RD-005

行動トピックベクトルと地域語特徴を用いた検索キーワードに対する行動タイプ付与

Action Type Assignment to Query Terms Based on Action Topics and Location Name Features

羽田野真由美† 数原良彦† 戸田浩之† 小池義昌†

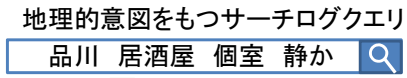
Mayumi Hadano Yoshihiko Suhara Hiroyuki Toda Yoshimasa Koike

1 はじめに

近年，スマートフォンやタブレット端末の普及に伴い，外出先で情報検索サービスが利用される場面が増えている．モバイル端末での情報検索サービス利用方法の特徴として，実世界で行動を行う際の意思決定のために利用されることが多いことが報告されている [1]．たとえば外出先で食事をとる際，情報検索サービスを用いて利用する飲食店を決めるという意思決定を行い，その後決定した飲食店に向かうケースがその代表である．

実世界での行動に関する検索では，飲食店や宿泊施設，観光地などの行動カテゴリごとに存在する専門検索 [1] を利用した検索サービスが利用されていることが多い．しかしながらこの場合，ユーザは行動カテゴリ (以後行動タイプという) ごとに異なった検索インターフェースを操作する必要がある．そこで，ウェブ検索エンジンに用いる検索窓のような 1 つのインターフェース上で入力されたクエリに対して，システムがその行動タイプを自動判定することができれば，行動タイプに応じた専門検索の結果を出力結果に利用することが可能となる．これにより，ユーザの操作量が減り，すばやく必要とする情報にたどり着くことが可能となる．また，行動タイプをキーワード単位で付与することができれば，キーワード間の関係性が把握でき，より詳細な検索意図を抽出することができる．たとえば，図 1 で示したクエリの例では，クエリを構成するキーワードごとに地名語 (geo)，属性語 (attribute)，行動語 (action) というキーワードタイプが付与され，行動語に対しては食事という行動タイプが付与されている．これによって，システムは「品川」に立地していて「個室」があり「静か」という条件に合う「居酒屋」を食事に関する専門検索への入力として与えることが可能となる．

本研究では，上記のように，行動タイプの自動判定によって 1 つのインターフェースで専門検索が利用できるシステムの実現を目標に，入力クエリを構成する各キーワードに対して (1) 地名語，(2) 属性語，(3) 行動語という 3 種類のアノテーションを行う機能の実現を目指す (図 1)．本研究においては，実世界における行動意図を持つクエリは地理的意図を含むクエリと仮定する．地理的意図とは，地理的な範囲を対象にして行動を起こす意思があることである．たとえばウェブ上でユーザが知らない知識について調べ物を行う場合には地理的意図を含まない．



	分類項目	品川	居酒屋	個室	静か
1	キーワードタイプ	*geo*	*action*	*attribute*	*attribute*
2	行動タイプ	-	2: 食	-	-

図 1 本研究の解く課題

検索クエリが地理的意図を含むかの判定は Yi ら [2] の方法によって可能であるため，本稿では検索クエリの中から地理的意図を含むクエリのみが選択され，システムへの入力として与えられるという状況を想定する．システムは，入力されたクエリをキーワードに分割し，それぞれのキーワードに対して地名語，属性語，行動語のいずれかのキーワードタイプを付与する (ステップ 1)．その後，行動語に分類されたキーワードに対して，事前に準備した 12 の行動タイプに分類する (ステップ 2)．本研究では，ステップ 1 のキーワードタイプ分類は所与のものと考え，ステップ 2 の行動タイプ分類の課題に取り組む．

本研究では，一定量の正解行動タイプが付与された訓練データが利用可能な状況を想定し，教師あり学習の枠組みを用いて，マルチクラス分類問題としてステップ 2 を定式化し，その解決に取り組む．マルチクラス分類問題においては，分類対象であるキーワードに対応する事例に特徴ベクトルを生成し，特徴ベクトルを入力として行動タイプに対応するクラスを判別する分類器を生成することで，クラスが未知なキーワードに対して特徴ベクトルが生成でき，行動タイプを予測することが可能となる．

クエリログを利用した従来の文書分類において，分類器を学習する際の特徴には，判定対象のクエリに含まれる語の bag-of-words を利用したり，クリックスルー率を用いたりすることが多い [3]．しかしながら，この手法は訓練データに出現しないキーワードに対して特徴を付与することが出来ないという課題があった．この問題は特に訓練データが少ない場合において，適切な特徴抽出ができないため，生成された分類器の精度低下につながるおそれがある．このような問題を本稿では未知語問題と呼ぶ．この問題を解決するために，正解ラベルが付与されていないクリックログデータを用いて，キーワード-ドメイン名クリック頻度情報を行列分解することによって計算した行動トピックベクトルを特徴生成に利用する手法を提案し，

† 日本電信電話株式会社
NTT サービスエボリューション研究所
NTT Service Evolution Lobaratories, NTT Corporation

[1] https://ssl.gstatic.com/think/docs/creating-moments-that-matter_research-studies.pdf

分類器の精度向上を目指す．

もうひとつの課題として，従来の手法では地域語を他の単語と区別した特徴として利用していないことが挙げられる．一般に出現するキーワードが表す行動タイプの分布は地域によって変化すると考えられるため，地域ごとの特徴を持たないことで，行動タイプの予測が適切に行われていないおそれがある．この問題を解決するために，地域語毎の行動タイプの分布を明示的に特徴として用いることによって，より高精度な行動タイプ判定を目指す．

本稿では，検索エンジンのクエリログに対して人手でキーワードアノテーションを行った評価用データとクリックログデータを用いて実験を行い，行動タイプ分類の評価を行った．まず単純な条件付き確率に基づくルールベースの手法を用いた予備実験を行い，行動タイプ分類において解くべき課題が特に未知語問題の解決であることを示した．その後，ベースラインとなる出現キーワードに基づく特徴抽出手法と提案手法とを比較した結果，提案手法が未知語においてより高い Accuracy と F1 値をとり，より高精度に行動タイプ分類を行うことができることを示した．

本研究の貢献は以下の 3 点である．

- クリックログデータを利用して作成した単語-ドメイン名のクリック頻度行列を行列分解して得られた行動トピックベクトルを，行動タイプ分類におけるキーワードの特徴として用いる方法を提案する．
- 訓練データに含まれる行動語とクエリ内で共起する地名語の情報を利用して，共起する地名語の行動タイプ分布情報をキーワードの特徴として用いる方法を提案する．
- 上記 2 つの手法を出現キーワードに基づく特徴のみを用いた手法と比較し，特に訓練データに出現しない未知のキーワードに対して，予測モデルの精度が向上したことを確認した．

2 関連研究

サーチログにおけるユーザの検索意図を分類する研究は今までにも数多く行われてきた．Broder[4] は informational，navigational，transactional という 3 つのユーザ意図を定義し，Rose ら [5] はそれを 11 のサブカテゴリに分類した．廣嶋ら [6] は，システム応答を変化させるための条件であるクエリの種別をクエリタイプと呼び，概念ベースを用いてクエリタイプを分類した．しかしながらこれらの研究は，検索意図をクエリ単位で付与するものであり，本研究のようにキーワード単位に細かくラベルを付与し，キーワード間の関係性を抽出するものではない．

本研究はキーワードごとにキーワードタイプを分類し，そのキーワード間の関係に着目するため，エンティティサーチに関する研究と関連する．Lin ら [7] は，あるエンティティに対しての行動に着目して，行動を抽出するという取り組みを行っている．ここでの行動例としては，商品名というエンティティに対して，レビューを「読む」，オンラインで「買う」，デモビデオを「見る」などが挙げられる．しかしながら，この研究は行動を自動抽出するものであり，本研究のように，事前に設定した行動タイプにマルチクラス分類するというものではない．

本研究が対象とする地理的意図をもつクエリを予め抽出するためには，以下の研究が利用できる．Jones ら [8] は地名を含むクエリの分析を行っており，IP アドレスで推定したユーザ位置とクエリ内地名間の距離分布がクエリのトピックによって変化することを報告している．Welch ら [3] は，検索結果にユーザの現在地に基づく情報が含まれることが暗黙的に望まれているかどうかを教師あり学習で判定するという研究を行っている．Yi ら [2] は，地名を含まないクエリにおいて，そのクエリが地理的意図をもつかどうかを教師あり学習で判定し，Accuracy で 89% の精度の結果を得ている．

本研究が対象とする行動タイプ付与の前処理として，地名語と属性語を判別する既存研究は以下が挙げられる．地名語の判別に関する研究としては，地名 DB を用いて入力文書の位置候補を取得したのちに，距離や有名度などの指標を用いて実際の位置を推定する手法が多くとられている．たとえば，平野ら [9] は，有名度として店舗 DB の店舗数を用いている．ほかにも，Li ら [10] は，地名の曖昧性解消のために地名階層の上下関係を用いている．属性語の判別に関する研究としては，サーチログの文脈パターンを用いる手法が多い．たとえば Paşca[11] は，サーチログの文脈パターンを用いて特定のクラスに属する属性を，インスタンスシードと属性シードを利用して自動抽出する手法を提案している．その結果たとえば，「会社」というクラスの属性として「場所」，「CEO」，「本社」，「株価」などの属性を自動抽出している．インスタンスシードと属性シードさえ準備することが出来れば，Paşca[11] の方法をそのまま本研究に用いることができる．

3 問題設定

3.1 用語の定義

定義 1 (地理的意図)．地理的な範囲を対象にして行動を起こす意思があること．本研究では，実世界で行動を行うユーザに対して適切な情報を提供し，ナビゲーションにつなげることを目標としているため，地理的意図を含む検索クエリを対象にする．なお，地理的意図を含む検索クエリを構成するキーワードは地名語，行動語，属性語のいずれかに分類される．本研究ではこの 3 つをキーワードタイプと呼ぶ．

定義 2 (地名語)．実世界の地域を指し示すキーワードのこと．具体的には，市区町村名や駅名，ランドマーク名などが挙げられる．図 1 のクエリの例では，「品川」が地名語にあたる．地名語はユーザの検索意図において地理的条件を示している．

定義 3 (行動語)．実世界の地域において，ユーザが行動を行う (予定がある) ことを示唆するキーワードのことである．図 1 のクエリの例では，「居酒屋」という単語からユーザが実世界で食に関する行動を行うことが示唆され

表 1　行動タイプ一覧

行動タイプ ID	行動タイプ	典型的なキーワード
1	宿泊	ホテル，民宿
2	食事	ランチ，居酒屋
3	移動	乗り換え，駐車場
4	観光	観光，お土産
5	買い物	セール，福袋
6	趣味・娯楽	映画，サッカー
7	受診	耳鼻科，診療時間
8	交際・付き合い	デート，接待
9	求職	バイト募集，派遣募集
10	物件探し	マンション，賃貸
11	アダルト	-
12	その他の行動	-

ていると考えられるため，「居酒屋」が行動語にあたる．

定義 4 (属性語). 他のキーワードに対して条件を付与する役割をもったキーワードのこと．図 1 のクエリの例では，ユーザは「個室」があって，「静か」という条件の「居酒屋」を検索していると考えられるため，「個室」「静か」の両者が属性語にあたる．

定義 5 (行動タイプ). 実世界で行う行動のカテゴリのこと．すべての行動語は必ずいずれかの行動タイプに分類される．本研究では，社会生活基本調査 [2]における生活行動分類を参考に，日常で行われやすいものを抽出・再分類した 12 種類の行動タイプを設定する (表 1)．

定義 6 (地理的意図アノテーション). キーワード列として与えられる地理的意図を含むクエリ $Q = (w_1, w_2, \cdots, w_n)$ を構成するキーワード $w_i (i = 1, \cdots, n)$ に対して地域語，属性語，行動語のいずれかのアノテーションを行う．行動語については行動タイプを付与する．すなわち入力されたキーワード列に対応するアノテーション列 $A = (a_1, a_2, \cdots, a_n)$ を出力する．アノテーション $a_i = (t_i, c_i)$ はキーワードタイプ $t_i \in \{\text{“}geo\text{”}, \text{“}attribute\text{”}, \text{“}action\text{”}\}$ と行動タイプ ID である $c_i \in \{0, 1, \cdots, 12\}$ を持つ．行動タイプ ID は，$t_i = \text{“}action\text{”}$ のとき $1 \leq c_i \leq 12$ であり，それ以外の場合は $c_i = 0$ とする．行動タイプ IDc_i の値は表 1 の行動タイプ ID と一致する．地理的意図アノテーションは，t_i の判定であるキーワードタイプ分類と，$t_i = \text{“}action\text{”}$ である場合に c_i の値を推定する行動タイプ分類の 2 段階の処理で構成される．

3.2　問題の設定

N 個のアノテーション付きクエリデータ $D = \{(Q^{(i)}, A^{(i)})\}_i^N$ が与えられた際に，入力された地理的意図を含むクエリ Q について地理的意図アノテーションを行う．本稿では，キーワードタイプ分類は正しく行われるものと仮定して，$t_i = \text{“}action\text{”}$ のキーワードに対する行動タイプ分類のみを対象とする．

[2] http://www.stat.go.jp/data/shakai/2011/

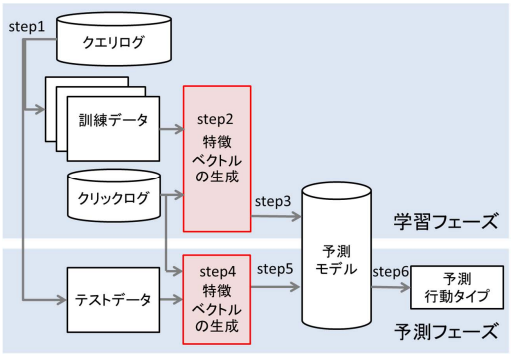


図 2　全体の処理の流れ

表 2　特徴一覧

特徴種類	特徴	説明
Basic Features	$P(t\|w)$	キーワード w における行動タイプ t の確率
	co_v	共起語の bag-of-words ベクトル
	$click_i$	キーワードのクリックドメイン頻度
	$r_{weekday}$	検索日の平日割合
	$time_k$	検索日の時間帯割合
Geo Features	$P(t\|geo)$	地名語 geo が出現した時の行動タイプ t の条件付き確率
SVD Features	s_d	SVD を用いて語彙空間を d 次元に削減した後のベクトル

4　マルチクラス分類を用いた行動タイプ判定

本章では，入力された検索クエリに含まれる行動語について，マルチクラス分類の枠組みで行動タイプ判定を行う方法を述べる．全体の処理の流れを 4.1 で述べ，教師あり学習に利用する特徴の抽出手法について，4.2 では既存の手法，4.3 と 4.4 では提案手法について説明する．

4.1　全体の処理の流れ

マルチクラス分類の枠組みで，行動タイプ判定を行う手法の全体の構成を図 2 で示す．まず，サーチログからサンプリングしたアノテーション付きクエリログを，行動タイプの予測モデル生成用に用いる訓練データと，生成した予測モデルの性能評価用のテストデータとに分割する (step1)．予測モデルの学習フェーズではまず訓練データとクリックログデータを用いて特徴ベクトルを計算する (step 2)．我々の提案手法はこの特徴ベクトルの生成方法に新たな方法を用いることである．用いた特徴ベクトルについては表 2 にまとめ，その生成方法については，4.2 で既存の手法，4.3 と 4.4 で提案手法を説明する．次に，計算した特徴ベクトルと正解行動タイプを教師あり学習アルゴリズムに入力し，予測モデルを生成する (step 3)．予測フェーズにおいては，テストデータから特徴ベクトルを計算し (step 4)，予め生成した予測モデルを用いて (step5), 行動タイプを予測する (step 6)．

4.2　出現キーワードに基づく特徴抽出

ここでは，マルチクラス分類を行う際の特徴ベクトルとして，既存手法 [3] でも一部が用いられている出現キーワードに基づいた特徴抽出方法について説明する．な

お，本節で説明する出現キーワードに基づく特徴を Basic features と呼ぶ．本稿では行動語と判定されたキーワードについて行動タイプ判定を行う問題設定であるため，これよりキーワードと呼ぶ場合には，行動語と判定されたキーワードのことを指す．

キーワード情報に基づく行動タイプ出現度合： $P(t|w)$

出現キーワード w が行動語であるとアノテーションされたとき，行動タイプ t にアノテーションされる条件つき確率 $P(t|w)$．この確率の計算は，行動タイプの集合を T とすると，以下の式で推定できる:

$$P(t|w) = \frac{Count(t,w)+1}{\sum_{t' \in T} Count(t',w) + |T|}. \quad (1)$$

ここで T は行動タイプ集合であり，$|T|$ は行動タイプ数を表す．本研究では，ゼロ頻度問題の回避のために，ラプラススムージング法 [12] を用いた．

共起語の bag-of-words： co_v

出現キーワード w と同じクエリで共起するキーワードの出現頻度を要素とする bag-of-words ベクトル．このベクトルの長さは，訓練データ中のクエリの語彙数となる．

キーワードのクリックドメイン頻度： $click_i$

クリックログデータを用いてキーワード w を含むクエリによってクリックされた URL のドメイン名とその頻度を集計する．上位 $i-1$ 件までの頻度に，i 位以上の頻度を合計したものを i 番目の結果とみなして追加し，i 次元ベクトルとする．本研究では，$i = 20$ とした．

キーワードの平日割合： $r_{weekday}$

クエリログからキーワード w を含むクエリ q の集合を取得して Q とする．Q の要素数を $Count(Q)$，Q に含まれるクエリのうち，検索日時が平日であるクエリ数を $Count(Q, weekday)$ としたとき，キーワード w の平日割合 $r_{weekday}$ は，以下の式で計算できる:

$$r_{weekday} = \frac{Count(Q, weekday)}{Count(Q)}. \quad (2)$$

検索日時の時間帯割合： $time_k$

クエリログからキーワード w を含むクエリ q の集合を取得して Q とする．Q の要素の数を $Count(Q)$，q の検索時間帯が k である場合のクエリ数を $Count(Q,k)$ としたとき，キーワード w の平日割合 $time_k$ は，以下の式で計算できる:

$$time_k = \frac{Count(Q,k)}{Count(Q)}. \quad (3)$$

なお，検索時間帯 k は $k = 1,2,3,4$ のときそれぞれ 03:00-08:59，09:00-14:59，15:00-20:59，21:00-02:59 であるとした．

4.3 クリックログを用いた行動トピックベクトル抽出

クリックログを用いて作成した単語-ドメイン名の頻度行列を特異値分解 [13] することで得られた各単語の潜在ベクトルを行動トピックベクトルと呼び，これを特徴として用いる方法を提案する．我々は，クリックログの検索クエリに含まれるキーワードとクリックされたドメイン名，およびその頻度の対応関係がユーザの行動に関する潜在的な意図を反映していると仮定し，この情報を元に生成されたキーワードに対応する潜在ベクトルは行動タイプ分類の特徴ベクトルとして有用だと考えた．

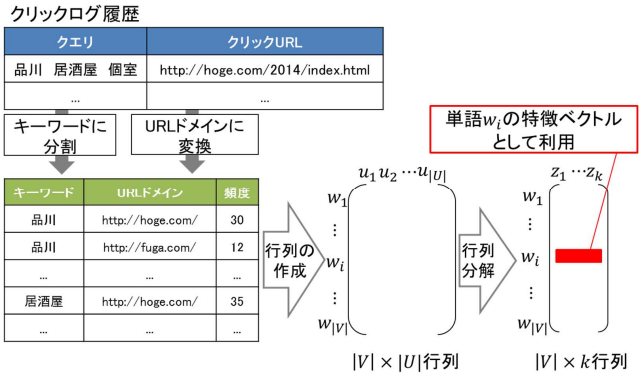


図 3　行動トピックベクトル抽出の流れ

4.2 で説明した従来の出現キーワードに基づく特徴抽出法では，訓練データに出現しない未知のキーワードが入力として与えられた場合，この入力に対して特徴を用意することができない．そのため特に訓練データが少量の場合，この問題が多く発生することになり，生成された分類器の予測精度が低下する問題があった．一方で特異値分解を用いて得られた行動トピックベクトルは，訓練データに出現しないキーワードについても用意することが可能であるため，この未知語問題を解決するものと考えられる．なお，本節で述べる次元削減によって得られる特徴を SVD features と呼ぶ．

クリックログを用いた行動トピックベクトル： s_d

行動トピックベクトルの計算の流れを図 3 に示した．まずクリックログのクエリを単語 v に分割し，クリック URL 情報からドメイン名 u を抽出する．その後，ドメイン名のクリック頻度を集計し，上位 n 件のドメイン名とその頻度のペアだけを抽出する．本研究では $n = 20$ と設定した．単語 v に対応する各ドメイン名情報について，(単語 v，ドメイン名 u，頻度) の 3 つ組要素を抽出し $|V| \times |U|$ の行列 S を作成する．このとき集合 U は，3 つ組要素のうち，ドメイン名に関する集合を表している．行列 S は行が単語 v に対応し，列がドメイン名 u に対応する．行列 S をこのように設定すると，i 行 j 列の要素は単語 v_i，ドメイン名情報 u_j の頻度情報 n_{ij} に該当する．

次に行列の次元削減方法について説明する．得られた行列 S において，ドメイン名集合 U の次元を任意の次元に削減することで，似たような概念の単語同士をクラスタリングすることができる．そこで本研究では特異値分解 (SVD) を行って $|V|$ 次元を予め定めた K 次元に変換する $|V| \times K$ の行列を生成し，各行の行ベクトル $n_i = \{n_{i1}, n_{i2}, \cdots, n_{i|U|}\}$ をキーワード w の特徴ベクトルとした．

ここで，特異値分解について説明する．特異値分解は，$|V| \times |U|$ の行列 X を $X = A\Sigma B^T$ の形に分解するものである [13]．この時，A は $|V| \times |V|$ の直交行列，B は

$|U| \times |U|$ の直交行列，Σ は $|V| \times |U|$ で対角成分が非負の対角行列である．なお，Σ を特異値行列，その非負の対角要素を特異値という．この際，特異値行列 Σ を特異値の降順に k 個を選択した場合の近似行列は，$X_k = A_k \Sigma_k B_k^T$ で得られる．この際，Σ_k は Σ における特異値を降順に k 個だけ残した $k \times k$ の行列を表し，A_k と B_k はそれぞれ A，B における最初の k 列からなる行列を表す．ここで

$$\min_{Y|rank(Y)=k} = \|X - Y\|_F = \|X - X_k\|_F \qquad (4)$$

が成立することが知られている [13]．すなわち特異値分解は，階数 k においてフロベニウスノルムの意味で元の行列を最良近似するような行列分解を実現している．このように分解された行列を用いると，行動トピック行列 Z は以下の式で表せる:

$$Z = A_k \Sigma_k. \qquad (5)$$

ここで行列 Z は $|V| \times k$ 行列で，各行が単語 v に対応する潜在ベクトルの情報を持つ．我々は URL ドメインに対するクリック情報はユーザの行動意図を表すと考え，本稿では各単語に対応する潜在ベクトルが行動意図を表していると見なし，キーワードに対応する潜在ベクトルを行動トピックベクトルとして新たな特徴として用いる．

4.4 地名特徴を用いた特徴抽出

地名情報に基づく特徴を抽出する手法を提案する．4.2 で説明した従来の出現キーワードに基づく特徴では，地名情報を区別して用いない．しかしながら一般に，地域によって行動タイプの分布が異なるため，地域ごとの特徴を持たないことで，行動タイプの予測が適切に行われないおそれがあった．そこで，本研究では地名情報に基づいて特徴を抽出する手法を提案する．なお，本節で説明した今地名情報に基づく特徴を今後 Geo features と呼ぶ．一般に，クエリ中に地名語が存在するとき，その行動タイプの出現分布は地名ごとに異なると考えられる．たとえば，オフィス街においては食事や移動といった行動タイプが特徴的に出現し，観光地では観光や買い物といった行動タイプがよく出現すると考えられる．評価実験に用いたデータの一部に対して地名語と共起するキーワードの行動タイプ分布を計算した結果を図 4 に示す．新宿駅では食事タイプや移動タイプの割合が多く，オフィス街やハブ駅という地域特徴が表れている．一方，おなじ 23 区でも代官山では半分が買い物タイプであり買い物の町という一般的な認識と一致する．東戸塚では前述の 2 つの都市と比較すると行動タイプの分布に偏りが少なく様々なタイプの行動が行われていると推測できる．このように地域によって行動タイプ分布は異なるため，地名情報を単語と区別して用意する特徴は，行動タイプ分類を行う上で有用であると考えられる．そこで，本研究では次の特徴を用いる．

地名情報に基づく行動タイプ出現度合：$P(t|geo)$

$$P(t|geo) = \frac{Count(t, geo) + 1}{\sum_{t' \in T} Count(t', geo) + |T|} \qquad (6)$$

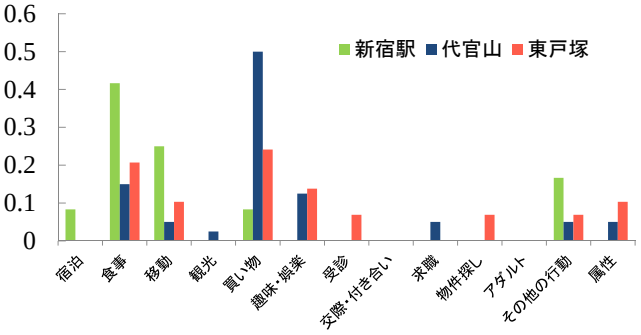


図 4　地域による行動タイプ分布の比較

本研究では，ゼロ頻度問題の回避のために，ラプラススムージング法 [12] を用いた．ここで $P(t|geo)$ の計算方法は Basic Feature における $P(t|w)$ と基本的に同じであるが，キーワード w そのものではなく，共起する地名語に対応する条件付き確率 $P(t|geo)$ を用いる点で異なる．

5 評価実験

5.1 使用データ

本実験では，商用検索エンジンのサーチログを用いた．データ取得の対象期間は 2012 年 10 月から 2013 年 9 月までの 1 年間である．また，クエリを構成するキーワードに対して人手で行動タイプを付与することで，行動タイプの正解ラベル付きデータを作成した．

正解ラベル付きデータの作成方法を説明する．まず前処理として，人手で用意した辞書を用いてアダルトワードが含まれるクエリを消去した．その後，地名辞書に含まれるキーワードを含み，2 キーワード以上から構成されるクエリを選択し，地名ごとにクエリ頻度が上位 50 位のものを抽出した．その後，ランダムに 15,610 クエリを選択したものをアノテーション対象とした．なお，今回は地名辞書として，日本国内の駅名と Wikipedia の見出し語情報を利用し，完全一致するものを地名語候補とした．一般に地名語には語義曖昧性の問題があるため，地名辞書と完全一致させるだけでは人名なども抽出してしまう．本稿では，地名語かどうかの判断を以下の判定 A) において作業者に依頼した．

行動タイプ判定のために，各クエリについて 3 名の作業者に以下の 3 つの判定を依頼した．判定 A)，B) では地理的意図を含むクエリを抽出するためのフィルタリングを行い，判定 C) で地理的意図アノテーションを行っている．

A) クエリに含まれる地名は，場所を示す単語として利用されているか．作業者の選択肢は {YES, NO, 判定不能 } のいずれかである．
B) A) で YES と判定された場合，そのクエリは地理的意図を含むか．作業者の選択肢は {YES, NO, 判定不能 } のいずれかである．
C) B) で YES と判定された場合，クエリを構成するキーワードごとに地名語・属性語・行動語のいずれかであ

表 3　行動タイプの事例数

行動タイプ ID	行動タイプ	事例数
1	宿泊	199
2	食事	1294
3	移動	1002
4	観光	515
5	買い物	703
6	趣味・娯楽	637
7	受診	284
8	交際・付き合い	60
9	求職	125
10	物件探し	217
11	アダルト	120
12	その他の行動	818

るかを判定する．また，行動語と判定された場合，その行動タイプを表 1 の 1-12 のうちから 1 つだけ選択する．この判定では地名語を 0，属性語を 13，行動タイプを 1 から 12 で判定することにしたため，作業者の選択肢は $\{0, 1, \cdots, 13\}$ のいずれかである．また，本研究ではクエリ内で共起するキーワードによって同一キーワードであっても違う行動タイプに判定することを許した．

判定 A)，B)，C) における 3 人の作業者間の回答一致度合いを確かめるために kappa 係数 [14] を計算した．作業者間の kappa 係数の平均値は作業者 a-b 間，b-c 間，c-a 間でそれぞれ 0.77，0.73，0.86 であり，高い一致がみられた．

評価データのうち，1 クエリあたりの地名語，属性語，行動語の平均キーワード数はそれぞれ 1.16，0.11，1.14 であった．また，各行動タイプを付与されたキーワード数を表 3 に示す．キーワード数が一番多いのは食事タイプの 1,294 で行動語全体の約 21% ほどであった．反対にキーワード数が一番少ないのは交際・付き合いタイプの 60 で全体の約 1% ほどであった．本研究では，3 人の作業者の回答が一致したクエリのみを実験に利用した．

5.2　未知語を含む事例に対する予備実験

行動タイプ分類問題の特徴を調べるために，単純な条件付き確率に基づくルールベースの手法を用いた予備実験を行った．ルールベースの手法は，キーワード w が与えられたときの行動タイプ t の条件付き確率 $P(t|w)$ が最大となる

$$\hat{t} = \underset{t}{\operatorname{argmax}} P(t|w) \tag{7}$$

を予測行動タイプとするものとした．なお，上記手法は訓練データに出現するキーワード (以後，既知語という) に対しては計算可能だが，出現しないキーワード (以後，未知語という) に対しては，計算できない．そこで，未知語を分類する際は，正解アノテーションデータの事例数が一番多い食事タイプに分類するというルールを設けた．ルールベース手法の既知語と未知語に対する精度評価を区別するため，交差検定の各試行においてテストデータに含まれるキーワードから既知語のみを判定対象とした評価 (Known) と未知語のみを対象にした評価 (Unknown) の 2 つの評価を行った．精度検証には 5 分割交差検定手法を用いた．評価指標には，Accuracy，Precision，Recall，F1 値を用いた．各クラスについて Precision，Recall，F1 値を計算し，12 クラスの平均を取るマクロ平均法で Precision，Recall，F1 値の平均値を計算した．

表 4　ルールベース手法の評価結果

	Accuracy	Precision	Recall	F1
Known	0.950	0.927	0.914	0.918
Unknown	0.205	0.017	0.083	0.028

表 4 は，実際にルールベース手法を用いて行動タイプ分類を行った結果である．Known においては Precision の平均値，Recall の平均値はともに 0.9 を超え，精度よく分類できたことがわかる．しかしながら，Unknown おいては，Precision の平均値は 0.017，Recall の平均値は 0.083 となり，ほとんど正確に分類できていないことが分かった．これにより，訓練データに出現するキーワード情報を明示的に使う特徴だけでは，未知語に対して適切に分類できないことを確認した．

Known においてルールベースの手法が高い精度を達成したことについて考察する．正解データを分析した結果によると，1 つの行動語が複数の行動タイプに分類される割合はおよそ 2% であり，1 つの行動語は 1 つの行動タイプのみにアノテーションされる例がほとんどであった．つまり，一度でも訓練データに出現するキーワードについては単純な条件付き確率に基づくルールベースの手法であっても，高い精度で分類できたものと考えられる．しかしながら，テストデータに含まれる未知語の割合は，5 分割したものの平均で 49.5% でおよそ半分を占めるため，未知語を無視することは全体の精度低下にもつながると考えられる．これらの理由により，行動タイプ分類において解くべき課題となるのは，訓練データに含まれない未知語に対する予測を高精度に達成することであることがわかった．

5.3　実験　各特徴を用いた際の分類精度比較

4 章で述べた 3 種類の特徴を利用した予測モデルをそれぞれ生成し，分類精度を比較することで，提案した特徴抽出手法が行動タイプ分類に有効であることを検証する．比較手法は Basic Features のみを用いるもの (Basic Features) と，Basic Features と Geo Features を用いるもの (BF + Geo Features)，Basic Features と SVD Features を用いるもの (BF + SVD Features)，3 種類全ての特徴を用いるもの (All Features) の 4 種類を用意した．精度検証には 5 分割交差検定手法を用いた．

分類器には，SVM の線形カーネル (SVM linear)[15]，RBF カーネル (SVM RBF)[15] と決定木 (Tree)，Random Forest (Forest)[16]，ロジスティック回帰モデル

(Logistic)[17] を用いた．実装には scikit-learn v.0.14[3] を利用し，分類器のパラメータの選択は，5 分割交差検定の各試行における訓練データに対して更に 3 分割交差検定を行い，グリッドサーチによって Accuracy が最大のものを選択した．それぞれ探索したパラメータとその範囲を説明する．SVM linear におけるペナルティパラメータ C は，$C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$ とした．SVM RBF におけるペナルティパラメータ C 及びカーネル係数 γ は，それぞれ $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$，$\gamma \in \{0.01, 0.001\}$ とした．Forest で用いる木の数 n_{est} は，$n_{est} \in \{5, 10, 20\}$，用いる特徴数 max_{feat} は，n_{feat} を全特徴数とすると $max_{feat} \in \{\sqrt{n_{feat}}\ ,\log_2 n_{feat}\}$ とした．Logistic における正規化項の逆数 C' は $C' \in \{0.01, 0.1, 1, 10, 100\}$ とした．評価指標は，クラスごとに F1 値を計算し，12 クラスのマクロ平均をとったものと Accuracy を利用した．また，実験評価対象となるキーワード群の種類は 5.2 と同様に既知語のみ (Known)，未知語のみ (Unknown) の 2 種類とした．

5.4 結果と考察

評価結果を表 5,6 にまとめた．表 5 は Accuracy，表 6 は F1 値の結果である．各項目は対象キーワード Known，Unknown に対して，各分類器，各特徴を用いて行動タイプを分類したときの評価指標値を示している．また，Best は，分類器の中で一番 F1 値が高かった値を示している．

対象キーワードを Known としたときの結果を述べる．Best の値で比較すると，提案手法を用いる BF+Geo Feature，BF+SVD Features，All Features の 3 手法が，既存手法である Basic Features と同程度の値を示した．5.2 で述べたとおり，一度でも訓練データに出現するキーワードについては単純な条件付き確率に基づくルールベースの手法であっても，Accuracy が 0.950，F1 値が 0.918 という結果が得られている．そのため，提案手法によって得られた特徴が，Known の行動タイプ分類精度向上の意味で条件付き確率以上の情報量を持っていないと考えられる．

対象キーワードを Unknown としたときの結果を説明する．Best の値で比較したとき，BF+SVD Features は既存手法の Basic Features と同程度の Accuracy と F1 値であった．そこで，分類器ごとに結果を見てみると，SVM linear と Logistic における結果において提案手法の BF+SVD Features が Basic Features よりも高い Accuracy および F1 値を示している．これより，計算時にクリックログデータを利用することで，適切な行動トピックベクトルを計算し，未知語に対して行動タイプ分類する上で適切な情報を与えていると考えられる．しかしながら，SVM Linear と Logistic 以外の分類器において評価指標の値が向上しておらず，この理由の考察は今後の課題としたい．

提案手法である BF+Geo Feature と All Features は既存手法の Basic Features より Accuracy，F1 値がともに高いという結果が得られた．BF+Geo Features の精度が Basic Features に比べて高い値を示した理由はキーワードと共起する地名語の行動タイプ分布を特徴として用いることで，キーワード自体が未知語であっても共起する地名語が訓練データに出現していれば，特徴を計算できたためと考えられる．Geo Features と SVD Features を合わせて追加した All Features においても Unknown の Accuracy において BF+Geo Feature, BF+SVD Feature に比べて僅かに高い精度を示しており，これにより 2 つの提案手法によって抽出された特徴量の情報量が異なることが示唆された．これらの結果により，提案手法に基づく特徴を用いることにより，行動タイプ分類において課題となっていた未知語問題を部分的に解決し，従来手法に比べて高精度な予測が可能になったといえる．

表 5　特徴ごとの Accuracy の比較

対象キーワード	分類器	Basic Features	BF + Geo Features	BF + SVD Features	All Features
Known	SVM linear	0.950	0.952	0.950	0.952
	SVM RBF	0.917	0.884	0.918	0.886
	Tree	0.949	0.949	0.948	0.939
	Forest	0.937	0.919	0.855	0.813
	Logistic	0.950	0.950	0.951	0.947
	Best	0.950	0.952	0.951	0.952
Unknown	SVM linear	0.188	0.215	0.205	0.227
	SVM RBF	0.313	0.336	0.314	0.336
	Tree	0.012	0.026	0.038	0.054
	Forest	0.178	0.228	0.212	0.213
	Logistic	0.219	0.246	0.238	0.260
	Best	0.313	0.336	0.314	0.336

表 6　特徴ごとの F1 値の比較

対象キーワード	分類器	Basic Features	BF + Geo Features	BF + SVD Features	All Features
Known	SVM linear	0.916	0.922	0.917	0.922
	SVM RBF	0.854	0.818	0.856	0.822
	Tree	0.915	0.914	0.903	0.895
	Forest	0.886	0.834	0.713	0.722
	Logistic	0.917	0.916	0.918	0.911
	Best	0.917	0.922	0.918	0.922
Unknown	SVM linear	0.051	0.076	0.058	0.082
	SVM RBF	0.138	0.173	0.140	0.177
	Tree	0.003	0.008	0.006	0.017
	Forest	0.052	0.088	0.054	0.079
	Logistic	0.082	0.110	0.102	0.131
	Best	0.138	0.173	0.140	0.177

6　おわりに

本研究では，ユーザが意図する実世界の行動に合わせた検索結果を提示するために，検索エンジンに入力された地理的意図を含むクエリに対して地名語，行動語，属性語を付与する問題において，特に行動語を事前に設定された行動タイプに分類する課題に取り組んだ．既存研究でも用いられているマルチクラス分類として定式化をし，訓練データに出現しない未知語に対しても適切に特徴を抽出する手法を 2 つ提案した．

1 つ目の提案手法は，クエリを構成するキーワードと，クリックされた URL のドメイン情報の関連性がユーザの行動意図を反映しているという仮定に基づいてて，キーワード-ドメイン名の頻度行列を行列分解することによっ

[3] http://scikit-learn.org/stable/

て得た行動トピックベクトルを特徴として用いる方法である．

2 つ目の提案手法は，キーワードと共起する地名語の行動タイプ分布を特徴として用いる方法である．この手法はキーワードと共起する地名語によって行動タイプの分布が異なるという仮説に基づいてており，この仮説が成立することはアノテーションデータを用いた予備分析で確認された．

商用検索エンジンのクエリログに対するアノテーションデータを利用して，上記 2 つの提案手法の評価実験を行った結果，提案手法を特徴として用いることで，特に訓練データに含まれない未知語に対して高精度に行動タイプ分類できることを確認し，未知語に対する提案手法の有効性を検証した．

今後は，地名語・属性語・行動語の分類と行動タイプの分類を同時に行うことにより，より現実に近い条件下で高い精度を予測モデルの生成を目指す．

参考文献

[1] Z. Nie, J. Wen, and W. Ma. Object-level vertical search. In *In Proc. of CIDR'07*, pp. 235–246, 2007.

[2] X. Yi, H. Raghavan, and C. Leggetter. Discovering users' specific geo intention in web search. In *Proc. of WWW'09*, pp. 481–490, 2009.

[3] M. J Welch and J. Cho. Automatically identifying localizable queries. In *Proc. of SIGIR'08*, pp. 507–514, 2008.

[4] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36, pp. 3–10, 2002.

[5] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proc. of WWW'04*, pp. 13–19, 2004.

[6] 廣嶋伸章, 戸田浩之, 松浦由美子, 片岡良治. 概念ベースに基づく web 検索のクエリタイプ判定手法とその評価. 情報処理学会論文誌. データベース, Vol. 3, No. 3, pp. 33–45, 2010.

[7] T. Lin, P. Pantel, M. Gamon, A. Kannan, and A. Fuxman. Active objects: Actions for entity-centric search. In *Proc. of WWW'12*, pp. 589–598, 2012.

[8] R. Jones, W. V. Zhang, B. Rey, P. Jhala, and E. Stipp. Geographic intention and modification in web search. *IJGIS*, Vol. 22, No. 3, pp. 229–246, 2008.

[9] 平野徹, 松尾義博, 菊井玄一郎. 地理的距離を用いた地名の曖昧性解消. 第 70 回情報処理学会全国大会, 2008.

[10] H. Li, R. K. Srihari, C. Niu, and W. Li. Location normalization for information extraction. In *Proc. of COLING'02*, pp. 1–7, 2002.

[11] M. Paşca. Organizing and searching the world wide web of facts–Step two: Harnessing the wisdom of the crowds. In *Proc. of WWW'07*, pp. 101–110, 2007.

[12] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. of ACL'96*, pp. 310–318, 1996.

[13] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, Vol. 41, No. 6, pp. 391–407, 1990.

[14] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational linguistics*, Vol. 22, No. 2, pp. 249–254, 1996.

[15] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[16] L. Breiman. Random forests. *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[17] H. Yu, F. Huang, and C. Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, Vol. 85, No. 1-2, pp. 41–75, 2011.