)}, x_2^{(S)}\right\}$を選択し，これは$D^{(S)} = \left\{\left(x_1^{(S)}, +1\right), \left(x_2^{(S)}, +1\right), \left(x_3^{(S)}, -1\right)\right\}$と表すこともできる．地理情報例示検索システムは$D^{(S)}$をクエリとして受け取り，ターゲットドメイン$\mathcal{D}^{(T)}$から得られた神戸の飲食店集合$X^{(T)} = \left\{x_1^{(T)}, x_2^{(T)}, x_3^{(T)}\right\}$をランク付けする．

既存のドメイン適応タスクとの大きな違いは，例示検索ではターゲットドメインのラベルをあらかじめ用意することが非常に困難である点である．適合データはユーザの検索意図に依存するため，すべての検索意図に対してあらかじめ正解（ラベル）を用意するのは現実的ではない．注意すべき点として，アドホック検索ではすべての検索対象データ（ターゲットドメインのインスタンス）を利用することはデータ量の問題から一般的に不可能である．そのかわり，たとえば検索対象データの一部を利用するか，近傍検索を行った後に上位のインスタンスを利用してリランキングすることが考えられる．ただし，地理情報例示検索においては，ユーザが検索対象とする地域に含まれるインスタンスはたかだか 1,000 件程度であり，すべてのターゲットドメインのインスタンスを利用してランク付けを行うことができる．

ドメインの違いは定義域の違いおよび周辺確率の違いによって特徴付けられる．地理情報検索の例では，どのドメインにおいても定義域は等しく$(\mathcal{X}^{(S)} = \mathcal{X}^{(T)})$，あるインスタンスは予算，カテゴリ，レストランに関する記述で表現される．一方で周辺確率はドメインによって異なり$(P(X^{(S)}) \neq P(X^{(T)}))$，たとえば銀座には高い店が多く，京都には和食の店が多いことが反映される．我々はドメイン間の違いを$\mathcal{A}$距離によって定量化する[1]．$\mathcal{A}$距離はドメイン間の違いを計る尺度として用いられ，主にドメイン適応で改善できる性能の上限値を推定するために用いられる．後に，我々が用いるすべてのドメインは互いに十分異なっていることを示し，これを不等式$\mathcal{D}^{(S)} \neq \mathcal{D}^{(T)}$によって表現する．

4. 相対的集約点に基づくドメイン適応

この節では RAP を用いたドメイン適応手法について述べる．この手法ではソースドメインとターゲットドメインにおいて同様の意味を持つインスタンスを発見し，それらとの類似度によってインスタンスのベクトル表現を拡張しドメイン適応を行う．我々は，あるクラスに所属するインスタンスの期待値である RAP が，そのような同じ意味を持つインスタンスであると仮定する．たとえば，飲食店の例では，平均的な店，最も高い店，ある地域でのみ頻出するカテゴリに属する店が RAP の例である．

4.1 相対的集約点に基づく特徴表現

RAP の厳密な定義を与えるために，ある地域で最も高い店を考える．あるドメイン$\mathcal{D}$からサンプリングによって得られたインスタンス集合Xが与えられた時に，最も高い店は容

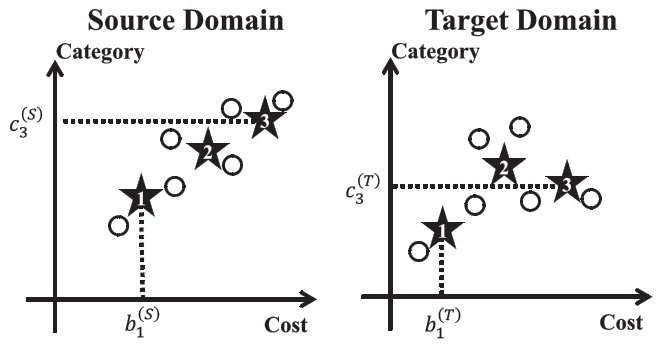

図 1　RAP とそれに基づく特徴表現．黒い星は RAP であり，白い丸はインスタンスを表している．同じ数字を持つ RAP は対応している．

Fig. 1　RAPs and construction of a new feature representation. Black stars are RAPs, and white circles are instances. Stars that have the same number are corresponding.

易に特定でき，それらは$\psi_{\mathrm{MAX}}(X) = \mathrm{argmax}_{x \in X} x_b$と表現することができる．ただし，$x_b$はインスタンス$x$の予算を表す．しかし，インスタンス集合$X$がそのドメインの典型的なインスタンス集合であるとは限らない．たとえば，ドメインでの平均予算が低いにもかかわらずXが異常に高級な店を含んでしまい，その地域で最も高い店に大きく左右される可能性がある．そのため，我々はあるドメインから得られるすべてのインスタンス集合における最も高い店の期待値を推定する．

あるドメイン$\mathcal{D}$と部分集合関数$\psi: 2^{\mathcal{X}} \to 2^{\mathcal{X}}$（$2^{\mathcal{X}}$は$\mathcal{X}$のすべての部分集合)が与えられた時，RAP は下記のように定義される：

$$a_\psi = \int_{x \in \mathcal{X}} \mathbf{x} P(x|\psi) dx$$

ただし，$\mathbf{x}$はインスタンスxを表すベクトルであり，$P(x \mid \psi)$はドメイン$\mathcal{D}$から得られる任意のインスタンス集合Xの部分集合$\psi(X)$から，インスタンスxが得られる確率を表している．周辺化を行うことにより$P(x \mid \psi)$は下記のように得られる．

$$P(x \mid \psi) = \int_{X \in 2^{\mathcal{X}}} P(x \mid \psi, X) P(X) dX$$

ただし，$P(x \mid \psi, X)$はインスタンス集合Xの部分集合$\psi(X)$から，インスタンスxが得られる確率を表している．直感的には，部分集合関数ψはインスタンスの決定的な所属度合いを表しており，一方，確率$P(x \mid \psi)$は確率的な所属度合いを表している．

図 1 では RAP を黒い星で示している．ただし，特徴空間は予算とカテゴリ(たとえば和食らしさ)の 2 つの次元で構成されているものとする．黒い星 1 は最も安い店，2 は平均的な予算の店，3 は最も高い店を表している．我々はこれらの RAP がソースドメインとターゲットドメインで対応し，またそれらの特徴も対応することを仮定する．図 1 の例では，最も安い店の予算が対応し$(b_1^{(S)} = b_1^{(T)})$，最も高い店のカテゴリの値もまた対応していると考えられる$(c_3^{(S)} = c_3^{(T)})$．京都の学生街と神戸の例では，両ドメインで最も高い店は予算もそのカテゴリも異なると考えられる．たとえば，京都の学生街では最も高い店が京料理を出す 10,000 円程度の店，神戸では神戸牛を出す 15,000 円程度の店であるとする．両ドメインで最も高い店を利用することで，我々は 10,000 円と 15,000 円，京料理と神戸牛が同じ意味を持つものであると考える．上記の仮定により，我々は各インスタンスを RAP との類似度によって表現することができる．すなわち，異なるドメインのインスタンスは RAP との類似の仕方が似ていれば類似したものであると考えることができる．

RAP による具体的な特徴表現を下記にて与える．$H = \{h_1, h_2, \ldots, h_m\}$を類似度関数の集合，$A = \{a_1, a_2, \ldots, a_n\}$を RAP の集合とする．たとえば，$H$には値段の類似度，カテゴリの類似度などが含まれる．我々は関数$\phi: \mathcal{X} \to \mathbb{R}^{mn}$を用いて，あるインスタンス$x$を$mn$次元のベクトル$\phi(x)$で表現でき，その定義は下記のようになる：

$$\phi(x) = \big(\phi_1(x), \phi_2(x), \ldots, \phi_n(x)\big)$$

ただし，$\phi_i(x) = \big(h_1(x, a_i), h_2(x, a_i), \ldots, h_m(x, a_i)\big)$であり，各次元はある類似度関数によって測られたある RAP とインスタンスxの類似度に対応している．インスタンスxのベクトル表現$\mathbf{x}$と RAP による特徴表現$\phi(x)$を組み合わせることで，拡張されたベクトル表現$\big(\mathbf{x}, \lambda\, \phi(x)\big)$が得られる．ただし，$\lambda$は RAP による特徴表現の重要度を表すパラメータであり，実際にはベクトル$\mathbf{x}$と$\phi(x)$はそれらのL^2ノルムによって正規化されているものとする．

RAP はソースドメインとターゲットドメインで別々に求められる．関数$\phi^{(S)}$をソースドメインで推定された RAP に基づく関数，$\phi^{(T)}$をターゲットドメインの関数であるとしたとき，ソースドメインのインスタンス$x^{(S)} \in X^{(S)}$は$\Big(\mathbf{x}^{(S)}, \lambda\, \phi^{(S)}\big(x^{(S)}\big)\Big)$に変換され，ターゲットドメインのインスタンス$x^{(T)} \in X^{(T)}$は$\Big(\mathbf{x}^{(T)}, \lambda\, \phi^{(T)}\big(x^{(T)}\big)\Big)$に変換される．したがって，我々の手法は直接的な類似度(すなわち，$\mathbf{x}^{(S)}$と$\mathbf{x}^{(T)}$の類似度)と RAP による相対的な類似度(すなわち，$\phi^{(S)}\big(x^{(S)}\big)$と$\phi^{(T)}\big(x^{(T)}\big)$の類似度)の両方を同時に考慮することが出来る．

たとえば，「比較的安いその土地独自の料理が食べられる店」は下記のように得られる．比較的安い店はあるドメインにおける最も安い店に予算の面で類似し，その土地独自の料理が食べられる店はそのドメインで相対的に多く出現するカテゴリの店とカテゴリの面で類似するはずである．また，最も高い店や平均的な店とのカテゴリ類似度もその土地独自の料理を特定するために重要な特徴になりうる．したがって，我々はこれらの RAP との類似の仕方がどれくらい似ているかを考えることで，「比較的安いその土地独自の料理が食べられる店」を異なるドメインで発見することができる．

4.2 相対的集約点の例

我々の提案手法はフレームワークであり，効果的な RAP はタスクに大きく依存する．ここでは地理情報例示検索に有効な RAP の例を与える．多くの RAP は期待値として表現され，これらはサンプリングによって得られるが，詳細な方法については本質的でなく紙面の制約のため割愛する [1]．

AVG

最も平均的な店は異なるドメインであっても対応するかもしれない．Average (AVG) RAP はあるドメインのインスタンスの期待値で与えられる．部分集合関数は$\psi_{\mathrm{AVG}} = \{x \mid x \in X\}$と定義され，このとき，確率$P(\mathrm{x} \mid \psi_{\mathrm{AVG}})$は周辺確

[1] たとえば，単純な期待値はサンプルされたインスタンス集合Xにより下記のように近似できる：$\int_{x \in \mathcal{X}} \mathbf{x} P(x)\, dx \approx 1/|X| \sum_{x \in X} \mathbf{x}$.

率$P(x)$と等しくなる．したがって，AVG RAP は$a_{\psi_{\mathrm{AVG}}} = \int_{x\in X} \mathbf{x}P(x)dx$で求められる．

MAX/MIN

先述したように最も高い店(MAX)(と最も安い店(MIN))は異なるドメインでも対応しうる．MAX RAP は予算最大の店の期待値として与えられる．部分集合関数は$\psi_{\mathrm{MAX}} = \mathrm{argmax}_{x\in X} x_b$と定義され，確率$P(\mathrm{x} \mid \psi_{\mathrm{MAX}})$は$P(\mathrm{x} \mid \psi_{\mathrm{MAX}}) = \int_{b\in\mathbb{R}} P(x \mid b)P(b|\psi_{\mathrm{MAX}})db$によって与えられる．ただし，$P(b|\psi_{\mathrm{MAX}})$は任意のインスタンス集合における最大予算の確率であり，$P(x|b)$は予算がbであるようなインスタンス集合からインスタンスxが得られる確率である．したがって，MAX RAP は下記のように定義される：$a_{\psi_{\mathrm{MAX}}} = \int_{x\in X} \mathbf{x} \int_{b\in\mathbb{R}} P(x \mid b)P(b \mid \psi_{\mathrm{MAX}})dbdx$.

MIN RAP も同じようにして得られるが，部分集合関数は$\psi_{\mathrm{MIN}} = \mathrm{argmin}_{x\in X} x_b$と定義される．

FREQ

相対的に頻出する（FREQ）カテゴリは異なるドメインにおいて対応するかもしれない．たとえば，京都の和食料理店と神戸の洋食料理店はその土地で人気のあるカテゴリという意味で類似する．部分集合関数は$\psi_{\mathrm{FREQ}} = \{x \mid x \in X \wedge (\forall c \in x_c)\mathrm{freq}_X(c) > \mathrm{freq}_A(c)\}$で定義される．ただし，$x_c$はインスタンス$x$のカテゴリ集合，$\mathrm{freq}_X(c)$はあるインスタンス集合$X$でのカテゴリ$c$の頻度，そして，$A$はドメイン全体からサンプリングされたインスタンス集合である．この部分集合関数はXにおいてAよりも多く出現するカテゴリをもったインスタンス集合を返す．MAX/MIN RAP と同様に FREQ RAP は下記のように定義される：$a_{\psi_{\mathrm{FREQ}}} = \int_{x\in X} \mathbf{x} \int_{C\in\mathcal{C}} P(x \mid C)P\left(C \mid \psi_{\mathrm{FREQ}}\right)dCdx$．ただし，$\mathcal{C}$は全カテゴリ集合の部分集合全体である．

CLS

RAP はクラスタリングによっても求められる．部分集合関数は$\psi_{\mathrm{CLS}} = \{x \mid x \in X \wedge x \in S\}$と定義され，インスタンス集合$S$はソースドメインとターゲットドメインのインスタンスをまとめてクラスタリングすることで得られる．確率$P(x \mid \psi_{\mathrm{CLS}})$はインスタンスのサンプリングとは独立であるため，CLS RAP は$a_{\psi_{\mathrm{CLS}}} = \int_{x\in X} \mathbf{x}P(x \mid S)dx$と簡潔に定義できる．

5. 実験

本節では実験を通して下記の 2 つの疑問に答える：(i)例示検索にドメイン適応は必要であるか，また，(ii) RAP によるドメイン適応は既存手法を有意に上回ることができるか．

5.1 テストコレクション

我々は地理情報例示検索で用いるデータとして飲食店情報を選択し，グルメ情報検索サイト「ぐるなび」[2] から得られた合計 46,945 件の飲食店情報をインスタンスとした．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，店舗名，カテゴリ，ジャンル，紹介文，予算を用いて各飲食店を特徴付けた．名前とジャンル，紹介文の 3 つの属性値はテキストで構成されており，これらは形態素解析 [3] を行い名詞および形容詞のみを用いた．カテゴリは複数の値から成り，たとえば，{和食，割烹}と表現される．そのため，これらも形態素解析で得られた語と同じように扱った．テキストとカテゴリはそれぞれ別々に処理され，tf-idf によって重み付けされた．

ドメインは日本の主要都市から 5 都市を選択し，その中でも特に，京都祇園(Kyoto)，東京銀座(Tokyo)，札幌駅周辺(Sapporo)，博多駅周辺(Fukuoka)，名古屋駅周辺(Nagoya)を選択した．またその中でも飲食店が 500 件以上含まれる範囲を選び，これらをドメインのインスタンス集合とした．

例示検索において言語化された検索意図は必要ないが，検索意図を表す文章はテストコレクション作成時に必要になる．そのため，我々は Yahoo!知恵袋 [4] の「飲食店」カテゴリから，検索意図として利用できる質問を手動で収集した．まず 100 件の飲食店に関する質問を検索意図として抽出し，これらをドメイン依存性という観点から 2 つのクラスに分類した．ドメイン依存性とは，ある検索意図に対する適合データが地域，すなわち，ドメインによって変わる可能性があるかを主観的に判断したものである．我々は 2 名の被験者に 100 件の検索意図のドメイン依存性を評価してもらった．

たとえば，ドメイン依存であるような検索意図には「今週，上司と男二人で<x>-<y>に出張に行きます．夕食のおすすめはどこでしょうか？金沢からで，場所は<y>，予算は一人 3000－5000 円ほど． <x>$らしいものを食べながら，お酒を飲みたいと思っています．ちなみに年齢は，30 歳と 40 歳，肉でも魚でもＯＫです．」があり，非依存であるような検索意図には「<x>-<y>で洋食の店を紹介してください．<y>に用があるのでその近郊の洋食，具体的にはイタリアン or フレンチのお店を教えて頂きたく思っております．予算はランチコースでワイン抜き 5000 円以内です．」がある．ただし，検索意図中の<x>および<y>は各ドメインごとにそのドメインの地名に置き換えられる．たとえば，京都祇園であれば<x>は「京都」に<y>は「祇園」に置き換えられる．

我々はオンラインアンケートによって 5 地域在住の 1,000 人の被験者を募集し，居住地と対応するドメインにおいて検索意図に適合する飲食店を選択してもらった．なお，被験者の性別，年齢が等しく分布するように募集し，男女と 20，30，40，50，60 代の組み合わせが各地域において 20 人になるように調整してある．結果として，適合であると判断された回数に基づき各インスタンスに対して段階評価を得た．

5.2 実験設定

我々はテストコレクション中のあるドメインをソースドメイン$\mathcal{D}^{(S)}$，別のドメインをターゲットドメイン$\mathcal{D}^{(T)}$として利用した．3 節で述べたように入力は二値のラベル（選択されたインスタンス：+1，選択されなかったインスタンス：−1）とインスタンスのペア集合$D^{(S)}$であり入力では段階評価は無視される．例示検索手法は$D^{(S)}$とターゲットドメインのインスタンス集合$X^{(T)}$を受け取り，$X^{(T)}$をランク付けするために最適なランク関数fを推定する．評価には nDCG を利用した．また一般的な検索エンジンは検索結果ページ中に 10 件の結果を含むため，nDCG の閾値は 10 とした．（nDCG@10 と表記する．）

ターゲットドメインの知識を利用しない単純なベースラ

[2] http://www.gnavi.co.jp/
[3] MeCab を利用した． http://mecab.sourceforge.net/
[4] http://chiebukuro.yahoo.co.jp/

表 1　各ドメイン間で SVM を用いて例示検索を行ったときの平均 nDCG@10. 太字は in-domain 設定を表している.
Table 1　The average nDCG@10 over all the search intents in each combination of the source and target domains.　The bold font indicates in-domain settings, while the others are out-domain settings.

		Target				
		Kyoto	Tokyo	Sapporo	Fukuoka	Nagoya
Source	Kyoto	**0.620**	0.463	0.520	0.511	0.431
	Tokyo	0.460	**0.596**	0.449	0.501	0.389
	Sapporo	0.492	0.432	**0.638**	0.524	0.506
	Fukuoka	0.505	0.472	0.510	**0.588**	0.477
	Nagoya	0.476	0.466	0.538	0.527	**0.599**

イン手法として，我々は最近傍検索(NNS)と 1 クラス SVM(OSVM)，SVM を選択した．NNS は$D^{(S)}$中の選択されたインスタンスの重心をクエリとし類似度の高い順にターゲットインスタンスをランク付けする．OSVM は$D^{(S)}$中の選択された例だけを用いて学習し識別関数をランク関数に利用した．NNS と OSVM は選択されなかった例を利用しておらず，従来の例示検索に最も近い手法であると考えられる．SVM では$D^{(S)}$を使って二値分類器を学習し OSVM 同様に識別関数をランク関数とした．

ドメイン適応を行うベースラインは transductive support vector machine (TSVM) と structural correspondence learning (SCL)，relative cluster mapping (RCM)である．Joachims によって提案された TSVM はドメイン適応のベースラインとして広く用いられている[6,9]．ランク関数には OSVM や SVM 同様に識別関数を利用した．最先端手法の 1 つであると考えられる SCL は，ピボットというソースおよびターゲットドメインで同じ意味を持つ特徴を利用することで，各インスタンスのベクトル表現を拡張している[3]．拡張した後は SVM を適用しランク関数を学習した．RCM はインスタンスをそのインスタンスが所属する集合の重心からそのインスタンスへのベクトルで表現し，そのベクトルに対して SCL と同様に SVM を適用する[7]．

我々の提案手法である RAP に基づくドメイン適応はパラメータλとどの RAP を使うかという選択肢がある．実験では$\lambda = 1$とし 4 節で述べたすべての RAP を利用して他の手法と比較を行った．CLS RAP は CLUTO[5]によって，両ドメインのインスタンスをクラスタリングすることで得られた．パラメータは規定値を用いクラスタ数は 30 とした．他のベースライン手法と同様にベクトル拡張を行った後は SVM を適用した．

5.3 例示検索におけるドメイン適応の必要性

ドメイン適応をせずに異ドメイン間で例示検索を行った場合，同ドメイン間の結果に比べ大きく劣ることを示す．我々はこの実験のために，ドメイン適応をしない手法 SVM を利用した．ソースドメインとターゲットドメインが同一であるとき，これを in-domain 設定($\mathcal{D}^{(S)} = \mathcal{D}^{(T)}$)と呼ぶ．このときには$k$交差検定と似たような方法をとり，あるドメインのインスタンス集合はk個に分割され，$k-1$個がソースに残りがターゲットに使われた．ソースドメインとターゲットドメインが異なるときは out-domain 設定($\mathcal{D}^{(S)} \neq \mathcal{D}^{(T)}$)と呼び，in-domain 設定と同じようにソースドメインとターゲットドメインのインスタンス集合をk個に分割し，ソースドメインの$k-1$個とターゲットドメインの1個で性能を評価した．本実験では$k = 5$としている．クエリはソースドメインと検索意図の組み合わせによって異なり 100 種類（5 ドメイン × 20 検索意図）ある．ターゲットドメインも 5 種類考えられるため，我々は合計 500 種類の検索を評価した．

表 1 に一方のドメインをソースドメインに他方をターゲットドメインにし全検索意図で平均を取った nDCG@10 を示す．SVM を用いて例示検索を行ったとき，すなわち，ドメイン適応を行わなかったとき，明らかに out-domain 設定($\mathcal{D}^{(S)} \neq \mathcal{D}^{(T)}$)の平均 nDCG@10 は in-domain 設定($\mathcal{D}^{(S)} = \mathcal{D}^{(T)}$)に比べ大きく劣っていることがわかる．また我々は平均 nDCG@10 と$\mathcal{A}$距離の間に負の相関を発見した(Pearson's coefficient $r = -0.678, p < 0.05$)．このことはドメインが異なれば異なるほど，検索精度が低下することを示唆している．両設定の平均 nDCG@10 は in-domain 設定で 0.608，out-domain 設定で 0.482 であった．ウェルチの t 検定を行ったところ，in-domain 設定と out-domain 設定には有意差が存在した($t(179) = 5.89, p < 0.001$)．また，Cohen's dで測られる効果量も中程度(0.588)の値を示している．ドメイン設定(in-domain および out-domain)が検索性能に有意に差を与えることは，例示検索においてドメイン適応が必要であることを示唆している．また同時に，同じ検索意図であっても検索のコンテキスト(検索する場所，検索対象データ)が異なれば正解となるデータも異なることを意味している．

5.4 相対的集約点に基づく手法と既存手法の比較

我々は RAP によるドメイン適応と他のベースラインの比較実験を行った．比較は out-domain 設定の 400 種類の検索(4 ドメイン × 100 クエリ)で行われた．表 2 は各検索意図ごとの平均 nDCG@10 を示している．一元配置分散分析を行った結果，nDCG@10 において手法による主効果が見られ($F(6,2793) = 40.0, p < 0.001$)，対データに対する t 検定によって，RAP によるドメイン適応と有意差が見られた手法にダガー†を付与している[6]．我々の手法は NNS と OSVM，TSVM と有意差がある．また，特にドメイン依存の検索意図に対して有効に働いているように見られる．

RAP による手法は特に検索意図 5, 11, 19 にて nDCG@10 を大きく改善させている．検索意図 5 は「外国からの客を接待するのに適切な 8,000 円ほどの店」，検索意図 11 は「クリスマスディナーに良い洋食の店」である[7]．両意図に共通するのは明確な予算またはジャンルの指定がなく，曖昧性をはらんでいる点である．この場合，予算やジャンルが地域によって変わりうるため，その変化を RAP によって捉えることで，検索性能を向上させたのではないかと考えられる．

6. まとめと今後の課題

本論文では例示検索におけるドメイン適応の必要性を実験的に示し，各ドメインにおける RAP を対応させることに

[5] http://glaros.dtc.umn.edu/gkhome/views/cluto/

[6] 有意水準は 0.05 とし，Holm の手法を使って多重比較を行っている．

[7] 本来の検索意図は長く詳細なものであるが，ここでは簡略化されている．

表 2　各検索意図ごとの nDCG@10.
Table 2　The average nDCG@10 for each search intent.

		Inductive			Transductive			
	Intent	NNS	OSVM	SVM	TSVM	SCL	RCM	RAP
Domain-dependent	1	0.191	**0.192**	0.175	0.176	0.171	0.177	0.173
	2	0.173	0.249	0.406	0.277	0.415	0.406	**0.420**
	3	0.149	0.123	0.084	**0.151**	0.103	0.084	0.086
	4	0.145	0.087	0.389	0.315	0.367	0.386	**0.395**
	5	0.182	0.167	0.242	0.250	0.260	0.242	**0.283**
	6	0.118	0.162	0.267	0.213	0.261	0.267	**0.268**
	7	0.024	0.046	**0.174**	0.121	0.170	**0.174**	0.168
	8	0.213	0.251	0.303	0.219	**0.311**	0.303	0.309
	9	0.199	0.184	0.254	0.223	0.258	0.254	**0.282**
	10	0.130	0.178	0.277	0.226	0.269	**0.278**	0.266
Domain-independent	11	0.208	0.172	0.589	0.582	0.617	0.589	**0.631**
	12	0.148	0.121	0.288	0.162	**0.296**	0.288	0.295
	13	0.201	0.198	0.295	0.262	0.283	0.295	**0.325**
	14	0.068	0.056	0.184	0.038	**0.194**	0.184	0.183
	15	0.285	0.230	0.483	**0.513**	0.509	0.483	0.492
	16	0.164	0.136	0.487	0.375	**0.496**	0.486	0.486
	17	0.107	0.135	0.418	0.230	**0.475**	0.418	0.440
	18	0.222	0.174	**0.368**	0.206	0.356	**0.368**	0.323
	19	0.305	0.345	0.492	0.378	0.493	0.492	**0.540**
	20	0.118	0.109	0.127	0.081	**0.128**	0.120	0.121
Total		0.168†	0.166†	0.315	0.25†	0.322	0.315	**0.324**

よりドメイン適応を行う方法を提案した．また，地理情報例示検索のためのテストコレクションを構築し，実験では例示検索にドメイン適応は必要であり，また，RAP によるドメイン適応は既存手法を有意に上回ることを明らかにした．将来的には，適切な RAP を自動的に発見する方法を開発したいと考えている．また，検索コンテキストが適合性判定に与える影響について，より詳しく調査する予定である．

[謝辞]

本研究の一部は，グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」（研究代表者：田中克己），文部科学省科学研究費補助金基盤 (A)「ウエブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究」（24240013，研究代表者：田中克己），文部科学省科学研究費補助金若手(A)「意味的に周辺にあるウェブ情報へのナビゲーションの研究 」（24680008，研究代表者：大島裕明），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」（22・4687，研究代表者：加藤誠）によるものです．ここに記して謝意を表します．

[文献]

[1] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In Proc. of NIPS, pp. 137–144, 2006.
[2] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proc. of ACL, pp. 440–447, 2007.
[3] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In Proc. of EMNLP, pp. 120–128, 2006.
[4] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Co-clustering based classification for out-of-domain documents. In Proc. of KDD, pp. 210–219, 2007.
[5] M. P. Kato, H. Ohshima, S. Oyama, and K. Tanaka. Search as if you were in your home town: geographic search by regional context and dynamic feature-space selection. In Proc. of CIKM, pp. 1541–1544, 2010.
[6] X. Ling, W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Spectral domain-transfer learning. In Proc. of KDD, pp. 488–496, 2008.
[7] S. Nakajima and K. Tanaka. Relative queries and the relative cluster-mapping method. In Proc. of DASFAA 2004, pp. 843–856, 2004.
[8] S. Pan and Q. Yang. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, 2010.
[9] G. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged pLSA for cross-domain text classification. In Proc. of SIGIR, pp. 627–634, 2008.

加藤 誠 Makoto P. KATO
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在学中．2009年京都大学大学院情報学研究科社会情報学専攻修士課程修了．主に情報検索の研究に従事．情報処理学会，日本データベース学会，人工知能学会，ACM各学生会員．

大島 裕明 Hiroaki OHSHIMA
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）．主にウェブ，情報検索，データベースの研究に従事．情報処理学会，電子情報通信学会，日本データベース学会，ACM各会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．京大工博．主にデータベース，Web 情報検索，Web マイニング，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．