アナロジーに基づく情報検索

加藤　誠†

† 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
E-mail: †kato@dl.kuis.kyoto-u.ac.jp

あらまし　知らない分野の情報を検索し理解することは難しい．これは，その分野について知識がなければ，適切な質問（クエリ）が思いつかず，また，得られた情報（検索結果）をうまく解釈できないためである．本講演では未知ドメインに対する新しい検索方法として，アナロジーに基づく情報検索について紹介する．

キーワード　アナロジー，情報検索

1. はじめに

一般に，知らない分野について何らかの情報を得ることは難しい．これは，ある情報を得るためにはどのような手段を使ったら良いのか（探索），また，得られた情報が本当に欲しかった情報なのか（判断），欲しかった情報だとして，その内容が何を意味しているのか（理解）わからないことがしばしばあるためである．我々はこの問題に対して，アナロジーに基づく情報検索手法を提案してきた．この検索手法では，知っている分野に関する情報を与えることで，それに対応する知らない分野の情報を検索することができる．本稿では比例アナロジーに基づく情報検索，アナロジーに基づく地理情報検索，アナロジーに基づく情報検索のためのドメイン適応について紹介する．

2. 比例アナロジーに基づく情報検索

現在，多くの情報検索ではキーワードクエリが採用されているが，検索対象に含まれている語や検索対象の名称を知っていなければ目的の情報にたどり着くことは難しい．このような制約の元では，知らない分野に対する情報検索は困難であると考える．そこで，我々はアナロジーの中でも比較的単純な部類に属する比例アナロジーを利用した情報検索手法を提案し，その 1 つの実現手法として共起を用いた手法について述べる．比例アナロジーは，4 つの項，A，B，C，D において，A と B の関係と C と D の関係が等しいときに成り立ち，このとき $A:B::C:D$ と表現されるものである．例として，月:地球::エウロパ:木星や大工:木::石工:石などが挙げられる．比例アナロジーに基づく検索とは，入力として A，B，C が与えられた時に，D を何らかの情報源から得る検索のことである．たとえば，A=“月”，B=“地球”，C=“エウロパ” が入力として与えられた時，D=“木星” が出力として得られる．D に対してもっともらしい順に結果は順位づけされ，正解である「木星」を上位に出力できれば，その検索アルゴリズムは優れているといえる．この例の場合，ユーザは「月」が「地球」にとって何であるのか知る必要はなく，3 つの項の例示だけによって言語化されていない情報が伝達できる点が比例アナロジーによる検索の利点である．

比例アナロジーについては Turney ら [20] が語 A と語 B の組が与えられたときに，予め用意された複数の語 C と語 D の組のうち，A と B の関係と最も近い C と D の組を選択するという比例アナロジーの問題に対して様々なアプローチによる研究を行ってきた．また，Bollegala ら [5] も同様の問題に Web からの言語パターン抽出と SVM による学習で一定の成果をあげている．これらの研究は，すでに与えられた解候補の中から最も適したものを選択する問題に焦点を当てており，入力として A と B，C を与えて，D を検索する我々の手法とは，関係の類似に着目する面では共通するものの，根本的に異なるものである．

2.1 提案手法

我々は語の共起を用いて，類似関係を検索する方法を提案した．この手法では語 A, B が共に出現する文書においてのみ，著しく出現確率が高くなる語を t とし，語 t と語 C が共に出現する文書において，$A:B::C:D$ を満たす語 D の出現確率が高くなることを仮定している．これは，語 t は語 A と B の関係を表していると考えられ，語 A と B と類似関係にある語 C と D は語 t と頻繁に共起すると考えられるためである．

より具体的な方法について以下で述べる．最初に，入力 A と入力 B を結びつけるような語集合 T を Web 検索を用いることによって取得する．ここで，A と B を結びつける語は，A と B を共に含む文書において出現確率が高くなると仮定している．

（1） 入力 A 及び入力 B に対して，A を含み B を含まない文書を検索するクエリと，B を含み A を含まない文書を検索するクエリで Web 検索を行い，上位 n 件の検索結果のタイトルとスニペットを取得する．

（2） 入力 A と入力 B を含む文書を検索するクエリで Web 検索を行い，検索結果のタイトルとスニペットを取得する．

（3） タイトルとスニペットに対して形態素解析を行い，形態素ごとに分割する．また，ストップワードリストを用いて不要な語を除去する．

（4） 品詞が「名詞」であると推定された各語 t_i に対して，“語 A を含み語 B を含まない文書と語 A と B を共に含む文書において，語 t_i の出現確率が等しい” という帰無仮説に対して χ^2 検定を行う．同様に，語 B を含み語 A を含まない文書と語

A と B を共に含む文書に対しても検定を行う．

（5） 有意水準 α の検定において棄却された語 t_i のうち，語 A，B が出現したときに出現確率が高くなるものを，入力された語 A と B を結びつける語として採用し，これを語集合 T とする．

次に，語集合 T を用いて，入力 A と B の関係と類似するような，入力 C に対する語 D を発見する．ここで，A と B の関係と類似する語は，A と B を結びつける語 t_i と C を共に含む文書において出現確率が高くなると仮定している．

（6） 語集合 T の全ての語 t_i に対して，入力 C を含み語 t_i を含まない文書を検索するクエリと，t_i を含み C を含まない文書を検索するクエリで Web 検索を行い，上位 n 件の検索結果のタイトルとスニペットを取得する．

（7） 語集合 T の全ての語 t_i に対して，入力 C と語 t_i を含む文書を検索するクエリで Web 検索を行い，検索結果のタイトルとスニペットを取得する．

（8） タイトルとスニペットに対して形態素解析を行い，形態素ごとに分割する．また，ストップワードリストを用いて不必要な語を除去する．

（9） 品詞が「名詞」であると推定された各語 d_j に対して，“語 C を含み語 t_i を含まない文書と語 C と t_i を共に含む文書において，語 d_j の出現確率が等しい”という帰無仮説に対して χ^2 検定を行う．同様に，語 t_i を含み語 C を含まない文書と語 C と t_i を共に含む文書に対しても検定を行う．両者の検定の結果，帰無仮説が真になる確率をそれぞれ $P_C(d_j)$，$P_{t_i}(d_j)$ とする．

（10） 有意水準 β の検定において棄却された語 d_j のうち，語 C，t_i が出現したときに出現確率が高くなるものに対して，$P_C(d_j)$，$P_{t_i}(d_j)$ の積を $P_{C,t_i}(d_j)$ とする．

（11） 語集合 T の全ての語 t_i に対する，$P_{C,t_i}(d_j)$ の全ての積を語 d_j のスコアとする．

語 d_j はそのスコアが小さい順に順位づけされユーザに提示される．これは，語に対するスコアが小さければ小さいほど，その語の出現確率が有意に高くなることを表しているためである．

2.2 評価実験

提案手法の有効性を示すために，我々は 33 の関係クラスと 854 の正解セットを用いて提案手法の性能評価を行った．正解セットの一例を表 1 に示す．全ての正解セットを用いて 2 つのベースライン手法と比較した結果を表 2 に示す．ベースライン 1 は TF-IDF 法を利用した手法で，ベースライン 2 は大島ら [15] による言語パターンを利用した手法である．@5，@10，@20 はそれぞれ 5，10，20 位に正解を含めることが出来たテストの割合を表している．提案手法と言語パターン，ベースライン手法を比較したとき，MRR，@5，@10，@20 において顕著な差が出ている．出力として期待される D を上位 20 件以内に含めることができた正解セットは全体の中の 49.8%で，これはベースライン手法よりも遙かに優れているが，まだ精度的に不十分であると思われる．

大島ら [15] による言語パターンを利用した手法と提案手法を組み合わせたより詳細な実験については，我々の論文 [11] を参照してほしい．

表 1 正解セットの例．

A(入力)	B(入力)	C(入力)	D(正解出力)
千葉	落花生	静岡	茶
日本	相撲	韓国	テコンドー
チョコレート	カカオ	かまぼこ	魚
長崎	カステラ	京都	八ツ橋
名古屋	手羽先	広島	お好み焼き
フランス	パリ	カナダ	オタワ
野球	バット	ゴルフ	クラブ
チョウ	イモムシ	トンボ	ヤゴ
クジラ	哺乳類	カエル	両生類
舞姫	森鴎外	人間失格	太宰治
金閣寺	足利義満	銀閣寺足	利義政
江戸幕府	徳川家康	鎌倉幕府	源頼朝
月	地球	エウロパ	木星
キャビア	チョウザメ	とびこ	トビウオ

表 2 比例アナロジーに基づく検索手法の実験結果．MRR は平均逆順位を，@k は k 位以内に正解を得られたテストデータの割合を表している．

	MRR	@5	@10	@20
ベースライン 1	0.045	8.3	10.8	18.7
ベースライン 2	0.150	24.5	30.9	33.8
提案手法	**0.249**	**34.4**	**42.1**	**49.8**

3. アナロジーに基づく地理情報情報検索

3.1 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(注1)や食べログ(注2)，ホテル情報サイトの Booking.com(注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，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，情報が爆発的に増大する昨今においては，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オブジェクト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

(注1)：http://www.gnavi.co.jp/
(注2)：http://tabelog.com/
(注3)：http://www.booking.com/

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オブジェクト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) 中の地理オブジェクトが検索される．検索結果は地図の下方に表示される．

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．

3.2　アナロジーに基づく地理情報情報検索システム

我々は地理オブジェクト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 [1]．例示検索は特にマルチメディア検索で用いられる検索方法 (Content-based retrieval [6], [7], [18]) 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に基づいて検索する，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．

例示して検索するという考えは，自然に地理オブジェクト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，右が地理オブジェクト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オブジェクト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オブジェクト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．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，現実的ではない．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．

3.3　アナロジーに基づく地理情報情報検索システムの問題点

入力された例に基づいて地理オブジェクトをランキングするために，オブジェクト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：

（1）　類似性尺度はユーザ，コンテキストによって異なる．また，

（2）　異なる分野におけるオブジェクト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，さらには，そのコンテキストによって変化する [21]．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オブジェクト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 つのオブジェクト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．

第 1 の問題については，ユーザが与えた例から類似性尺度を推定する手法の一つに石川らの MindReader [8] がある．しかしながら，MindReader を我々の問題に適用した場合，過学習の問題から十分な推定精度が得られないことが予想される．検

索のために入力されるキーワードなどのように，検索時にユーザから入力される情報は一般に多くない．特に Content-based search や例示検索の場合には入力方法が難しいことから，多くの入力をユーザから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と考えられる．そのため，距離尺度推定手法は少数の入力であっても高い精度を達成しなくてはならない，すなわち，推定手法は「頑健」でなくてはならない．我々は MindReader を拡張することによって，選択された例と選択されなかった例の情報を用い，両者の差異を増幅させる，頑健な類似性尺度推定手法を提案した．この結果については我々の論文 [12] に詳細が書かれている．以降，本稿では第 2 の問題について述べる．

4. アナロジーに基づく情報検索のためのドメイン適応

本節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．

4.1 問題設定

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 [16]．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 $x_i^{(T)} \prec x_j^{(T)} \Leftrightarrow f(x_i^{(T)}) < f(x_j^{(T)})$ で定義される．また，得られる順序集合 $(X^{(T)}, \prec)$ 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 $X^{(S)} = \{x_1^{(S)}, x_2^{(S)}, x_3^{(S)}\}$ は周辺確率 $P(X^{(S)})$ 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 $Q = \{x_1^{(S)}, x_2^{(S)}\}$ を選択し，これは $D^{(S)} = \{(x_1^{(S)}, +1), (x_2^{(S)}, +1), (x_3^{(S)}, -1)\}$ 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 $X^{(T)} = \{x_1^{(T)}, x_2^{(T)}, x_3^{(T)}\}$ 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 [2], [17]．$\mathcal{A}$ 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．

4.2 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．

4.2.1 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

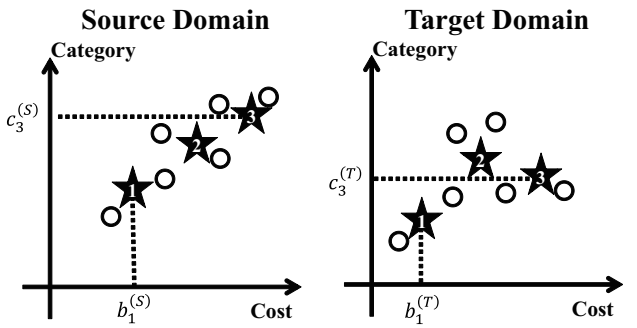

図 2　RAP とそれに基づく特徴表現．黒い星は RAP であり，白い丸はインスタンスを表している．同じ数字を持つ RAP は対応している．

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が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：

$$a_\psi = \int_{x \in \mathcal{X}} \mathbf{x} P(x|\psi) dx, \qquad (1)$$

ただし，$\mathbf{x}$ はインスタンス x を表すベクトルであり，$P(x|\psi)$ 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 $P(x|\psi)$ は下記のように得られる．

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

ただし，$P(x|\psi, X)$ 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 $P(x|\psi)$ は確率的な所属度合いを表している．

図 2 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 2 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．$H = \{h_1, h_2, \dots h_m\}$ を類似度関数の集合，$A = \{a_1, a_2, \dots a_n\}$ 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：

$$\phi(x) = (\phi_1(x), \phi_2(x), \dots, \phi_n(x)) \qquad (3)$$

ただし，$\phi_i(x) = (h_1(x, a_i), h_2(x, a_i), \dots, h_m(x, a_i))$ 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 $(\mathbf{x}, \lambda\phi(x))$ が得られる．ただし，λ 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 $(\mathbf{x}^{(S)}, \lambda\phi^{(S)}(x^{(S)}))$ 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 $(\mathbf{x}^{(T)}, \lambda\phi^{(T)}(x^{(T)}))$ 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，$\phi^{(S)}(x^{(S)})$ と $\phi^{(T)}(x^{(T)})$ 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．

具体的な RAP の例は我々の論文 [10] に書かれている．また，テストコレクションの詳細についても上記の論文を参照してほしい(注4)．

4.3 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．

4.3.1 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

(注4)：Geographic Object Retrieval Dataset http://www.mpkato.net/jp/datasets/

表 3　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．

		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**

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 MAP および 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 [13], [22]．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], [4]．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 [14].

4.3.2　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 3 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 ($t(179) = 5.89$, $p < 0.001$)(注5)．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．これまでの研究ではユーザによって [19]，また，検索デバイスによって [9] 正解となるデータが異なることが示されてきたが，今回の結果から検索のコンテキストもまた適合データを左右する要素の 1 つであると考えられる．

4.3.3　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 4 および 5 は各検索意図ごとの平均 MAP および平均 nDCG@10 を示している．一元配置分散分析を行った結果，MAP と nDCG@10 において手法による主効果が見られ ($F(6, 2793) = 16.1$, $p < 0.001$ および $F(6, 2793) = 40.0$, $p < 0.001$)，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(注6)．我々の手法はすべての手法と MAP において有意差があり，nDCG@10 では 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

(注5)：我々は同様に MAP でも in-domain 設定 (0.609) と out-domain 設定 (0.479) の有意差を確認した ($t(157) = 5.35$, $p < 0.001$, Cohen's $d = 0.585$).

(注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．

表 4　各検索意図ごとの平均 MAP.

		Inductive			Transductive			
	Intent	NNS	OSVM	SVM	TSVM	SCL	RCM	RAP
Domain-dependent	1	0.539	**0.546**	0.490	0.417	0.482	0.490	0.490
	2	0.387	0.452	0.522	0.439	0.512	0.521	**0.574**
	3	**0.347**	0.319	0.317	0.344	0.321	0.317	0.320
	4	0.409	0.350	0.672	0.558	**0.688**	0.672	0.668
	5	0.488	0.458	0.512	0.506	0.515	0.513	**0.522**
	6	0.198	0.227	0.327	0.329	0.324	0.327	**0.333**
	7	0.078	0.083	**0.155**	0.150	0.153	**0.155**	0.144
	8	0.192	0.219	0.254	0.206	0.251	0.254	**0.260**
	9	0.488	0.491	0.548	0.561	0.548	0.548	**0.570**
	10	**0.261**	0.293	0.335	0.305	0.329	0.335	**0.337**
Domain-independent	11	0.442	0.401	0.630	0.581	0.631	0.630	**0.654**
	12	0.180	0.177	0.280	0.233	**0.284**	0.280	**0.284**
	13	0.586	0.571	**0.645**	0.627	0.638	**0.645**	0.640
	14	0.115	0.114	**0.161**	0.109	**0.161**	**0.161**	0.156
	15	0.354	0.293	**0.545**	0.506	0.541	**0.545**	0.538
	16	0.416	0.399	0.702	0.609	0.686	0.701	**0.705**
	17	0.270	0.270	0.379	0.317	0.384	0.379	**0.385**
	18	0.256	0.226	**0.401**	0.312	0.398	**0.401**	0.399
	19	0.586	0.595	0.641	0.532	**0.646**	0.641	0.645
	20	**0.288**	0.286	0.238	0.203	0.236	0.238	0.239
Total		$0.344^{\dagger}$	$0.338^{\dagger}$	$0.438^{\dagger}$	$0.392^{\dagger}$	$0.436^{\dagger}$	$0.438^{\dagger}$	**0.443**

表 5　各検索意図ごとの平均 nDCG@10.

		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		$0.168^{\dagger}$	$0.166^{\dagger}$	0.315	$0.250^{\dagger}$	0.322	0.315	**0.324**

スディナーに良い洋食の店」である(注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．

5. ま と め

本稿では未知分野における検索の困難性に対して提案された，アナロジーに基づく情報検索手法を紹介した．特にこれまで発表してきた，比例アナロジーに基づく情報検索，アナロジーに基づく地理情報検索，アナロジーに基づく情報検索のためのドメイン適応についてそれぞれ説明をした．

文　献

[1] D. Angluin and C. Smith. Inductive inference: Theory and methods. *ACM Computing Surveys*, 15(3):237–269, 1983.

[2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Proc. of NIPS*, pp. 137–144, 2006.

[3] 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.

[4] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proc. of EMNLP*, pp. 120–128, 2006.

[5] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring the similarity between implicit semantic relations from the web. In *Proc. of WWW 2009*, pp. 651–660, 2009.

[6] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[7] V. Gudivada and V. Raghavan. Content based image retrieval systems. *Computer*, 28(9):18–22, 2002.

[8] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. In *Proc. of VLDB 1998*, pp. 218–227, 1998.

[9] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *Proc. of WWW*, pp. 801–810, 2009.

[10] M. P. Kato, H. Ohshima, and K. Tanaka. Content-based retrieval for heterogeneous domains: domain adaptation by relative aggregation points. In *Proc. of SIGIR 2012*, pp. 811–820, 2012.

[11] M. Kato, H. Ohshima, S. Oyama, and K. Tanaka. Query by analogical example: relational search using web search engine indices. In *Proc. of CIKM 2009*, pp. 27–36, 2009.

[12] M. 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 2010*, pp. 1541–1544, 2010.

[13] X. Ling, W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Spectral domain-transfer learning. In *Proc. of KDD*, pp. 488–496, 2008.

[14] S. Nakajima and K. Tanaka. Relative queries and the relative cluster-mapping method. In *Proc. of DASFAA 2004*, pp. 843–856, 2004.

[15] H. Ohshima and K. Tanaka. High-speed Detection of Ontological Knowledge and Bi-directional Lexico-Syntactic Patterns from the Web. *Journal of Software*, 5(2):195–205, 2010.

[16] S. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[17] P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian. Domain adaptation meets active learning. In *Proc. of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pp. 27–32, 2010.

[18] 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):1349–1380, 2000.

[19] J. Teevan, S. Dumais, and E. Horvitz. Potential for personalization. *ACM Transactions on Computer-Human Interaction*, 17(1):1–31, 2010.

[20] P. Turney and M. Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning Journal*,

(注7)：本来の検索意図は長く詳細なものであるが，ここは簡略化されている．

60(1-3):251–278, 2010.

[21] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[22] 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.