ソーシャルメディアデータからの
体験マイニングに関する研究

倉島 健

要旨

本論文では，ソーシャルメディアに存在する人々の都市での体験に関する記述を構造化情報として蓄積し，さらに，大量の体験データに埋もれている有用な傾向を発見し活用するまでを扱う体験マイニング手法について提案を行っている．

近年，ブログに代表されるソーシャルメディアの普及が著しい．現在のところ，ソーシャルメディアは都市を実際に訪れた人の体験や，実体験に基づく主観的な記述を頻度高く含むという他メディアとは異なる重要な特徴を持っている．しかし，自然言語形式で書かれた非構造データであり，その意味や属性を掴みづらく，また，日々，膨大な量のコンテンツが生成され，発信されているといった理由から，これらの情報は充分に利活用されていない現状がある．これまで，ソーシャルメディアコンテンツを解析対象として“情報が新鮮である”観点に着目した話題分析技術や，特定の商品・サービスの“良い/悪い”に着目した評判分析技術はさかんに行われてきたが，個々人のうちで直接的に得られた都市における体験を扱うことを指向し，“時間的・空間的な背景に紐づけられた都市における人間行動”を軸に分析をした研究はほとんど存在しない．

そこで，本論文では，ソーシャルメディアに存在する人々の体験を構造化情報として蓄積し，そこから有用な傾向を獲得して個人や企業の意志決定に利活用するまでの処理プロセスを支援するための体験マイニング技術に関する提案を行っている．具体的には，本論文では「ブログ文書からの体験情報抽出と構造化」,「人々の体験に関する有用な相関ルール（association rule，アソシエーションルール）の選択手法」,「写真共有サイトのジオタグ情報を利用したトラベルルート推薦」に関する研究を行い，提案した手法の評価を行っている．

本論文は全 6 章から構成されている．その概要は以下の通りである．

第 1 章は序論であり，本論文の研究の背景，本論文の研究を行うに至った動機，および，本論文の研究の全体の概要について述べている．

第 2 章では，第 3〜5 章において説明するそれぞれの研究課題に特に深く関連する従来研究について整理し，本研究との位置づけを議論している．

第 3 章では，自然言語で記述された非構造データであるブログ文書から個人の体験情報を，時間，空間，行動属性から成る構造化情報として抽出する手法を提案し，これについて論じている．フィルモアの格文法解析に加え，時間的・空間的要因によって出現が規定されることを手がかりに，人間の行動内容を示す表現を適切に選択し，抽出する特徴がある．実際のブログ文書からの体験情報抽出精度を評価する実験により，提案法の有効性を示している．また，提案手法に基づき，ユーザが体験情報集合を柔軟に検索・要約することのできる体験ブログマップ（Blog Map of Experiences）を実現している．

第 4 章では，構造化した人々の体験情報集合から有用な知識を抽出する手法を提案している．具体的には，アソシエーションルール抽出手法によって得られた大量のルールの中から “ある特定の状況（時空条件）において特徴的に出現する行動の発見”，“人々に認知されている行動と都市で実際に人々がしている行動の差異発見” につながるものをそれぞれ選択する手法を述べている．評価実験では，時空間条件によって，また，出現するメディアの種類・性質によって行動の出現傾向がどの程度変化するかを評価する提案手法により，効率良く有用な知識が得られることを示している．

第 5 章では，蓄積した人々の体験情報を未来の意志決定に利活用する試みとして，旅行者のトラベルルートを自動拡張する推薦手法を提案している．評価実験では，過去の場所訪問履歴をもとにユーザが好む場所の潜在的な特徴を推定すること，また，ユーザが置かれた状況（現在地，空き時間）を考慮して訪問先を絞り込むことで，提案法が旅行者の訪問先を高精度に予測できることを示している．

第 6 章では，本研究で得られた研究成果をまとめ，さらに今後の展開について述べている．

目 次

図目次

表目次

第1章　序論

1.1　背景

現在のところ，ブログ，Twitter[1]，Flickr[2]，Facebook [3] などのソーシャルメディアはインターネット上の他メディアとは異なる重要な特徴を持っている．それは，都市を実際に訪れた人の実体験や，実体験に基づく主観的な記述を頻度高く含むことである．具体的には，2つの観点においてソーシャルメディアに存在する人々の体験情報は価値があると考える．1つ目は，都市に生きる人々の多様な行動内容を知るための重要な情報源であることである．これまでも旅行ガイドブックやインターネット上の観光サイトなどの情報源に目を通すことで，都市における定番，有名な体験を知ることが出来たが，そこには反映されていない，いわゆる“ロングテール”な体験を知ることは出来なかった．2つ目は，都市に生きる人々の実態が直接的に反映されていることである．従来，このような人間の動きに関する情報は，新聞やテレビなどのメディアから間接的ともいえる方法でしか得ることができなかった．これらのメディアの発信情報には広告的な意図を持ったものも少なからず含まれるため，人々の実態を正しく把握できるとは限らない．また，メディアや企業に情報提供をする調査会社は，大規模なアンケート調査を実施して都市の生活者の実態を把握していたが，主に，被験者負担の観点から継続的な調査が困難であった．

このような理由から，ソーシャルメディアに発信される都市における人々の体験情報に注目が集まっているが，観光サイトや店舗のホームページなどのコンテ

[1] https://twitter.com/
[2] https://www.flickr.com/
[3] https://www.facebook.com/

ンツに比べて品質が保証されず，玉石混淆であった．また，日々，新たな情報が生成，発信され続け，扱うデータ量が膨大であるという理由から，その利活用が進まない現状がある．現在，Web 上のコンテンツを広く網羅している Web 検索エンジンを利用した情報収集が主流となっている．Web 検索エンジンは，ユーザの入力した検索クエリに基づき膨大なページやコンテンツを何らかの観点でランキングし提示する．情報が情報として多くの人々に広く再利用されるためには，Web 検索エンジンにインデックスされ，かつ，上位にランキングされる必要があるが，ソーシャルメディア情報は玉石混淆であるという理由から，検索エンジン下位にランキングされるか，インデックスされることもなく再利用の機会を失っている現状がある．

これまで，自然言語処理分野やデータマイニング分野を中心に，玉石混淆なソーシャルメディアから “玉” を拾うための研究が非常にさかんに行われてきた．しかし，“情報が良質である” ことを評価することを目的とした情報検索技術や，“情報が新鮮である” という観点に着目した話題分析技術，特定の商品や人物の “良い/悪い” に着目した評判分析技術がほとんであり，“誰が，いつ，どこで，どのような行動をし，その結果としてどのような知見が得られたのか” という都市における個々人の体験を中心とした分析を行った研究はほとんど行われていない．本研究の目的は，ソーシャルメディアに存在する人々の体験情報を理解，分析，活用するために役立つ情報を抽出する技術を開発し，個人の行動計画や，企業活動で生じる意志決定を容易にすることにある．

1.2　本研究の概要

本研究で取り組む研究課題を図 1.1 に示す．ソーシャルメディアの 1 つであるブログ上には都市の生活者である個人の体験内容が頻度高く記述されている特徴があるが，自然言語で記述された非構造データであるため，その意味や属性を掴みづらい．また，日々，膨大な量のコンテンツが発信されているといった理由から，これら人間の体験情報は充分に利活用されていない現状がある．たとえば，ブロ

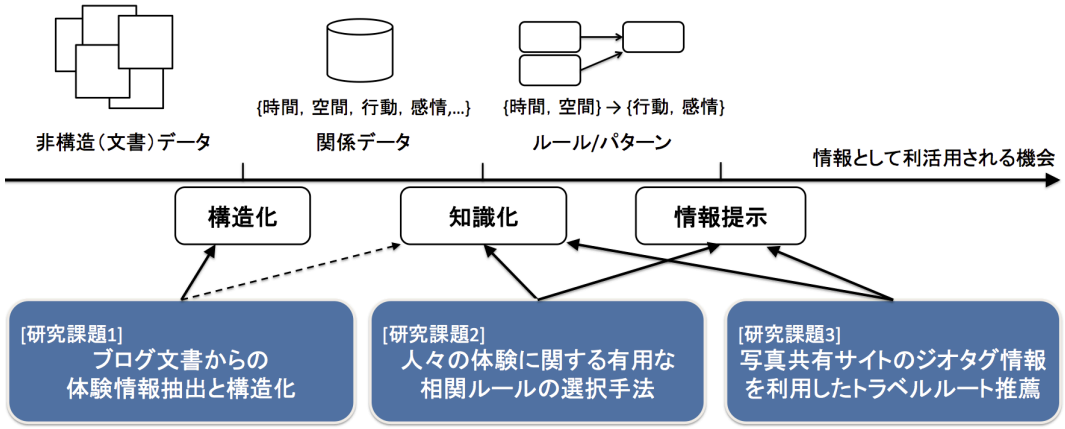


図 1.1: 本研究で取り組む研究課題

グホスティングサービスが提供するテキスト検索機能で“清水寺”などのキーワードで検索した結果は膨大であり，さらには，そのすべてが体験記とは限らない．すべてのブログ文書に目を通すことは困難である上，その一部に目を通しただけでは地域の人々の行動に関する偏った見解に陥ってしまいかねない．そこで，本研究では 1 つ目の研究課題として，自然言語で記述された非構造なブログデータを対象とし，人間の体験を表現する最小構成要素として，時間属性，場所属性，行動属性の組合せ情報の抽出に取り組む．提案手法においては，係り受け解析により，動詞，名詞句，格助詞の組合せ情報を広く抽出した後，フィルモアの格文法解析，動詞の意味解析をすることで，行動内容を示す表現を順に取捨選択していく．さらに，人間の行動は時間的・空間的要因によって規定されているという点に着目し，全ブログデータ中で時間，空間，行動属性の共起しやすい組合せパターンをアソシエーションルールとして抽出し，抽出処理に反映させることで，精度高く体験情報を抽出する．提案手法に基づき，体験情報を構成する属性を指定することで，柔軟に人々の体験情報を検索可能な体験ブログマップ（Blog Map of Experiences）を開発した．提案システムは，データベースに格納された体験情報への構造化されたアクセス方法を提供するものである．

2 つ目の研究課題は，蓄積した大量の体験情報から，人々の行動傾向に関する有用な知識を獲得するための仕組みを構築することである．日々，変化する人々の行動傾向に関して適切な仮説を立てることは決して容易なことではなく，仮説検証型のアプローチでは多くの重要な傾向を見落としてしまう可能性があった．その一方で，データ傾向を説明するパターンやルールを網羅的に自動抽出する方法も考えられるが，抽出されたパターンやルール自体の数が膨大である，さらには，そのほとんどがユーザにとって既知のものである，目的にそぐわないものであるといった問題があった．一般に，ある情報の価値は，その情報の利用者の背景知識や目的，利用シーンによって異なる．そこで，第 2 の研究テーマでは，都市のトレンドに関心を持つ調査会社やマーケッターをユーザとして想定し，ある時間に，ある場所を人々が訪れる理由や目的の説明となる情報として，“ある特定の状況（時間，空間）において，人々が特徴的にしている行動” を表現するルールを “興味深い知識” として抽出する．GPS 機能を標準搭載したモバイル端末やカーナビゲーションシステムの普及などを背景に，比較的容易に人々の移動や集中に関する情報を把握できるようになってきたが，なぜ，あるいは，何をするために訪れたかの説明となる情報は，アンケート調査などを実施して都市の生活者に聞くしかほかに方法がなかった．提案手法においては，人間の体験を構成する属性の中でも特に 5 属性（時間，空間，動作，対象，感情）をブログから抽出する．そのように生成した構造化データから，一般性が低いものも含めたアソシエーションルールを幅広く抽出した後，時間的・空間的な条件づけによって行動の出現傾向がどの程度変化するかを評価することで，ある場所を人々が訪れる理由，目的の説明につながるルールのみを抽出する．

また，第 2 の研究テーマでは，旅行ガイドブックや Web 検索エンジンなどのメディアを用いて地域情報を収集したことのあるユーザを想定し，一般のメディアへの露出度が高く人々に認知されている行動と，都市で実際に人々がしている行動の差異発見につながるルールを “興味深い知識” として抽出する手法も提案する．提案手法では，Web 検索エンジンの検索結果ページでの出現頻度や出現位置を分析することで情報の認知度を推定し，さらに，ソーシャルメディアにおける出現

傾向と比較することで，ソーシャルメディアに特徴的な傾向を発見する．評価実験においては，過去に観光したことのある都市に関する新たな体験を発掘するタスクにおいて提案手法の有効性を示す．

3つ目の研究課題は，過去に都市を訪れた人々の体験情報を利用しユーザ行動の自動拡張を行う仕組みを構築し，ユーザの地域情報検索の入り口を支援することである．Web検索エンジンに代表される，ユーザの能動的な検索クエリ入力を前提としたシステムの場合，ユーザが欲する情報が曖昧なほど，検索クエリの言語化が困難である．さらに，実世界に置かれたユーザは，現在地，現在時間，空き時間など，考慮すべき要因も多く，検索クエリが複雑化しがちである．たとえば，“予期せず空き時間ができたので暇つぶしをしたい”といった場合など，ユーザが求める情報，欲する情報が，特定の場所や行動内容という形で顕在化しているケースは多くない．提案手法においては，ユーザが過去にどの場所を訪れたかを示す移動履歴をもとにユーザがどのような特徴を持つ場所を好むかを分析し，さらに，ユーザが置かれた状況も考慮しながら，次に行く確率の高い場所を予測する．また，ユーザ自身の空き時間を入力とすることで，単一の場所としてではなく，より具体的な旅行計画（トラベルルート）としてユーザに情報提示を行う．提案手法は，他人の過去の体験を自身の意志決定に反映する処理を自動化するものであり，ユーザは，自分が情報収集（地域情報検索）を繰り返すことで次第に明確化していく様々な選択肢に，自らの情報収集なくして気づくことができる．

第2章　関連研究

2.1　体験情報の定義

本研究では，個々人のうちで直接的に感得された都市における体験を扱う．固有な出来事としての体験を自然な形で表現するために，以下の情報で構造化する．

- **動作主**: 行動をした動作主の属性
- **状況**: 行動をした**時間**と**空間**
- **行動**: 人間が行う**動作**とその**対象**
- **主観**: 動作をともなう対象に対する**評価**と，行動の結果，動作主が抱いた**感情**

動作主，時空的な文脈（状況），行動内容，そこから得られた知見（主観）と合わせ，1 つの固有な体験を表現する．それぞれの要素はさらに細分化でき，実際には動作主，時間，空間，動作，対象，評価，感情の 7 属性で人間の体験を表現する．たとえば，ある動作主 A によって 2013 年 11 月 1 日に投稿された “嵐山で紅葉を見ましたが，きれいで感動しました．” と書かれた文は｛動作主，時間，空間，動作，対象，評価，感情｝＝｛A，2013 年 11 月 1 日，嵐山，見る，紅葉，きれい，感動｝のように表現できる．

ある体験をした都市におけるランドマーク，建物，寺社，店舗，公共施設などを示す地名や位置情報（緯度・経度）が空間属性値であり，ある体験をした時刻や日付が時間属性値である．これらの状況属性は，個々の体験の文脈，背景を保存する．本来，連続的な空間・時間領域に対して人間の体験の紐付けを行うほうが自然ではあるが，その粒度を統一的に定めることは難しい．本研究では，扱うソー

シャルメディアデータの性質に応じて時間，空間属性値の粒度を決定している．“食べる”，“見る”，“買う” など，人間の動作内容を扱うのが動作属性であり，人間の動作が作用する対象となる都市の具体物が対象属性である．具体的には，自然物，自然現象，食物，動植物，生産物，実世界イベントを対象属性の値として扱う．たとえば，“紅葉を見る” という行動を示す言語表現は，動作=“見る”，対象=“紅葉” として表現される．ここで，人間の行動を示す表現の中でも特に，都市空間に対する紐付けが可能なもののみを扱う．たとえば，“テレビを見る”，“ブログを見る” や “ニュースを見る” などは都市における特定の空間に紐づけることが困難な行動情報であり，都市の体験記とは異なる文脈で出現し分析のノイズとなるため本研究では扱わないこととする．

本論文では主に，動作主属性と評価属性を除く，5 属性について考える．まず，第 1 の研究課題において，これらの中の時間属性，空間属性，行動（動作，対象）属性の値をブログ記事から抽出し，第 2 の研究においては，さらに，感情属性の値も抽出する．評価属性の抽出は，既存の評判情報抽出技術の主要課題であるため本研究のスコープから外す [1, 2, 3, 4, 5]．また，体験をした人物を一意に識別するためのユーザ ID を動作主属性としている．性別，居住地など，動作主がどんな人物か，を示す属性情報は，通常テキスト中に明示されないため，推定する研究も行われている [6]．将来的にはこれらの分野の知見を参考にして扱っていきたい．

2.2　ソーシャルメディアデータのマイニング技術

2.2.1　コンテンツ分析

評判情報抽出: ソーシャルメディアの中でも特に，自然言語で記述されたブログや Twitter の本文情報を対象とした研究が非常にさかんに行われてきた．評判情報抽出技術は，商品，サービス，人や組織などに対する人々の評価，評判を知ることができるというソーシャルメディアの特徴に着目した技術である．評判情報抽出技術の主要な技術課題は，{評価対象，属性，評価} という 3 つ組を自然文から抽

出することである．立石らは，評価対象，属性，評価に関する共起パターンを介して，属性表現と評価表現をブートストラップ的に抽出する手法を提案した [1]．小林らは，領域に固有な属性，評価表現を共起パターンに基づいて効率的に収集する手法を提案している [2]．Liu らは，アソシエーションルール抽出技術を利用し，3 つ組のうちの属性要素を抽出するための言語パターンを自動生成する手法を提案している [3, 4, 5]．また，“A よりも B のほうがデザインが良い” といった比較評価表現に着目し，{評価対象，比較対象，属性，評価} という 4 つ組から成る比較評価情報として抽出する研究も存在する [7, 8, 9]．これらの評判情報抽出技術が，商品，サービス，人物や組織などの “評価対象” を軸とした情報抽出を試みているのに対して，本研究は，“いつ（時間）”，“どこで（空間）”，“何をする（動作，動作の対象）” という人間の行動内容を軸とした情報抽出を行っている．

人間の主観に関する表現は，“評価” と “感情” とに大別できる．“良い” や “悪い” といった人間の “評価” は，ある対象に対する主観的な価値付けを示す表現であるのに対して，“嬉しい” や “悲しい” は，動作主体の心的状態を示す表現である．体験情報を構成する 1 要素である感情属性の抽出において，熊本らは，Plutchik ら [10] の感情カテゴリを参考にし，“悲しい-嬉しい”，“怒る-喜ぶ”，“悲しみ-怒り”，“受容-嫌悪” という 4 軸に対する評価値を含む感情表現の抽出手法を提案した [11]．また，福原らは，新聞記事から人手で感情語を収集し，感情辞書を作成している [12]．本研究においては，感情属性情報は，動作主がある体験を失敗だったと考えているか，それとも成功だったと考えているかを知るための鍵となる表現だと考えている．そのため，第 2 の研究課題において，感情表現抽出分野の知見を利用して感情属性値を抽出する．

経験情報抽出: 本論文で提案する体験マイニングを一般公開した後に発表された研究ではあるが，関連が深い研究として乾らの経験マイニングがある [13, 14]．経験マイニングは，個人の経験情報を {トピック，経験主，事態タイプ，事実性情報，事態表現} の構造化情報として抽出することを目指している．トピックは，商品，サービスなど，どの利用物に関する経験かを示す情報であり，経験主は経験の主体であるとしている．事態タイプは，経験の核となる事態表現の種類であり，ポ

ジティブ/ネガティブな出来事，状態，動作（食べる，見る，買う）などに分類される．事実性情報は，その事態が実際に起こったことなのか，可能性に言及しただけなのか，を表す情報である．

経験マイニング研究は，状態，性質から関連する人間の動作まで，トピック（商品，場所，サービスなど）に関連する経験を広く扱うものとして定義されているが，主に商品，サービスの分析を目的としているため，評判分析技術の発展的な研究領域ととらえることができる．それに対し，本研究で扱う体験マイニングの主な関心は，ソーシャルメディアを通して透けてくる都市における人々の生活にある．そのため，都市における個人の体験のみに分析を限定し，それを自然な形で表現する構造として体験情報を定義している．たとえば，行動内容と行動をした状況（時間，空間）との組合せ情報として人間の体験をとらえることで，ある時間，ある空間で切り取った都市の一側面を人々の行動内容から描くことが可能になる．さらに，食べる，買う，見るといった動作の種類ではなく，“何を” 食べたのか，“何を” 見たのかといった行動内容を抽出することにより，より具体的に都市の生活者の姿を描き出そうとしている．特定の利用物ではなく，一連の体験に対する評価として感情表現に注目している点も体験マイニングの特徴である．

また，経験マイニングが自然言語処理技術に基づく情報の構造化のみを研究領域としているのに対して，本研究テーマである体験マイニングは，構造化だけではなく，構造化データから有用な傾向やパターンを抽出し提示する処理までの領域を広く扱う．提案する体験マイニングは，自然言語処理技術とデータマイニング技術が融合した研究領域であり，主観的，断片的な個々の体験の蓄積データから，有用な傾向を知識として発見するまでのプロセスを扱う．

なお，提案手法の情報抽出処理としての特徴は，教師データを必要としない点，体験内容が記述される際の時空的な文脈を活用する点にある．関連研究 [13, 14, 15] は，どの表現が何の属性を示すものかをラベル付けする必要があり，実現のためのコストが高い．提案手法は，扱う情報が人間の体験であることに着目し，言語表現の出現傾向を時間的・空間的に分析することで正しく行動語を抽出する．

話題・トピック分析: 話題抽出技術は，リアルタイムに人々の反応を知ることができるというソーシャルメディアの特徴に注目した技術であり，ニューストピックや商品などに人々の関心が集中している状態を自動検出する技術である．Kleinberg らは，時間軸上で集中的にドキュメントが到着する“バースト状態”を発見する手法を示した [16]．藤木らは，Klenberg らの手法を拡張し，ソーシャルメディアにおいて文書の到着間隔が時間変化する点について対処する手法を提案している [17]．また，リアルタイムに話題語を提示するサービスとして kizasi.jp[1]，blogWatcher[18, 19, 17] や BLOGRANGER[20, 21] などがある．最近では，誰が，いつ，どの話題に言及したかを手掛かりとして，メディアから人へ，人から人への情報伝搬を引き起こす潜在的な情報伝搬ネットワークを推定する研究も行われている [22, 23, 24, 25]．前述の通り，これら話題抽出のための研究やサービスは情報が“新鮮である”という観点でソーシャルメディアをするためのものである．

2.2.2 リンク構造分析

ブログ間のリンク構造に着目したコミュニティ分析手法がこれまでにいくつか提案されてきた．Kumar らはブログのリンク構造から，ブログコミュニティの成長と変化を分析した [26, 27]．Bar-ilan らは，ブログ投稿とリンクとの関係を統計的に分析した [28]．藤村らは，ブロガーのハブスコアとオーソリティースコアをリンク構造から推定する *EigenRumor* アルゴリズムを提案している [29]．中島らもブログのリンク構造を分析してコミュニティ内で重要な役割を果たすブロガーを特定する手法を提案した [30]．これらの研究はブログ間の関係性に着目したブログコミュニティ分析であり，ブログ記事の内容（コンテンツ）は扱っていない．

[1] http://kizasi.jp/

2.3　体験情報を活用したアプリケーション

ソーシャルメディアにおける人々の体験情報を用いて，情報推薦 [31, 32, 33, 34, 35, 36]，情報検索・コンテンツブラウジング [37, 38, 39, 40, 41, 42, 43]，自動アノテーション [44, 45, 46] などに関連して様々な試みが存在する．これらの研究が利用しているのは GPS 機能を搭載した端末から得られる人々の位置情報であるのに対して，本研究の第 1，第 2 の研究テーマは，自然言語で記述された行動内容に着目している．また，本研究の第 3 の研究テーマは，ソーシャルメディアに存在する人々の移動履歴情報を用いて都市のトラベルルート推薦を実現した点に新規性と技術的貢献がある．以降，それぞれの関連研究を紹介する．

情報推薦: Yahoo! TRAVEL[2] は，システムユーザの空き時間と，店舗やランドマークに対する過去の評点情報をもとに旅行先を推薦するサービスである．Horozov らも，協調フィルタリング技術を適用することでユーザの好みに合うレストランを推薦する手法を提案している [31]．推薦システムの性能は予測精度で評価するのが一般的であるため [47]，GPS 軌跡データを用いた行動予測技術も，本研究と関連が深い．Ashbrook らは，GPS 履歴をもとに，マルコフモデルで個人の行動モデルを生成した [32]．Krumm らは複数のドライバーの GPS 軌跡をもとに，ドライバーの行き先を予測する手法を提案した [33]．Zheng らは，107 名の被験者の GPS 軌跡データにグラフマイニング手法を適用し，人々が関心を寄せるランドマーク，及び，都市における典型的な移動パターンを抽出した [34]．本研究の第 3 の研究テーマにおいて，店舗・ランドマーク情報や，評点などの個人情報データベースがなくとも，写真共有サイト上のフォトグラファーの体験を利活用することで，地域情報の自動推薦が可能となることを示す．[35, 36] は，本研究と同様，ソーシャルメディアに存在する人々の時空情報を利用した研究である．しかし，都市を代表する旅行プランを自動生成することを目的としており，情報の個人化を行うことはできない．

[2] http://www.travel.yahoo.com/

情報検索・コンテンツブラウジング: 画像検索分野においては，時空情報は重要なメタデータとして利用されてきた．Kennedy らは，画像に付与されたテキストタグ情報，画像特徴量，そして時空情報に基づいて，ある場所の典型的な画像集合，あるいは多様な視点での画像集合を抽出する手法を提案している [37, 38]．地図インタフェース上で，位置情報が付与されたコンテンツやコンテンツに付与されたタグ情報を集約し可視化する手法がいくつか提案されている [39, 40, 41, 42, 43]．*World Explorer* は地図インタフェース上で，ある地域を代表するタグ情報を抽出し可視化するシステムである [39]．また，Crandall らは，全世界で撮影され画像情報を，一枚の世界地図に集約し提示する手法を提案した [41]．Popescu らは，位置情報が付与された画像集合と Wikipedia のデータを用いて，多言語対応の地名辞書を自動抽出する手法を示した [48]．Snavely らは，あるランドマークについて撮影された画像集合をつなぎ合わせ，3D ナビゲーションを可能とする *Photo Tourism* を提案している [42, 43]．

自動アノテーション: コンテンツの特徴と，それに付与された位置情報との関係性を分析し，コンテンツに自動で位置情報を付与する研究も存在する．Hays らは，位置情報が付与されていない画像と類似したシーンが存在する画像を抽出し，その類似画像に付与された位置情報から元の画像の位置情報を推定する手法を提案した [46]．Kalogerakis らは，撮影者の時空軌跡情報を事前知識として導入し，画像特徴量に基づく類似画像探索と組合せることで，ある画像が撮影された位置情報を高精度に推定する手法を示した [44]．Backstrom らは，Facebook におけるソーシャルネットワーク上で友人関係にある人々の居住地（住所）を解析し，特定のユーザの居住地（住所）を推定する手法を示した [49]．誰がどの店舗にいるかを示す Foursquare[3] のチェックインデータが，GPS 精度の低さ，都市部の店舗が密集しているといった理由から，正しく結びつけられていない点に注目し，経路情報，ユーザ情報，店舗に付与されたメタ情報を用いて，ユーザが実際にチェックインした店舗を推定する研究もある [50]．

[3] http://foursquare.com/

2.4 アソシエーションルール抽出

本研究に関する基本的事項として，アソシエーションルール抽出技術について説明する．アソシエーションルールはある商品 A を購入したらほかの商品 B も同時に購入する，というような観測データに潜む傾向を把握するための手法である．アソシエーションルール分析では，アイテム集合を一つのトランザクションとみなす．たとえば，スーパーマーケットでの買い物を例にすると，顧客の商品カートがトランザクションであり，商品カートの中に入っている商品（商品集合）がアイテム（アイテムセット）である．全顧客に関するトランザクションデータからアソシエーションルールを抽出する．アソシエーションルールは $X \Rightarrow Y$ の形式で表現され，X は条件部の，Y は結論部のアイテム集合である．条件部と結論部は，それぞれ複数のアイテムであってもよく，その場合は，もし条件部のすべてのアイテムがトランザクション中に現れれば，そのトランザクション中には結論部のすべてのアイテムが現れやすいことを示す．

アソシエーションルールの価値を決める指標は様々に提案されているが，最も典型的な指標が支持度（$support$）と確信度（$confidence$）と呼ばれる指標である．支持度はルールの一般性を評価する指標の 1 つであり，トランザクションデータで X と Y がどちらも出現する同時確率 $P(X, Y)$ である．テキストマイニング分野においては，全文書中で単語 X と単語 Y を同時に含む文書の割合として表される場合もある．確信度はルールの信頼性を評価する指標の 1 つであり，アソシエーションルール $X \Rightarrow Y$ の確信度は，トランザクションデータで X が出現したという条件の下で Y が出現する条件付き確率 $P(Y|X)$ である．ある一定値以上の支持度（最小支持度）と確信度（最小確信度）を持つアソシエーションルールを高速に抽出する手法が Agrawal らによって提案された *Apriori* アルゴリズムである [51, 52].

第3章　ブログ文書からの体験情報抽出と構造化

3.1　緒言

人間の行動は時間的・空間的要因によって規定されている．秋には多くの人が紅葉を見にいくし，地域特有の食べ物が存在すれば旅行者がそれを好んで食べる．ブログの普及により，地域を実際に体験した個人の情報発信が活発になり，さらには，記述された日時が記録されているというブログの特性によってこのような人間の動きに関する情報が得られるようになった．従来，このような人間の動きに関する情報は，地方紙や定期刊行物などのメディアを通して，間接的ともいえる方法でしか得ることができず，そのすべてを把握することは不可能であった．一般の人々が発信するブログから直接的に得られる個人の体験は，その地域を訪れようと考えている潜在的な訪問者や，地域の流行に興味のあるマーケッターにとって有用である．

現在，ある場所を実際に訪れて書いた体験記のような文書を，位置情報に基づいて一般のユーザが投稿するシステムが存在する．しかし，それらのシステムが広く普及しているとは言い難く，体験型ブログの多くはブログホスティングサービス上で，単に一個人の蓄積情報の 1 つとして扱われているのが現状である．人々の体験情報を網羅的に知るためには，ブログホスティングサービスが提供するキーワード検索機能を利用してブログ記事を検索し，それぞれの記事に目を通す必要がある．しかし，たとえば“清水寺”などの地名をキーワードとして検索した結果は大量であり，さらにはそのすべてが“清水寺”を実際に訪れて書いた体験記とは限らない．このように，ある地域に関して書かれた体験記ブログが Web 上に分散

して存在し，さらにはその量が膨大であるという現状においては，すべてのブログ記事に目を通すことは困難であるといえる．その一部に目を通しただけでは，都市における人々の行動傾向に関して偏った見解に陥ってしまいかねない．

本節では，ブログ文書から人間の体験を表現するための最小構成要素として，時間，空間，行動属性から成る体験情報を抽出する手法を示す．提案法においては，係り受け解析により，動詞，名詞句，格助詞の組合せ情報を広く抽出した後，フィルモアの格文法解析，動詞の意味解析をすることで，行動内容を示す表現（行動属性値）を順に選択していく．さらに，人間の行動は時間的・空間的要因によって規定されているという点に着目し，時間，空間，行動属性間の相関ルール（assotiation rule，アソシエーションルール）を抽出し属性間のつながりの強さを評価することで，時間，空間，行動属性値の意味ある組合せを選択し，抽出する．提案法に基づき，体験情報を構成する属性を指定することで，柔軟に人々の体験情報を検索，要約可能な体験ブログマップ（Blog Map of Experiences）を開発した．提案システムは，データベースに格納された体験情報への構造化されたアクセス方法を提供するものである．評価実験においては，体験情報抽出手法を構成する各ステップにおいて，ブログ記事からの体験情報抽出精度が順に改善することを示す．

3.2　提案手法

本節では，2 章で定義した体験情報の最小構成要素として時間，空間，行動属性（動作とその対象）を抽出する手法を述べる．提案手法は，次の 3 つのステップから構成される．

[1] ブログ記事の収集
[2] 本文情報からの行動属性値抽出
[3] 空間・時間・行動属性間のアソシエーションルール抽出

最初に，地名を検索クエリとしてブログ検索を行い，その地名に関連する記述が存在するブログ記事を収集する．次に，本文情報から人間の行動内容を示す表現

を抽出する．人間の行動内容を最も単純に表現するのは，動詞とその動詞を係り先に持つ名詞句の組合せである．提案法では，単純に，動詞，名詞，格助詞の組合せをブログ記事から抽出した後，日本語における格分析と動詞の意味解析に基づくアルゴリズムにより，人間の行動内容を示す表現を獲得する．最後に，人間の行動が時間的・空間的要因によって規定される点に着目し，時間，空間，行動属性間のアソシエーションルールを抽出し属性間のつながりの強さを定量的に評価することで，時間，空間，行動属性値の意味ある組合せのみを選択し，抽出する．次項からその処理の詳細を述べる．

3.2.1 ブログ記事の収集

ブログ記事の収集は，既存のブログホスティングサービスが提供するキーワード検索機能を利用して行う．ブログホスティングサービスは検索結果を RSS 形式で配信するサービスを行っており，この機能を利用して検索結果を取得する．RSS とはブログホスティングサービスが提供する RDF rich Site Summary であり，メタデータが付与される形でブログ記事のタイトル，本文などの情報が整理されている．地域に関連するランドマーク名，寺社名などの地名をキーワード検索の検索クエリとして，定期的にブログ記事を収集する．検索クエリとする地名情報は，既存の GIS を利用するなどして事前に収集し，登録しておく．また，この地名情報が，空間属性値として格納される情報となる．RSS で定義された情報の中でも特に，地名（検索クエリ），タイトル，リンク，本文，日付情報の組合せをデータベースに蓄積する．以下にそのステップの詳細を示す．

[1] 地名をブログホスティングサービスのキーワード検索エンジンの検索クエリとして投入する
[2] 検索結果を RSS 形式で取得する
[3] RSS を解析し，ブログ記事のタイトル，本文，投稿日時を抽出する
[4] ブログデータベースに格納する
[5] 一定時間経過後，(2) に戻る

3.2.2　体験情報抽出

収集したブログ記事の本文情報を解析し，人間の行動内容を示す表現（行動属性値）を抽出する．実世界の体験に基づく体験型ブログには，多くの場合，“何をしたか” という行動の内容を示す文が含まれているが，たとえば，感想を述べた文，得られた知識や，動作主の状態を述べた文など，異なる観点で書かれたものも数多く含まれている．“何をしたか” を示す文を特定し，動作属性値と対象属性値の組合せからなる行動属性値として抽出する必要がある．提案手法においては，係り受け解析により，動詞，名詞句，格助詞の組合せ情報を広く抽出した後，フィルモアの格文法解析，動詞の意味解析をすることで，行動内容を示す表現を順に取捨選択していく．

係り受け解析

ブログ記事の本文中に出現する検索語の周辺テキストから，動詞，名詞句，格助詞の組合せから成る情報を抽出する．組合せ情報の中でも，動詞が動作属性値の，名詞句が対象属性値の候補となる．格助詞は，対象属性値を抽出する際の手掛かりとして用いる．ブログ記事の本文情報を係り受け解析し，その結果に基づいて抽出処理を行う．係り受け解析とは，各単語の形態素を解析した後，各単語の係り先を決定するものである．提案手法においては，まず，本文中に存在する動詞を特定する．次に，抽出した動詞を係り先に持つ名詞節を抽出する．さらに，名詞節の中でも，“名詞句＋格助詞” という組合せから構成されるもののみを抽出し，最終的に，名詞句，格助詞，そして係り先の動詞の組合せ情報として保存する．単一の動詞に，複数の “名詞句＋格助詞” が紐づけられる可能性があるが，その場合は，動詞と “名詞句＋格助詞” の全組合せを考慮して抽出するものとする．

Refinement 1: フィルモアの対象格抽出

フィルモアの格文法に基づき，動作が作用する対象を示す名詞句を抽出する．フィルモアは，動詞とその深層格との組合せから文を分析する理論を提案した [53]．格

は，単語と単語，主には，動詞と名詞（名詞句）の間の意味的関係を示すものである．表現レベルで考えるか，それとも意味的レベルで考えるかで，表層格と深層格とに分類される．さらに，深層格は，動作主格，経験者格，道具格，対象格，源泉格，場所格，時間格の 8 つに分類される．本研究においては，動作の対象を表現するものとして，深層格の中でも特に，対象格を抽出する．日本語における対象格は，“〜を” という形式で現れる「を格」と，“〜に” という形式で現れる「に格」によって表現される．したがって，格助詞として，“を” か “に” が用いられている動詞，名詞句，格助詞の組合せのみを抽出する．

Refinement 2: 動作動詞の抽出

一般に，文は用いられる動詞によって，次の 3 種類に分類することができる．

- する文（行為）
  例）紅葉を見る，お守りを買う，湯豆腐を食べる
- なる文（過程）
  例）花火大会が延期になる
- である文（状態）
  例）桜が五分咲きである

行動内容を示す表現は，この中の “する文” に該当する．提案手法においては，“見る”，“聞く”，“食べる” などの動作動詞と “拝観”，“観覧” などのサ変名詞に着目して “する文” を特定する．サ変名詞は “サ変名詞＋する” のように用いられ，行為を示す動詞的な役割を果たす品詞である．

Refinement 3: 移動を示す動詞の削除

ある場所における具体的な行動内容を含まない表現を削除する．たとえば，“行く” や “来る” などの移動を示す動詞は，“清水寺に行く” などのように，表層格の

「に格」や「へ格」と組み合わさることで，移動の着点を意味する文となる．情報としては，その場所に動作主が訪れたということだけであり，その場所で何をしたかまでは含んでいないため，移動を示す動作動詞を削除する．

Refinement 4: アソシエーション分析

ブログ記事に付与された日付情報は，記事が投稿された日時であり，行動をした日時とは異なる場合がある．また，検索クエリとした地名が，必ずしも本文から抽出した行動内容をした場所とは限らない．提案法では，ブログ記事全体の傾向に基づいて，時間，空間，行動属性値の尤もらしい組合せのみを抽出し，情報抽出処理に反映する．

前節の解析結果に基づいて，アソシエーションルール抽出を適用するトランザクションデータを作成する．生成するトランザクションのスキーマは，

$$T = \{d, time, location, activity_1, ..., activity_{M^d}\}$$

である．d はブログ記事（動作主）を一意に識別するための ID，$time$ は d が投稿された日付情報（時間属性），$location$ は，d を収集した際に検索クエリとして用いた地名（空間属性），$activty_i$ は d の中で i 番目に抽出された行動情報（行動属性），M^d は d の中で抽出された行動情報の総数である．$activity_i =< verb_i, particle_i, noun_i >$ であり，$verb_i$，$particle_i$，$noun_i$ は i 番目の行動情報の動詞，格助詞，名詞句である．

収集したすべてのブログ記事からトランザクションを生成した後，時間，空間，そして行動属性間のつながりの強さを評価するために，ある閾値以上の支持度，確信度を持つ，

$$\{\text{空間，時間}\} \Rightarrow \{\text{行動（動作，対象）}\}$$

形式の属性間アソシエーションルールを抽出する．このルールは，人々が，ある時間にある場所を訪れた場合に，ある行動をする傾向があることを意味する．全ブログ記事を分析して得られたルールは，一般性（支持度）と信頼性（確信度）が高く，時間，空間，行動属性値の意味ある組合せを表現している．アソシエーショ

ンルール抽出によって得られた時間，空間，行動属性値の意味ある組合せは体験表現辞書に格納する．

体験情報抽出処理の最後のステップは，体験表現辞書に格納された表現との照合である．まず，Refinement 3 までの処理結果に基づき，個々のブログ記事から得られた時間，空間，行動属性値の組合せ情報を取得する．次に，取得した組合せ情報が体験表現辞書に格納されているかどうかを照合する．照合が成功した場合にのみ，つまり，尤もらしい組合せ情報であると判定された場合にのみ，体験情報として出力を行う．

3.3 体験ブログマップ

構造化した人々の体験情報をもとに地域情報検索を支援する体験ブログマップ（Blog Map of Experiences）を提案する．ユーザが体験を構成する時間，空間，行動のいずれかの属性を指定すると，システムは未指定の属性を人々の体験情報をもとに補完し，ユーザに出力する．さらに，提示された体験情報を，情報検索の入り口として利用することも可能である．本節では，体験ブログマップの機能をその利用方法とともに説明する．

3.3.1 体験情報の可視化機能

ユーザの典型的な利用シーンは，現在地や目的地周辺での行動の選択肢を知りたい，とった場合である．ユーザが検索フォームに現在地や目的地を示す地名を入力すると，システムは，

$$\{\text{空間}\} \Rightarrow \{\text{行動（動作，対象）}\}$$

形式のアソシエーションルールを抽出し，ルールの結論部（行動属性情報）を確信度の降順にソートして提示する．体験ブログマップのインタフェースを図 3.1 に示す．地図上のアイコンは各場所を表しており，アイコンの吹き出しの中に，その場所に関して抽出された人々の行動情報がリスト表示される．観光客が使うこ

とを想定し，行動情報は人気順（確信度の降順）にソートして提示するように設定しているが，検索結果の下位の行動を眺める，あるいは確信度の昇順に行動情報をソートさせることで，マイナーな行動を発掘するための用途としても使うことができる．

ユーザはプルダウンメニューから時間属性を指定することも可能である．たとえば，ユーザが旅行で訪れる場所と期間が決まっており，その場所でどのような行動の選択肢があるのかを調べたいという状況では，時間属性を指定した検索が有効である．その場合，システムは，

$$\{\text{空間，時間}\} \Rightarrow \{\text{行動（動作，対象）}\}$$

形式のアソシエーションルールを抽出して提示する．提案システムにおいては，過去の時間属性のみ指定可能だが，人間の体験は四季によって規定される場合も多く，一年前，同じ場所で人気があった行動は，再び人気となる可能性も高い．また，訪問日が近い場合は，直近一ヶ月間を指定して，最近の傾向を参考にするのも有効である．さらに，行動属性の一部である動作属性を指定した検索も可能である．たとえば，ユーザが“食べる”と指定すれば，“湯豆腐”や“八つ橋”などの都市で人気の食べ物に関する行動内容のみをシステムは出力する．ユーザが，ある特定の時間にある場所にいて，“見る”，“食べる”，“買う”などの動作の内容も決定しているが，その対象を何にするか決めかねている場合にこの種の検索が有効である．

3.3.2　場所ランキング機能

もう 1 つの有効な利用シーンは，目的の行動内容が予め決まっており，それを実行するための場所を知りたい，といった場合である．たとえば，京都周辺でホタルを見たいが，どこに行けば見ることができるか，といった検索要求をユーザが持つ場合である．その場合，システムは，

$$\{\text{行動（動作，対象）}\} \Rightarrow \{\text{空間}\}$$

図 3.1: 体験ブログマップのユーザインタフェース

形式のアソシエーションルールを抽出して提示する．図 3.1 の地図右側のリスト内で，ルールの結論部を確信度の降順にソートして提示する．行動をする時間を検索結果に含めることも可能であり，その場合は，

$$\{\text{行動（動作，対象）}\} \Rightarrow \{\text{空間，空間}\}$$

形式のアソシエーションルールを抽出する．たとえば，京都周辺でホタルを見たいが，いつ，どこに行けば見ることができるか，といった検索要求に応えることができる．

3.3.3　ブログ検索機能

図 3.1 の地図上に表示された行動語（行動属性値）をクリックすると，地名（空間属性値）と行動語（行動属性値）を検索クエリとした場合のブログ検索結果の画面に遷移する．別の言い方をすると，特定の体験情報を検索クエリとして用いたブログ検索が可能である．体験情報の抽出元の文書に目を通すことにより，目的の体験に関する感想や評判を効率的に知ることが可能である．また，ブログ記事の筆者とブログ記事のコメント欄を通じて情報交換することにより，ブログ記事に書かれていない情報を引き出す，といった使い方も可能である．

3.4　評価実験

3.4.1　プロトタイプシステム

提案手法に基づきプロトタイプシステムを実装した．システムの構成を図 3.2 に示す．ブログ記事の収集に利用したブログホスティングサービスは goo ブログ [1] と Bulkfeeds[2] である．日本語の形態素解析は ChaSen[54] を，係り受け解析は CaboCha[55] を，動作動詞の判別には日本語語彙体系 [56] を利用した．日本語

[1] http://blog.goo.ne.jp/
[2] http://bulkfeeds.net/

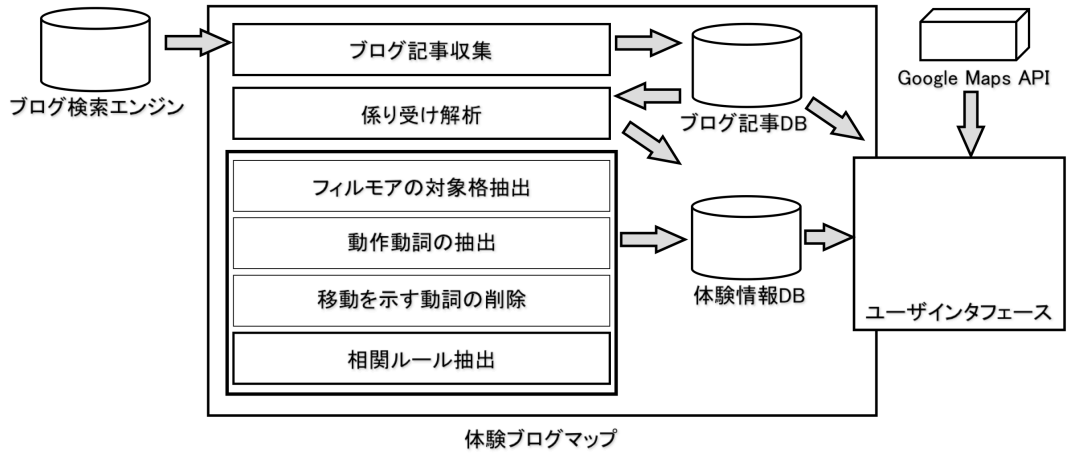


図 3.2: 体験ブログマップのシステム構成

語彙体系においてはいくつかの観点で動詞がカテゴリー分けされており，動作動詞も判別可能である．“行く”や“来る”などの移動を示す動詞，及びその同義語は，手動で選定した．ブログの日付情報（時間属性情報）は月ごとに集計して用いている．地図インタフェースは Google Maps API [3] を使って実装した．

3.4.2 情報抽出精度の評価

本節では，提案手法の体験情報抽出手法としての妥当性を検証する．前述の通り，提案法は以下のステップで構成される．

[1] 係り受け解析による動詞，格助詞，名詞句の組合せ情報の抽出（ベースライン手法）
[2] Refinement 1: フィルモアの対象格抽出
[3] Refinement 2: 動作動詞の抽出
[4] Refinement 3: 移動を示す動詞の削除
[5] Refinement 4: アソシエーション分析

[3] https://maps.google.co.jp/

最も単純に，係り受け解析によって動詞，格助詞，名詞句の組合せ抽出を行った場合をベースラインとし，各ステップで体験情報の抽出精度が順に向上することを確認する．実験設定を以下に示す．

- 検索クエリとして用いた地名: 京都市における主要な 8 個の地名（清水寺，嵐山，平安神宮，南禅寺，伏見稲荷大社，貴船神社，京都駅，京都市美術館）
- データセット: 25,320 のブログ記事
- ブログ記事が投稿された期間: 2005 年 8 月 15 日〜12 月 15 日（4ヶ月間）
- ルールのタイプ: { 空間 } ⇒ { 行動（動作，対象）}（最小確信度 = 0.001）

アソシエーションルール抽出において設定する最小確信度 0.001 とは，たとえば，ある地名に関して収集したブログ記事が 5,000 件だった場合，5 件以上の記事で共起した空間属性，行動属性値の組合せのみをルールとして抽出することを示す．ただし，その共起回数の閾値が 2 件未満となった場合は，2 件以上で共起した組合せのみを扱うこととしている．被験者に対して，ブログ記事から，正しく地名と行動の組合せ情報が抽出された場合に正解ラベルを付与するように求めた．正解ラベルを付与する際には，体験情報抽出結果とともに情報抽出元のブログ文書も同時に提示し，ある場所でその行動をした事実が認められる場合にのみ，正解ラベルを付与するようにした．評価指標としては，適合率と再現率の調和平均である F 値（F-measure）を用いた．適合率（Precision）は，提案法が体験情報として抽出した体験情報の中で，正解ラベルが付与されたものの割合である．再現率（Recall）は，正解ラベルが付与された体験情報の中で，提案法が実際に抽出したものの割合である．ある検索クエリ q（地名）に関して得られた体験情報に基づいて適合率 $Precision_q$，再現率 $Recall_q$ を計算した後，以下の式で検索クエリ q に関する F 値 $F\text{-}measure_q$ を計算する．

$$F\text{-}measure_q = 1/\left(\frac{1}{2}\cdot\frac{1}{Precision_q} + (1-\frac{1}{2})\cdot\frac{1}{Recall_q}\right) \tag{3.1}$$

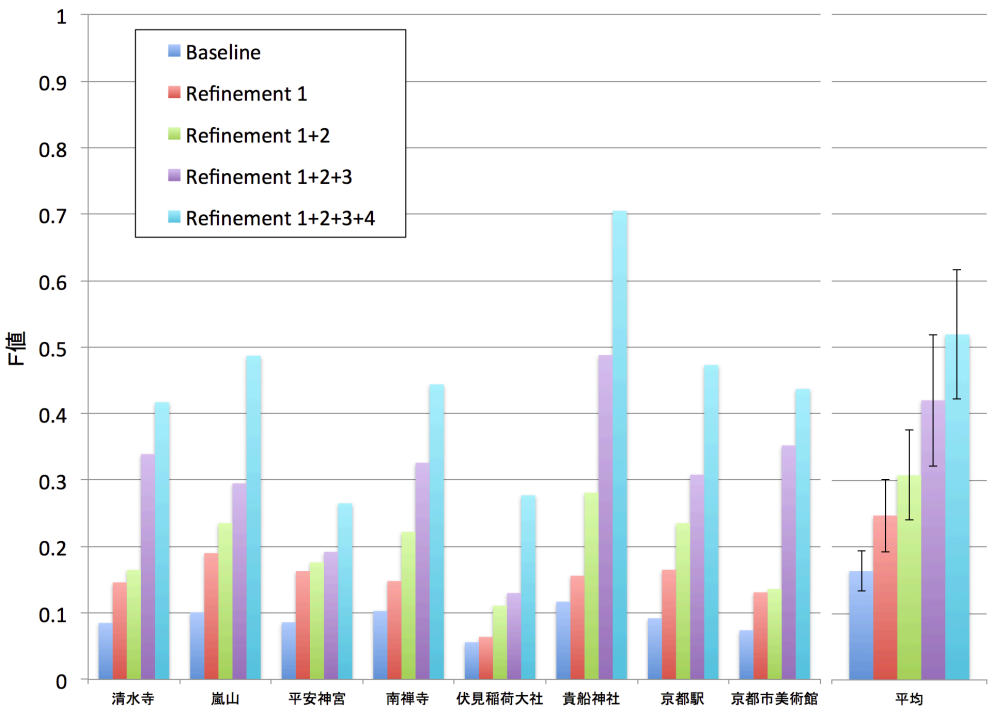


図 3.3: 体験情報抽出精度（F 値）

各検索クエリ（地名）ごとの F 値，及び，それらを全検索クエリで平均した値を図 3.3 に示す．図に示すように，各ステップが，ベースライン手法からの精度改善に寄与していることが分かる．参考までに，適合率，再現率についても，それぞれ図 3.4，図 3.5 に示す．提案手法の各ステップは，体験情報として適切な表現を，その特徴をもとに絞り込んでいく手法である．そのため，再現率は各ステップで低下する傾向があるが，適合率は上昇するため，総合的には精度向上の方向に提案手法が作用している．最終的に，適合率の平均値は 0.5 未満となったが，refinement 4 のアソシエーション分析における最小確信度の設定値次第で，適合率を重視するか，再現率を重視するかをコントロール可能である．また，各検索クエリごとに，最小確信度や最小支持度を適切に設定することで，精度（F 値）向上も期待できる．

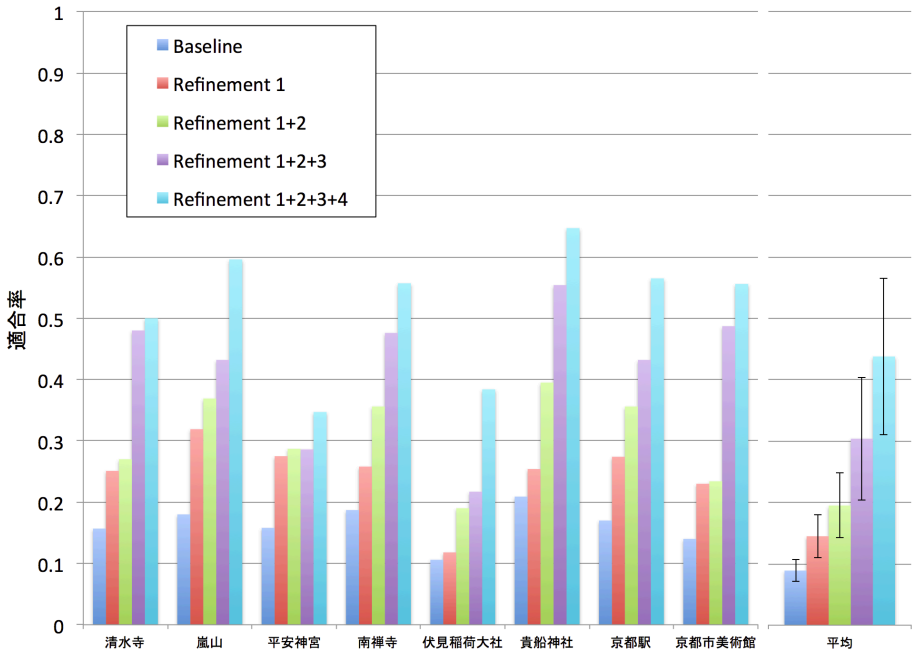


図 3.4: 体験情報抽出精度（適合率）

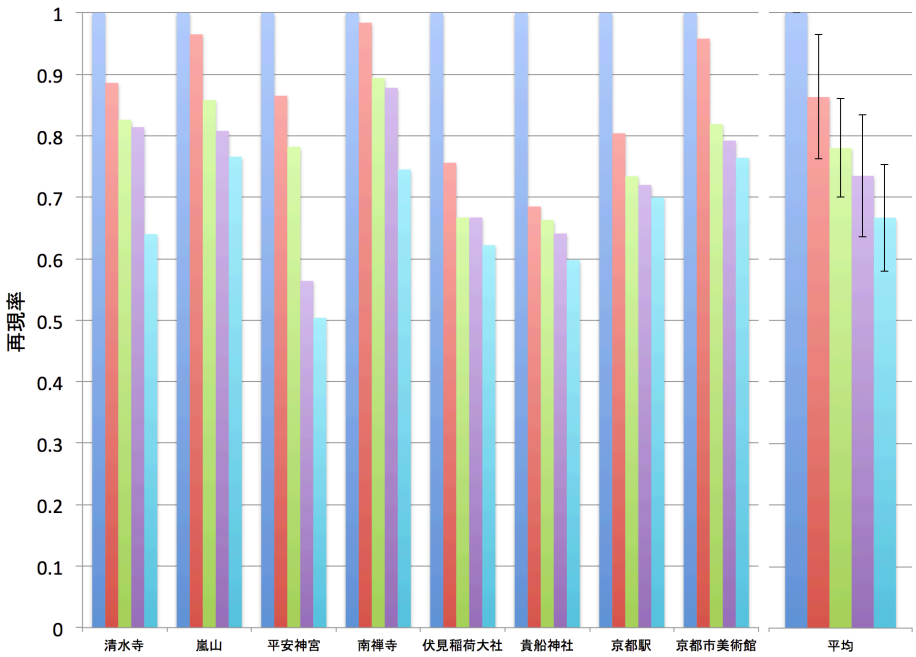


図 3.5: 体験情報抽出精度（再現率）

3.4.3 行動間のアソシエーション分析

アソシエーションルール抽出の最も典型的な応用先はマーケットバスケット分析である．マーケットバスケット分析においては，商品の POS（Point of Sales）データを分析して，一度に購入されやすい商品の組合せを発見する，得られた知識は，店舗内の商品配置の決定やクロスセル戦略などに活用される．従来のマーケットバスケット分析は“商品を購入する”という人間行動の一側面をとらえるものにすぎなかったが，本研究成果により，都市における人々の実世界行動を対象としたマーケットバスケット分析が可能となる．具体的には，提案する体験情報抽出手法によって得られた人々の体験情報の中から，

$$\{\text{行動（動作，対象）}\} \Rightarrow \{\text{行動（動作，対象）}\}$$

の形式のルールを抽出する．得られた知識は，個人の旅行計画やマーケッターによる人々の行動調査，分析などに活用可能であると考える．

本実験においては，年末年始休暇期間の初詣に関する人々の行動傾向を分析した．実験に用いたデータセットの詳細を以下に示す．

- 検索クエリとして用いた地名: 日本国内の 139 の寺社
- 収集したブログ記事の総数: 20,593
- ブログ記事が投稿された期間: 2006 年 1 月 1 日〜1 月 15 日（15 日間）

アソシエーションルール抽出においては，最小支持度を 5 件，最小確信度を 0.30 に設定した．最終的に 15 個のルールが得られ，その平均確信度は 0.50 であった．得られた 15 個のルールを可視化したものが図 3.6 である．エッジの太さがルールの支持度を，エッジに付与された数値が確信度を表している．条件部に複数の値を持つルールは，塗りつぶし四角形でルールの条件部を結合することで表現している．たとえば,「お守りを買う ⇒ おみくじを引く」というルールが存在することが分かれば，お守りを売っている場所とおみくじを引く場所を隣接させる，などの意志決定につなげることができる．また，初詣客が行動計画をする際にも参考になる情報である．

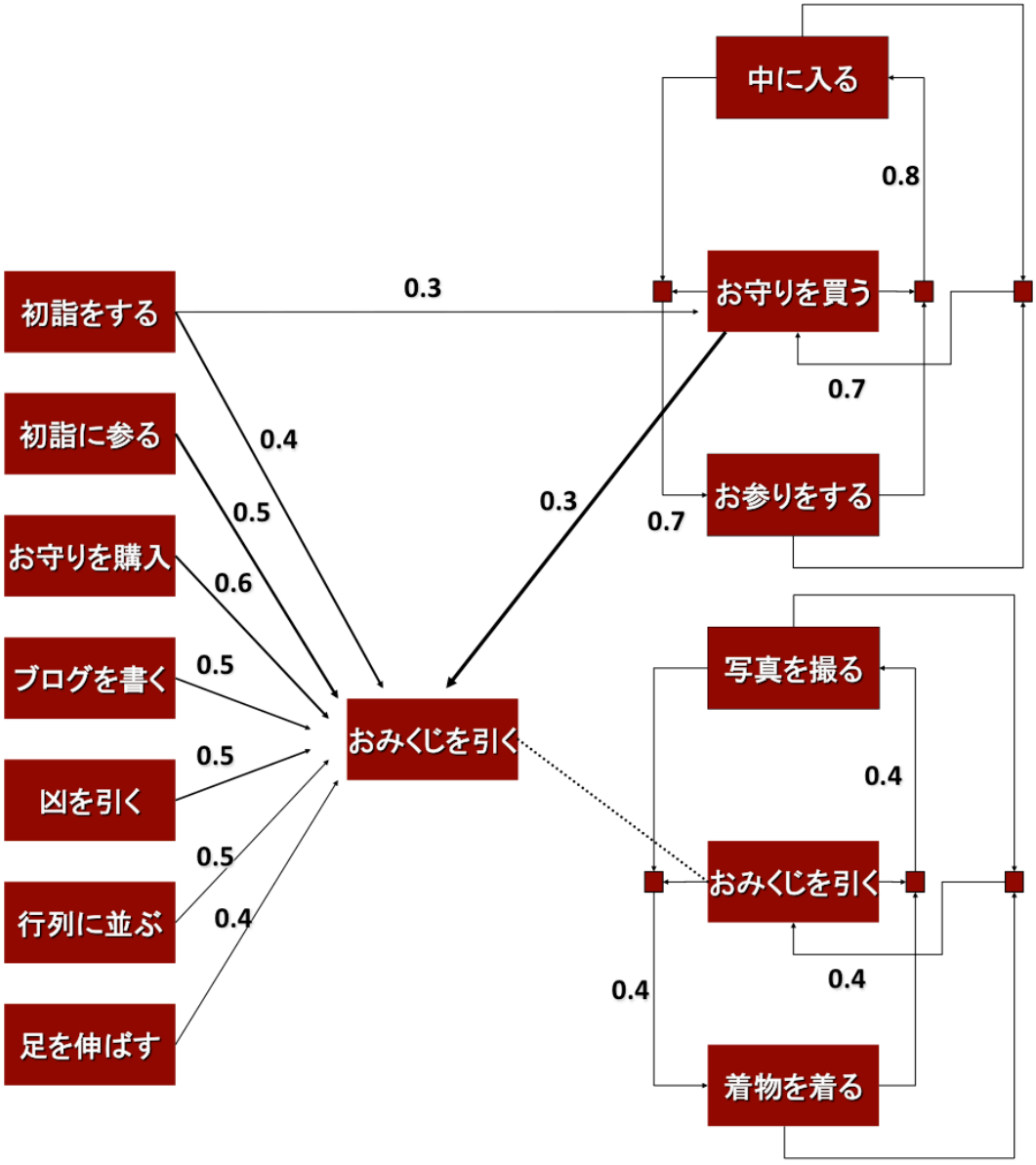


図 3.6: 行動間ルールの抽出結果

3.5 結言

本章では，ブログ上に自然言語で記述された体験を，時間属性，空間属性，行動属性から構成される情報として抽出する手法を示した．提案手法は，自然言語処理技術とデータマイニング技術を融合させた，テキストマイニング技術である．動作とその対象とから構成される行動属性情報を，係り受け解析，フィルモアの格文法解析，動詞の意味解析などの自然言語処理技術を用いて抽出する．さらに，人間の行動は時間的・空間的要因によって規定されているという点に着目し，全ブログデータ中で時間，空間，行動属性情報の共起しやすい組合せをルールとして抽出し，行動属性情報の抽出処理に反映させることで，精度の高い体験情報抽出を実現した．評価実験においては，動詞，名詞句，格助詞の組合せ情報を単純な係り受け解析器で抽出するベースライン手法と比較して，大きな精度向上を確認した．また，提案技術の応用として，体験情報を構成する属性を指定することで，柔軟に人々の体験情報を検索可能な体験ブログマップ（Blog Map of Experiences）を提案し，その機能を述べた．

今後の課題は，収集した体験情報データから発見性の高い知識（パターン，ルール）を抽出することである．単純にアソシエーションルール抽出技術を適用することで得られる“北野天満宮 ⇒ お守りを買う”や“お守りを買う ⇒ おみくじを引く”などのルールは，多くの人々の行動を直接的に反映している点，精度の高い体験情報抽出を実現する点で価値がある．しかし，ガイドブックや地域情報サイトにも掲載されている情報であり，新たな知識を発見したとは言い難い．次章では，数多くのパターンやルールの中から，有用なものを選択するための仕組みについて議論する．

第4章　人々の体験に関する有用な相関ルールの選択手法

4.1　緒言

本章では，体験情報抽出手法によって構造化した大量の体験情報集合をもとに，人々の傾向に関する有用な知識（パターンやルール）を獲得する仕組みを検討する．日々，変化する人間の行動傾向に対して，適切な仮説を立てることは容易ではないため，仮説検証型で知識を獲得する方法では多くの重要な傾向を見落としてしまう可能性がある．その一方，相関ルール（association rule，アソシエーションルール）抽出に代表されるデータマイニング技術は，データ傾向を説明するパターンやルールを網羅的に自動抽出する手法であり，分析者が仮説を立てる必要がない．しかし，単純に，収集した大量の体験情報を集約しデータ傾向を説明するパターンやルールを抽出した場合，誰もが知っている知識を表現するパターンやルールが少量得られる，もしくは，ユーザの目的にそぐわないものが多く含まれる膨大な量のパターンやルールが得られるだけである．人々の体験情報を分析することで得られる有用な知識とは何かを議論し，その知識を効率的に抽出するための仕組みを構築することが本章の目的である．

一般に，ある情報の価値は，その情報の利用者の知識量や状況などによって異なる．本章では，構造化された体験情報の活用に関する 2 つの利用シナリオを想定した．第 1 の利用シナリオでは，ある時間に，ある場所を人々が訪れる目的に関心を持つユーザ，たとえば，都市の調査会社を想定した．現在，GPS 機能を標準搭載したモバイル端末やカーナビゲーションシステムの普及を背景に，調査会社は比較的容易に人々の移動や集中に関する情報を把握できるようになってきた．

しかし，なぜ，そこを訪れたかの説明となる情報は，アンケート調査などを実施して都市の生活者に聞くしかほかに方法がなかった．そこで，ある時間に，ある場所を人々が訪れる理由や目的の説明となる情報として，“ある特定の状況（時間，空間）において，人々が特徴的にしている行動” を表現するルールを “興味深い知識” として抽出する．たとえば，“5 月” という時間，“銀閣寺” という場所で人々が特徴的にしている行動は “ホタルを見る” であることを示す知識は，ルール “5 月，銀閣寺 ⇒ ホタルを見る” で表現でき，5 月に銀閣寺に人が集まる理由や目的を説明する知識として価値がある．提案手法は，人間の体験を構成する属性の中でも特に 5 属性 {時間，空間，動作，対象，感情} をブログから抽出した後，構造化した体験情報集合から少数派なものも含めたアソシエーションルールを幅広く抽出する．さらに，特定の時間や空間条件において，行動の出現傾向がどの程度変化するかを数値化することでルールの価値を評価する．評価実験においては，ブログ記事約 4,800 万件から “ある特定の状況（時間，空間）において，人々が特徴的にしている行動” を示すルール発見を試み，提案手法の有効性を確認した．

第 2 の利用シナリオで想定したユーザは，旅行ガイドブックや Web 検索エンジンなどのメディアを用いて地域情報を収集する旅行者である．旅行者は，これらのメディアでは紹介されていないロングテールな情報を知りたい．また，メディアに含まれる主にサービス提供者側が生成した広告コンテンツの影響により，都市における人々の行動傾向に対して事実とは異なるイメージを抱いた場合に，その誤りを認識したい．これらの情報要求に応えるためには，都市に生きる人々の多様な体験が直接的に反映されているソーシャルメディアを有効に活用する必要がある．そこで，一般のメディアへの露出度が高く人々に認知されている行動と，都市で（ソーシャルメディアで）実際に人々がしている行動の差異発見につながるルールを，第 2 の利用シナリオにおける興味深い知識として抽出する．本研究では，Web 検索エンジンの検索結果ページでの出現回数や出現位置を分析し，ある行動がどの程度メディアを通じて認知されているかを推定する．その後，消費者側の観点で書かれたソーシャルメディアにおける出現傾向と比較し，そこに乖離があるかを分析する．

4.2 前処理: 体験情報集合の作成とアソシエーションルール抽出

自然言語で記述された非構造なソーシャルメディアデータから体験情報を構造化して抽出する．構造化データとは，コンピュータが処理できるように，その属性や意味を規定した情報である．人間の体験を構造化データとして関係データベースに格納すれば，SQL のようなデータベースに対する問合せ言語の検索機能，集約演算，ソート機能を用いて人間の体験に柔軟にアクセスすることができる．前章で述べた時間，空間，行動属性値の抽出に加えて，新たに感情属性値の抽出も試みる．さらに，それぞれの体験情報が，動作主にとって，成功だったのか，それとも失敗だったのかという観点で，成功/失敗属性値を付与する．本研究においては “成功/失敗は主に動作主の感情に因る” との仮定に基づく．時間，空間，行動属性値の抽出手法については前章で述べたため説明を省略し，感情，成功/失敗属性値の抽出について述べる．

4.2.1 感情情報の抽出

人間の主観に関する表現は，“評価” と “感情” とに大別できる．“良い” や “悪い” といった人間の “評価” は，ある対象に対する主観的な価値付けを示す表現であるのに対して，“嬉しい” や “悲しい” は，動作主体の心的状態を示す表現である．本手法では，後者の感情を示す表現を人手で収集し，感情語辞書を構築した．また，すべての感情語を，福原ら [12] の研究に基づき，喜び，驚き，困惑，怒り，悲しみ，疲労，不安，不満という 8 カテゴリに分類した．表 4.1 に，収集した感情語の一例を示す．括弧内の数値は，それぞれのカテゴリに含まれる感情語数であり，その総数は 121 である．

表 4.1: 登録した感情語の一例

カテゴリ	感情語
喜び（30）	嬉しい，笑う，爆笑，満足，感動，満喫
驚き（10）	衝撃，驚く，混乱，動揺，びっくり
困惑（12）	困る，悩む，苦悩，苦渋，落胆，凹む
怒り（6）	怒る，憤り，苛立つ，腹が立つ，非難
悲しみ（11）	悲しい，涙，嘆く，悲痛，号泣，切ない
疲労（19）	疲れる，疲労，ぐったり，がっかり
不安（14）	心配，気がかり，おびえる，怖い，恐い
不満（19）	不満，不平，後悔，悔しい，つまらない

4.2.2　成功/失敗情報の付与

テキストから抽出したそれぞれの体験情報 {時間，空間，行動（動作，対象），感情} に対して，動作主にとって成功だったのか，それとも失敗だったのかという観点で，成功/失敗情報を付与する．成功か失敗かの判断は，動作主の主観（“評価” と動作主が抱いた “感情”）に因ると考えられるが，特に，“感情” に因るところが大きい．たとえば，以下に示す体験情報の抽出例について考える．

（例）昨日，清水寺に紅葉を見に行きました．確かに紅葉はきれいだったのですが，観光客で大混雑 ··· がっかりです．→ { 時間，空間，動作，対象，評価，感情 }={ 昨日，清水寺，見る，紅葉，きれい，がっかり }

この例において，紅葉（対象）に対する評価は肯定的（きれい）だが，この行動を行った結果，動作主が抱いた感情は否定的（後悔）である．つまり，この動作主は，“清水寺に紅葉を見にいく” という行動選択を後悔しており，失敗だったと考えている．評価が対象に対する価値付けであるのに対して，人間の感情は，時間と空間に紐付けられた，ある行動に対する価値付けであるといえる．この観点で付与された成功/失敗情報は，他人がその行動を行うべきか否かの判断に役立てることができる．

本手法においては，ある行動を起こした結果，動作主が肯定的な感情を持った場合は成功であり，否定的な感情を持った場合は，失敗とみなし，得られた体験情報の感情情報から成功/失敗を導き出す．具体的には，感情カテゴリが“喜び”の感情語に対しては成功，“困惑”，“怒り”，“悲しみ”，“疲労”，“不安”，“不満”に対しては失敗，また，“驚き”に関しては，肯定，否定の判断が難しいため，肯定/否定以外とした．たとえば，“びっくり”のような感情語は，良い意味と悪い意味の両方で用いられることがある．

4.2.3 属性間アソシエーションルールの抽出

状況，行動，主観との間に内在する規則性をアソシエーションルールとして抽出する．アソシエーションルールは，$A \Rightarrow B$ という形式で表現され，“A が起こったという前提のもとで，B も同時に起こる”ことを示す．A 及び B は，それぞれ体験情報を構成する属性値の集合であり，$A \cap B = \phi$ である．具体的には以下に示す 4 タイプの属性間ルールを抽出対象とする．

タイプ 1: 状況と行動 { 空間，時間 }⇒{ 動作，対象 }
例 1）{ 北海道，5 月 }⇒{ 見る，桜 }

タイプ 2: 状況と主観 { 空間，時間 }⇒{ 感情 }
例 2）{ 京都，9 月 }⇒{ 嬉しい }

タイプ 3: 行動と主観 { 動作，対象 }⇒{ 感情 }
例 3）{ 引く，おみくじ }⇒{ がっかり }

タイプ 4: 状況と行動と主観 { 時間，空間 }⇒{ 動作，対象，感情 }
例 4）{9 月，京都 } ⇒ { 食べる，川床料理，不満 }

状況を構成する時間，空間属性については，どちらか一方が含まれていれば抽出対象とする．本章では，状況，行動，主観との間に内在するルールの中でも特に，状況（時間，空間）に特有な行動を表現するタイプ 1 のルールに焦点を当てる．

4.3　人々の体験情報集合からの知識発見

構造化した大量の体験情報集合から得られたアソシエーションルール集合の中から，都市における人々の行動傾向に関する有用な知識といえるもののみを選択する手法を述べる．構造化された体験情報の活用に関する 2 つの利用シナリオを想定して興味深い知識を定義し，それを抽出するための仕組みを述べる．

4.3.1　シナリオ 1: ある状況で特徴的に出現する行動情報の抽出

ある時間に，ある場所を人々が訪れる目的に関心を持つ，調査会社やマーケッターをユーザとして想定し，“ある特定の状況（時間，空間）において，人々が特徴的にしている行動” を表現するアソシエーションルールを人々が訪れる理由や目的の説明となる知識として抽出する．提案手法においては，特定の時空間条件において，行動の出現傾向がどの程度変化するかを数値化することでルールの価値を評価する．たとえば，“ホタルを見る” は希少な行動である．その行動をしたいと感じる人は多くても，見ることが可能な場所が限られるために実現に至る人は少ない．しかし，銀閣寺という場所では “ホタルを見る” 人たちがほかの場所と比べて相対的に多いため，銀閣寺に特徴的な行動であるといえる．また，“アジサイを見る” 行動をする人は，年間を通して 5 月前後が最も多いために 5 月に特徴的な行動である．このように，ある特定の状況において，ほかの状況と比較して相対的に出現確率が高くなる性質を示すルールを抽出する．

提案手法は，条件部にある条件を加えることで，どの程度結論部の出現確率が変化したかを評価する．条件部の部分集合 $A' \subseteq A$ を加えることで，ルール $A \Rightarrow B$ の結論部 B がどの程度変化したを示すスコアを $Score([A \Rightarrow B], [A' \subseteq A])$ と表す．また，本項では説明の簡略化のため，空間属性値，時間属性値，行動属性値をそれぞれ，空間，時間，行動と呼ぶ．提案手法では，ある空間 g における行動 x の出現傾向を示すルール $g \Rightarrow x$ を，“空間 g に特徴的に出現する行動 x であるか” という観点で評価する場合，条件 g を加えることで x がほかの場所と比べて何倍起

こりやすくなるかを以下の式で評価する．

$$Score([g \Rightarrow x],[g]) = \frac{P(x|g)}{\frac{1}{|G|-1}\sum_{\{g'|g' \in G, g' \neq g\}} P(x|g')} \propto \frac{P(x|g)}{\sum_{\{g'|g' \in G, g' \neq g\}} P(x|g')} \quad (4.1)$$

ここで，G は空間 g の集合，$|G|$ はその要素数であり，分母は（g を除いた）行動 x の出現確率の平均である．条件付き確率 $P(x|g)$ はアソシエーションルール $g \Rightarrow x$ の確信度である．同様に，ある時間 t での行動 x の出現傾向を示すルール $t \Rightarrow x$ を，“時間 t に特徴的に出現する行動 x であるか” という観点で評価する場合，条件 t を加えることで x がほかの時間と比べて何倍起こりやすくなるかを以下の式で評価する．

$$Score([t \Rightarrow x],[t]) \propto \frac{P(x|t)}{\sum_{\{t'|t' \in T, t' \neq t\}} P(x|t')} \quad (4.2)$$

ここで，時間 t は 1 月，2 月など，一定の区間で時刻をまとめ離散値に変換してある．T はその集合である．ある状況（空間 g と時間 t の組合せ）で行動 x が出現することを示すルール $g,t \Rightarrow x$ が，“空間 g と時間 t の組合せに特徴的に出現する行動 x であるか” という観点で評価する場合も同様である．

$$Score([g,t \Rightarrow x],[g,t]) \propto \frac{P(x|g,t)}{\sum_{\{g'|g' \in G, g' \neq g\}} \sum_{\{t'|t' \in T, t' \neq t\}} P(x|g',t')} \quad (4.3)$$

次に，ある空間 g と時間 t をともに条件部に含むアソシエーションルール $g,t \Rightarrow x$ が得られた場合の分析方法を示す．ある空間 g における行動 x の出現に関して，時間 t という条件がどの程度寄与しているかを評価する場合，時間 t を条件部から除いたルールと確信度（条件付き確率）を以下の式で比較する．

$$Score([g,t \Rightarrow x],[t]) = \frac{P(x|g,t)}{\frac{1}{|T|-1}\sum_{\{t'|t' \in T, t' \neq t\}} P(x|g,t')} \quad (4.4)$$

次に，ある時間 t における行動 x の出現に関して，空間 g という条件がどの程度関与しているかを評価する場合，空間 g を条件部から除いたルールと確信度（条件付き確率）を以下の式で比較する．

$$Score([g,t \Rightarrow x],[g]) = \frac{P(x|g,t)}{\frac{1}{|G|-1}\sum_{\{g'|g' \in G, g' \neq g\}} P(x|g',t)} \quad (4.5)$$

最後に，数式（4.4）と数式（4.5）で得られた値を比較し，各条件部がどの程度，行動の出現確率の上昇に寄与していたかを判断する．たとえば，“嵐山，4 月 $\Rightarrow$ うどんを食べる” が，高い確信度を持つルールとして抽出された場合，ルールの利用者は，“嵐山，4 月” という組合せにおいて，うどんを食べる人が特徴的に多いと解釈する可能性がある．しかし，上記の分析から，数式（4.4）で算出した値よりも，数式（4.5）で算出した値のほうが大きいことがわかれば，4 月という条件が加わらなくても嵐山では年中，うどんを食べる人が多い，といった事実に気づくことができる．条件部の各要素が結論部の予測にどの程度，寄与しているのかを丁寧に評価することにより，ルールの誤った解釈，認識を回避できる．

データマイニングの知識発見分野においては，アソシエーションルールの興味深さ（Interestingness）を測る様々な指標が提案されている，それぞれの指標は，データの異なる側面を評価するため，得られる結果も異なり，多様な観点からのデータ分析が可能である．以降，提案手法のその中での位置付けを述べる．Geng ら [57] によると，ルールの興味深さを測る指標は，客観的指標（$objective\,measure$）と主観的指標（$subjective\,measure$）とに大別できる．客観的指標が，データのみに依存する指標である一方，主観的指標は，データ自身に加え，ユーザの知識や背景をも考慮する．提案手法は，主に客観的指標に関する．客観的指標によるルールの評価において，最も重視されているのは，ルールの一般性（$generality$）と，信頼性（$reliability$）の側面である．一般性とは，データの特徴をどの程度反映しているかという観点でルールを評価するものであり，支持度（$support$）や被覆度（$coverage$）などがこれに該当する．全レコード数を N，集合 A を含むレコード数を $n(A)$，集合 A と集合 B をともに含むレコード数を $n(A,B)$ としたとき，ルール $A \Rightarrow B$ の支持度は $P(A,B) = \frac{n(A,B)}{N}$ で，被覆度は $P(A) = \frac{n(A)}{N}$ で表される．信頼性の評価には，確信度 ($confidence$) やリフト ($lift$) などの指標が用いられる．確信度は，集合 A が与えられたときの条件付き確率 $P(B|A) = \frac{n(A,B)}{n(A)}$ で表され，条件付き確率が大きいほど信頼性の高いルールとする．しかし，確信度には盲点がある．たとえば，ルール $A \Rightarrow B$ について確信度が 75%である場合を考える．これは，一見，高い数値のようであるが，そもそもの集合 B の出現確率 $P(B)$ が 80%であれ

ば，むしろ，結論部の予測に，条件 A がマイナスに働いていることになる．ほかのルールとの相対性に基づいてルールを評価する指標が，リフトである．リフトは，$\frac{P(B|A)}{P(B)}$ で定義され，条件 A を加えることで B が何倍起こりやすくなるかを示しているといえる．提案手法は，複数ルールを比較することで大局的な観点でルールの価値を判断する点で，リフトと共通している．なお，ここに示した尺度以外にも，統計に基づく指標である χ^2 値や，情報量に基づく指標である J-measure[58] など，統計学，情報理論，情報検索などの分野に起因し，様々な客観的な指標が提案されている．評価実験においては，これらの指標を比較手法とし，提案手法の有効性を議論する．

4.3.2 シナリオ 2: 人々に認知されている行動とユーザの実世界行動の差異発見

主に，旅行ガイドブックや Web 検索エンジンなどのメディアを用いて地域情報を収集する旅行者をユーザとして想定し，それらのメディアへの露出度が高く人々に認知されている行動と，都市で実際に人々がしている行動の差異発見につながるアソシエーションルールを興味深い知識として抽出する．ソーシャルメディアは都市の利用者が情報交換をする場であり，都市に生きる人々の多様な体験が直接的に反映されているため，その差異には，多くの人に知られていないロングテールな行動が数多く含まれていると考える．また，“自社の商品を知ってもらいたい，選んでもらいたい” と考えるサービス提供者は，テレビコマーシャル，街頭の看板広告，新聞，雑誌，自社 Web サイト，Web 検索エンジン，Web ポータルサイトなど，様々なメディアを利用して，商品，商品に関する体験の認知度を高める戦略をとる．すべての情報を信用すると，都市の人々の行動に対して，実際と異なるイメージを抱いてしまう危険性があるが，ソーシャルメディアを利用することで実態を正しく把握することができる．提案手法は（1） Web 検索エンジンを用いた認知度評価と（2） 認知度とソーシャルメディアにおける出現傾向の比較，の 2 つのプロセスから構成される．以下，順にそのプロセスを説明する．

Web 検索エンジンを用いた認知度評価

ここでは，“清水寺で紅葉を見る” のような空間属性，行動属性値の組合せで表現できる体験情報の認知度を評価する．人が何を知っているか/知らないかを把握することは困難であるが，人々がどのように情報収集するか，の入手経路を考えることで，多くの人に知られている可能性が高い情報を知ることは可能だと考える．本研究では，多くの旅行者の重要な情報源として Web 検索エンジンの利用を考える．Web 検索エンジンの検索結果の出現位置を考慮して，情報の認知度を評価する．

一般に，情報検索ユーザは，検索結果の上位の Web ページから順に目を通していくと考えられるため，Web 検索エンジンの検索結果の上位に含まれるほど，また，出現する回数が多いほど，ある情報の認知度が高いとみなす．本手法においては，体験情報集合中に出現する体験情報，ここでは空間属性，行動属性値の組合せ情報を抽出し，各組合せに対して認知度を付与する．空間 g と行動 x の組合せに対する認知度スコア $V(g,x)$ は，ある空間 g を検索クエリとして Web 検索した際に，行動 x を知る確率 $P(x|g)$ として，以下の式で定式化する．

$$V(g,x) = P(x|g) = \frac{1}{C} \sum_{i \in D_x^g} \frac{1}{r_i} \tag{4.6}$$

D_x^g は，空間 g を検索クエリとした検索結果中で行動 x を含むページ集合であり，r_i はページ $i \in D_x^g$ の検索結果中の順位である．検索結果中の順位の逆数を足し込んでいくことで，検索結果の上位に含まれる情報に対して高いスコアを与える．C は正規化項であり，以下の式で計算する．

$$C = \sum_{x' \in X} \sum_{i' \in D_{x'}^g} \frac{1}{r_{i'}} \tag{4.7}$$

X は行動 x の集合である．なお，Web ページがある行動を含むかを判断する際，完全マッチだけを扱うと該当する Web ページが極端に少なくなるため，動作対象レベルで判定する．たとえば，行動 “紅葉を見る” の構造化表現である {動作，対象} = {見る，紅葉} においては，“紅葉” を含むかどうかで判定する．

現在，サービス提供者は様々な媒体を通して広告を発信するが，特に，ユーザが能動的に自身の関心をキーワードとして入力するという性質から，効率良くターゲットにリーチするための媒体としてWeb検索エンジンへの関心が高まっている．一般に，検索エンジンでのキーワード検索結果として，上位ページと下位ページではクリック率に大きな差があるため，自社サイトの上位表示を目指すための最適化を実施する検索エンジン最適化（SEO）も，サービス提供者のマーケティング活動の一環として定着しつつある．本手法で計算される認知度は，このような広告コンテンツの影響を多く受けていると考えることができる

認知度とソーシャルメディアにおける出現傾向の比較

ある体験情報（空間属性，行動属性値の組合せ）に関して算出した認知度と，ソーシャルメディアにおける一般性の差異を定量化する．ここで，ソーシャルメディアにおける出現頻度の高さ，何人が体験したか，に基づく評価値を“一般性”と呼ぶ．認知度と一般性の傾向に差異があるほど，興味深い知識であるとして抽出する．

まず，認知度が高いが，一般性が低い体験情報を抽出したい場合が考えられる．これは，メディアへの情報露出が過多の状態であることを示唆する知識である．たとえば，ある空間 g と行動 x の組合せが，認知度が高いが実際にしている人が少ないことを示すスコア $Score(g,x)$ は以下の式で計算できる．

$$Score(g,x) = \frac{V(g,x)}{S(g,x)} \tag{4.8}$$

ここで，$V(g,x)$ は，ある空間 g で行動 x をする体験情報の認知度であり，$S(g,x)$ は，ある空間 g で行動 x をする体験情報の一般性である．たとえば，ある場所を訪れた人の中で何人がその行動をするかを示す一般性 $S(g,x)$ は，体験情報集合に基づいて計算した空間 g における行動 x の条件付き出現確率 $P^E(x|g)$ で計算できる．

$$S(g,x) = P^E(x|g) \tag{4.9}$$

認知度が高いほど，また，一般性が低いほど，数式（4.8）で計算した値は大きくなる．この値の降順で空間 g と行動 x の組合せをソートすることで，認知度が高すぎる，メディア露出が過多の情報を発見することが可能である．

逆に，認知度は低いが，一般性の高い体験情報を抽出したい場合も考えられる．この場合は，数式（4.8）の分母と分子を入れ替えれば良い．

$$Score(g,x) = \frac{S(g,x)}{V(g,x)} \tag{4.10}$$

数式（4.10）で計算した値の大きい体験情報を選定することにより，まだ認知度が低く，Web 検索エンジン上位の結果には出現しないが，ソーシャルメディアには存在する情報に目を向けることができる．ロングテールな体験や都市における新しいトレンドの発見につながることが期待される．また，ここでは，単純な定式化で認知度とソーシャルメディアにおける出現傾向の比較を行う方法を示したが，$V(g,x)$ と $S(g,x)$ の差を計算する方法など，ほかの計算式で両者のスコアの差異，乖離を評価しても良い．提案手法の基本アイディアは，認知度と一般性を比較し，その相対的な関係性を評価することである．

4.4　評価実験

4.4.1　実験 1: 状況に特徴的な行動情報抽出

データセット

解析を行ったブログ記事と抽出した体験情報の詳細を表 4.2 に示す．解析したブログ約 4,800 万件は，5ヶ月間分のデータであり，BLOGRANGER 2.0 API[1] を利用して収集した．ただし，これらのブログ記事中には，ニュース記事を引用したものが含まれている．これらの文書群から，体験に関係のないルールが生成されることを避けるために，ニュースサイト名をストップワードとして登録し，これらを含む文書は解析対象から外している．なお，1 記事当たりの平均抽出要素数は，時間属性を除くものであり，これを含めると，平均で 5.811 ということになる．また，今回の実験では，時間，空間，行動，感情という 4 属性間の共起頻度のみを算出した．時間属性に関しては，ブログを 5 日間づつ 60 区間に分け，それぞれの区

[1] http://ranger.labs.goo.ne.jp/TG/webapi.php/

表 4.2: 解析したブログ記事と抽出データに関する情報

記事数	48,112,100
収集期間	2007/1/1〜5/31
1 以上の要素が抽出できた記事数	29,778,231
1 記事当たりの平均抽出要素数	4.811
1 記事当たりの平均抽出要素数（行動属性）	3.777
1 記事当たりの平均抽出要素数（空間属性）	0.845
1 記事当たりの平均抽出要素数（感情属性）	0.259
空間属性値の種類数	34,932
行動属性値の種類数	664,914
感情属性値の種類数	121

間で，空間，行動，感情という 3 属性間の共起頻度を算出している．全区間トータルでの正確な共起頻度を算出することは膨大な計算量が必要となるため，各区間単独で，共起頻度が 3 以上の組合せを取得し，それらを集計することで計算量を削減した．

システムの実装

ブログ文書の形態素解析器には日本語形態素解析ソフト JTAG[59] を，空間属性値の抽出には多言語固有表現抽出器 Namelister[60] を使用した．固有表現抽出技術は，入力として与えられた文書から人名，地名や組織名といった固有表現を抽出する技術であり，固有表現抽出技術を用いて地名，組織名と判定された語を空間属性値として抽出する．前章で示した体験情報抽出手法においては，事前に空間属性値の候補語を人手で与える必要があった．本章では，クエリ非依存で収集したブログ記事，いわばブログ空間全体を対象とし，出現するすべての空間属性値を扱っている．アソシエーションルール抽出には，*Apriori* アルゴリズムを用いた [51, 52]．*Apriori* アルゴリズムは，ユーザが指定した最小支持度，最小確信度以上のルールを効率的に求める手法であり，頻出アイテムセットの導出，ルール導

出，の 2 ステップから成る．最初のステップで，最小支持度以上の支持度を持つアイテムの組合せを求め，次のステップで最小確信度以上の確信度を持つルールを生成する．本システムにおいては，ブログの収集，体験情報の抽出，体験情報の成功/失敗属性値の付与と，計算コストが高い最初のステップをバッチ処理にて行う．2 番目のステップと，ルールの興味深さの指標を用いたランキングは，ユーザとのインタラクションに応じて，リアルタイム処理にて行う．

得られるルールと客観的指標との関係性

支持度，確信度，提案法，の 3 つの指標で得られるルールについて考察する．人間の “主観” を含むタイプ 2 からタイプ 4 のルールは，定量的な評価が困難なため，タイプ 1 に該当する以下の形式のルールを用いて，それぞれの指標で得られる結果を比較する．なお，ルール抽出の際に設けた制約条件は以下の通りである．

[1] 空間 ⇒ 行動（動作，対象）

- 最小支持度: 1.00E-07
- 最小確信度: 1.00E-04
- 空間属性: お台場，ディズニーランド，横浜，沖縄，京都，大阪，北海道，名古屋
- 動作属性: 食べる，見る，買う
- ルール総数: 1,659

[2] 時間 ⇒ 行動（動作，対象）

- 最小支持度: 1.00E-05
- 最小確信度: 1.00E-04
- 時間属性: 2007 年 1 月，2 月，3 月，4 月，5 月
- 動作属性: 食べる，見る，買う
- ルール総数: 1,330

表 4.3: 支持度でルールをソートした結果の例

順位	ルール	支持度	確信度	提案法
1	京都 ⇒ 桜を見る	2.06E-05	9.79E-03	10.26
2	大阪 ⇒ ご飯を食べる	2.04E-05	7.71E-03	2.06
3	京都 ⇒ ご飯を食べる	1.48E-05	7.02E-03	1.88
4	沖縄 ⇒ 料理を食べる	1.05E-05	8.19E-03	14.30
5	横浜 ⇒ ご飯を食べる	9.97E-06	9.98E-03	2.67
6	沖縄 ⇒ そばを食べる	9.60E-06	7.49E-03	39.64
7	大阪 ⇒ 顔を見る	8.66E-06	3.27E-03	1.61
8	大阪 ⇒ 姿を見る	7.99E-06	3.02E-03	1.36
9	京都 ⇒ お土産を買う	7.99E-06	3.80E-03	10.30
10	ディズニーランド ⇒ パレードを見る	7.99E-06	1.23E-02	262.59

次に得られたルール集合を支持度，確信度，提案法の値の降順にソートする．[1] において，支持度，確信度，提案法の値でソートした場合の上位の結果をそれぞれ表 4.3，表 4.4，表 4.5 に示す．支持度はルールの条件部と結論部の単純な共起頻度に基づく評価尺度であるため，“京都 ⇒ 桜を見る” や “大阪 ⇒ ご飯を食べる” のような，一般的な組合せを評価する傾向にある．つまり，そもそも条件部と結論部の語の出現頻度が高い組合せを評価する．確信度はルールとしての信頼性を評価する尺度であり，条件部を満たした場合の結論部の生起確率を示す．つまり，確信度による評価値の高いルール “ディズニーランド ⇒ パレードを見る” は，ディズニーランドに行く人の多くはパレードを見ていることを意味する．しかし，確信度によるソートの上位の結果から分かるように，確信度は，“ご飯を食べる” といった，出現確率がそもそも高く，多くの場所で行われている行動を評価してしまう傾向にある．前述の通り，条件部を加えたときの結論部の出現確率が，全体集合におけるそもそもの出現確率と比較して，どの程度上昇しているかを評価するのが提案法である．実際，提案法による評価では “お台場 ⇒ ご飯を食べる” や “横浜 ⇒ ご飯を食べる” の評価値は低くなっている．提案法によるソートの上位の結果を見ると，結論部に，“スプリングロールを食べる” や “シンデレラ城を見る” など

表 4.4: 確信度でルールをソートした結果の例

順位	ルール	支持度	確信度	提案法
1	ディズニーランド ⇒ パレードを見る	7.99E-06	1.23E-02	262.59
2	お台場 ⇒ ご飯を食べる	5.27E-06	1.08E-02	2.87
3	横浜 ⇒ ご飯を食べる	9.97E-06	9.98E-03	2.67
4	京都 ⇒ 桜を見る	2.06E-05	9.79E-03	10.26
5	ディズニーランド ⇒ ショーを見る	5.57E-06	8.55E-03	36.21
6	ディズニーランド ⇒ ご飯を食べる	5.37E-06	8.24E-03	2.20
7	沖縄 ⇒ 料理を食べる	1.05E-05	8.19E-03	14.30
8	大阪 ⇒ ご飯を食べる	2.04E-05	7.71E-03	2.06
9	沖縄 ⇒ そばを食べる	9.60E-06	7.49E-03	39.64
10	名古屋 ⇒ ご飯を食べる	7.05E-06	7.25E-03	1.94

表 4.5: 提案法でルールをソートした結果の例

順位	ルール	支持度	確信度	提案法
1	ディズニーランド ⇒ スプリングロールを食べる	1.01E-07	1.55E-04	1533.62
2	ディズニーランド ⇒ シンデレラ城を見る	2.01E-07	3.09E-04	1533.62
3	ディズニーランド ⇒ エレクトリカルパレードを見る	1.01E-07	1.55E-04	1150.21
4	お台場 ⇒ ベール展を見る	1.34E-07	2.74E-04	815.84
5	お台場 ⇒ コルベール展を見る	1.34E-07	2.74E-04	815.84
6	ディズニーランド ⇒ スティッチのパレードを見る	3.36E-07	5.15E-04	766.81
7	お台場 ⇒ インポートカーショーを見る	1.01E-07	2.05E-04	764.85
8	お台場 ⇒ グレゴリーコルベール展を見る	1.01E-07	2.05E-04	679.87
9	お台場 ⇒ ｓｎｏｗを見る	2.35E-07	4.79E-04	648.97
10	お台場 ⇒ カーショーを見る	1.01E-07	2.05E-04	556.26

の出現確率がそれほど高くない行動に関するルールを評価していることが分かる．つまり，提案法により，一般的な行動に関するルールを除去し，ある空間（時間）的条件に特徴的に出現する行動を示すルールのみを抽出することができる．

次に，ある空間（時間）に特有な行動を得るための指標としての提案法の有用性を定量的に検証する．比較対象とする指標は，支持度，確信度に加え，χ^2 値，J-measure[58] の 4 つである．J-measure は，情報量に基づく指標であり，$P(A,B) * \log(\frac{P(B|A)}{P(B)}) + P(A,\bar{B}) * \log(\frac{P(¥ barB|A)}{P(\bar{B})})$ で計算する．この式から，J-measure は，支持度 $P(A,B)$ とリフト $\frac{P(B|A)}{P(B)}$ を統合した指標であることが分かる．[1] においては “空間に特有な行動（対象）”，[2] においては “時間に特有な行動（対象）” がルールの結論部として得られた場合に正解とする．正解セットは，人手で作成した．作成に関わった人数は 2 人であり，今回は，両者の意見が一致したもののみを正解としている．[1] の正解数は 129（ルール総数は 1,659）であり，[2] の正解数は 55（ルール総数は 1,330）である．それぞれの指標でルールをソートし，[1] においては上位 129 件の，[2] においては上位 55 件の適合率を評価した．適合率は，全ルールの中で，適合したルールの割合である．表 4.6 にその結果を示す．実験の結果，提案法がそのほかの指標よりも高精度に空間（時間）に特有な行動を抽出していることが分かる．特に，支持度，及び確信度との比較においては，大きな精度改善が見られる．また，χ^2 値によるルールの評価は，提案法に次ぐ精度を示した．χ^2 値は，条件部と結論部の方向性を考慮せず，その組合せの独立性を測る尺度である．具体的には，条件部の集合 A と結論部の集合 B が独立しており，同じトランザクションに含まれるのが単なる偶然であると仮定したときに期待される期待度数と，実際に観測されたトランザクション数（観測度数）との乖離を，$N\frac{(P(A,B)-P(A)P(B))^2}{P(A)P(B)(1-P(B))(1-P(A))}$ で計算する．もし，算出値の絶対値が大きければ，相関（correlation）が強いということができる．本実験においては，方向性を考慮した “状況 $\Rightarrow$ 行動” という形式のルールを評価対象としたが，条件部と結論部の方向性を問わず，その組合せの依存性，独立性の評価も重要な指標となることがわかった．J-measure は，3 番目に高い精度であった．J-measure は，支持度とリフトを統合した指標である．支持度による精度が低かったため，これらの影響が精度低下

表 4.6: ルールの評価尺度と適合率との関係

	支持度	確信度	提案法	χ^2 値	J-measure
[1]	0.085	0.109	0.643	0.596	0.357
[2]	0.018	0.018	0.727	0.618	0.491

表 4.7: ルール形式: {空間} ⇒ {感情} の一例

ルール	提案法	確信度
ディズニーランド ⇒ 喜び	1.166	0.583
ディズニーランド ⇒ 疲労	0.916	0.117
ディズニーランド ⇒ 不満	0.906	0.113
ディズニーランド ⇒ 不安	0.866	0.068
ディズニーランド ⇒ 悲しみ	0.740	0.043
ディズニーランド ⇒ 驚き	0.847	0.034
ディズニーランド ⇒ 困惑	0.606	0.027
ディズニーランド ⇒ 怒り	0.653	0.016

につながったと考えられる．複数の客観的指標を統合した指標は，適切な最小支持度と最小確信度を人手で設定する作業が不要であり，事前知識を必要しないという利点がある一方，今回の実験の様に，ある特定の指標が興味深いルールの発見に寄与していない場合には，精度を落とすという欠点がある．

同様のアプローチを用いて，人間の “主観” に関するルールの抽出結果について述べる．主観に関して得られたルールの一例を表 4.7，表 4.8 に示す．表 4.7 は，{ 空間 }⇒{ 感情 } という形式のルール，表 4.8 はその条件部に時間属性が加わった { 時間，空間 }⇒{ 感情 } という形式のルールであり，いずれも空間属性の値が “ディズニーランド” である．表 4.7 から，ディズニーランドがほかの場所に比べ，多くの人にとって “喜び” をもたらす場所であることが分かる．一方，表 4.8 から，5 月，3 月に訪れた場合には，“怒り” や “不満” が多く失敗の傾向が強いことが分

表 4.8: ルール形式: {時間，空間} ⇒ {感情} の一例

結果	ルール	提案法	確信度
失敗	5 月，ディズニーランド ⇒ 怒り	1.146	0.017
	3 月，ディズニーランド ⇒ 悲しみ	1.191	0.051
	1 月，ディズニーランド ⇒ 不満	1.112	0.126
成功	2 月，ディズニーランド ⇒ 喜び	1.023	0.600
	4 月，ディズニーランド ⇒ 喜び	1.013	0.591

かる．これらの背景には，3 月の春休みや 5 月のゴールデンウィークによる園内の混雑が関係していると直感的に理解できる．これらの結果は，ある場所が人間にもたらす感情が，時間属性値によって変化する例である．なお，空間，行動，感情属性値をすべて含むタイプ 4 のルールは，一部の有名な場所に関してのルールを除き，多くのルールを発見することができなかった．これは，空間，行動，感情属性値を 1 記事中に同時に含むものがそもそも少ないためである．2007 年 1 月 1 日〜1 月 5 日までの期間で，3 記事以上に出現した 3 属性間の組合せは 1,362 種類であり，2 属性間の組合せの 259,259 種類と比べて少なかった．今回の実験では 5ヶ月間分のブログデータを扱ったが，さらに抽出期間を広げ，解析データの量を増やすことでこの問題は解決可能だと考えている．

4.4.2 実験 2: Web 検索エンジンとソーシャルメディアの差異分析

データセット

ソーシャルメディア Twitter におけるジオタグ付きのツイートデータから体験情報抽出を行った．Twitter では，投稿記事にメタデータを付与してアップロードする機能を提供している．ジオタグ（緯度・経度）は，投稿記事に付与可能なメタデータの 1 つであり，GPS を搭載したデジタルカメラや携帯電話により自動的に付与されるため，各ユーザの時空情報を細粒度で取得できる利点がある．ツイートに付与されたジオタグ情報を空間属性の値，ツイートを投稿した日付情報を時

間属性の値とし，ツイートの本文情報から抽出した行動属性値をこれらの空間属性，時間属性の値に紐づけることで，時間，空間，行動属性値の組合せから構成される体験情報を抽出した．

以下，具体的なデータ収集方法についてを述べる．まず，日本国内における 12 個のランドマーク（六本木ヒルズ，東京スカイツリー，銀座，東京国際展示場，品川駅，横浜赤レンガ，羽田空港，鎌倉駅，平安神宮，河原町駅，京都駅，渡月橋）を選定し，各ランドマークの代表地点を中心とした半径 1km のエリアで投稿されたジオタグ付きのツイートデータを収集した．各ランドマーク周辺エリアを 1 つのデータセットとし，計 12 個のデータセットを生成した．なお，ジオタグ付きツイートの日付情報が 2012 年 10 月 1 日から 2013 年 5 月 9 日のものに限定して解析を行った．また，ジオタグ付きツイートから抽出した体験情報の空間属性値は緯度・経度情報であるが，地理情報辞書を用いて，可能な限り，駅名，店舗名などの地名に変換した．

行動の認知度と一般性評価

本実験では，各ランドマークを中心としたエリアごとに行動情報の認知度を評価した．旅行者があるエリアの情報収集を行う場合，エリア内で人気の場所（地名）を順に選択し，各場所（地名）を検索クエリとして Web 検索を実行すると考えられる．このシナリオに沿って行動情報の認知度を評価した．まず，体験情報集合に基づき，各エリアで出現数の多かった上位 30 個の空間属性値を取得する．次に，各空間属性値（地名）を検索クエリとして Web 検索を行い，検索結果上位 100 件にランキングされた Web ページのタイトル，及び，スニペットを取得する．取得した 3,000 件のページのタイトル，スニペット情報に基づいて，各エリアにおける行動 x の認知度を，

$$V(G,x) = \frac{1}{C}\sum_{g\in G}\sum_{i\in D_x^g}\frac{1}{r_i} \tag{4.11}$$

で計算した．G はエリアを代表する 30 個の空間属性値の集合，D_x^g は，空間 $g \in G$ を検索クエリとした検索結果中で行動 x を含むページ集合であり，r_i はページ $i \in D_x^g$

表 4.9: 正解ラベル付きデータ

エリア（データセット）	正解ラベル付き / 行動の種類数
六本木ヒルズ	174 / 517
東京スカイツリー	47 / 192
銀座	344 / 976
東京国際展示場	27 / 138
品川駅	104 / 388
横浜赤レンガ	88 / 290
羽田空港	59 / 222
鎌倉駅	48 / 170
平安神宮	25 / 106
河原町駅	86 / 336
京都駅	72 / 303
渡月橋	13 / 63

の検索結果中の順位である．C は正規化項であり，

$$C = \sum_{x' \in X} \sum_{g \in G} \sum_{i' \in D^g_{x'}} \frac{1}{r_{i'}} \tag{4.12}$$

で計算する．X は各エリアの行動属性値の集合である．検索結果上位の Web ページに含まれる行動ほど，また，多くの Web ページに含まれるほど認知度スコアが高くなる．さらに，複数の検索クエリ（空間属性値）に関する検索結果に含まれるほど，認知度スコアが高くなる．

また，各エリアにおける行動情報の一般性は，体験情報集合に基づいて計算した．単純に，あるエリアにおける行動 x の一般性は，各エリア内の空間属性値を含む体験情報の中で行動 x を含むものの割合とした．

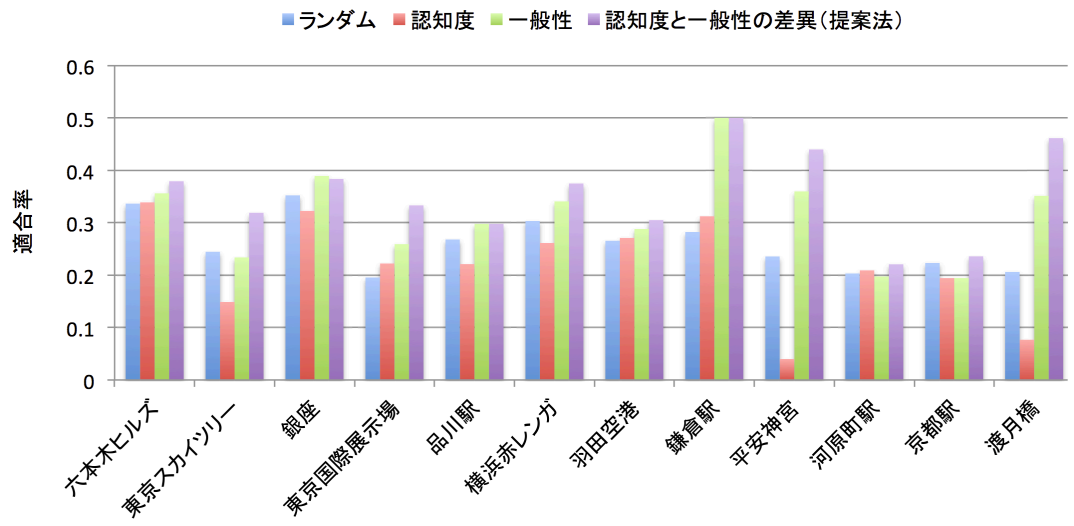


図 4.1: 適合率の比較

定量評価

本実験では，過去に観光であるエリアを訪れたことがあるユーザが“過去に聞いたことがない，あるいは体験したことはないが興味をもった行動” を探し出すタスクにおいて提案法の有効性を示す．この種の情報に対して適切に高いスコアを与えることができるかどうかで手法の有効性を評価する．提案法として用いたのは，数式（4.10）であり，認知度が低いほど，また，ソーシャルメディアで人気が高いほどスコアを与える手法である．ランダムに情報提示する手法，認知度に基づいて情報提示する手法，ソーシャルメディアにおける一般性に基づいて情報提示する手法との比較を行った．正解データは人手で作成した．各エリアを観光で訪問したことがある被験者を選び，行動情報のリストを提示する．提示情報の中で「過去に聞いたことがない，あるいは体験したことはないが興味をもった行動」に正解ラベルを付与するよう求めた．なお，各エリア（各データセット）で被験者 2 名でラベル付けを行い，両者の意見が一致したもののみを正解としている．各エリアで並び替えを行う行動属性値の種類数と，その中で正解ラベルが付与されたものの種類数を表 4.9 に示す．評価指標として用いたのは適合率である．各手法で行動情報をソートし，上位 K 件（K は正解数）の中で正解が含まれる割合で評

価した．たとえば，六本木ヒルズであれば，上位 174 件の適合率を評価する．図 4.1 にその結果を示す．12 個中 11 個のデータセット（エリア）で提案法が最も高い適合率を示した．また，12 個中 7 個のデータセットで認知度に基づく手法が最も低い適合率を示した．表 4.10 に，渡月橋周辺エリアにおける行動情報を各手法でソートした場合の上位の結果を示す．本実験においては，過去に聞いたことがない，あるいは過去に体験したことがないものを正解としている．認知度に基づく手法の適合率が低かったのは，“桜，見る” や “庭，見る” など渡月橋周辺エリアで有名で，観光客であれば誰もがする行動を上位にソートする傾向があるからである．この目的においては，むしろ，ランダムな情報提示のほうが認知度に基づく手法よりも高い適合率を示す傾向にある．その一方で，人々の実際行動に基づく一般性に基づく手法や，認知度が低いが一般性の高い行動を評価する提案手法は，“ゆばプリン，食べる” や “ジオラマ，見る” などロングテールな行動を評価する傾向があった．これらの情報は，最近，その場所を訪れた人々が偶然的に発見し，ソーシャルメディア上で共有したものであると考えられる．なお，認知度の上位結果と一般性の上位結果が異なる場合，提案法の上位結果は一般性による結果とほぼ等しくなる．一般性と認知度がほぼ同じ値を示しているデータセットが散見されるのはそのためである．本実験により，過去に観光であるエリアを訪れたことがあるユーザにとって，ソーシャルメディアは観光ガイドブックやニュースなどの従来型のメディアには存在しない多様な体験情報を知るための重要な情報源であること，また，そのような従来型のメディアコンテンツを頻度高く含む Web 検索エンジン上位の結果と対比させることで，より効率的にソーシャルメディアに特徴的な傾向を発見できることを示した．

4.5 結言

本章では，ソーシャルメディアに存在する人々の体験情報を分析することで得られる有用な知識とは何かを議論し，その知識を抽出するための仕組みを提案した．具体的には，2 つの利用シナリオに沿って “ある特定の状況（時間，空間）に

表 4.10: 渡月橋周辺エリアで得られた行動情報を各手法でソートした結果

認知度	一般性	認知度と一般性の差異
順位. 行動（対象，動作）	行動（対象，動作）	行動（対象，動作）
1. 桜, 見る	1. 紅葉, 見る	1. 紅葉, 見る
2. 桜餅, 買う	2. 蕎麦, 食べる	2. 送り火, 見る
3. 庭, 見る	3. 桜, 見る	3. 豆腐ソフトクリーム, 食べる
4. 景色, 見る	4. うどん, 食べる	4. ゆばプリン, 食べる
5. デザート, 食べる	5. 送り火, 見る	5. 松花堂弁当, 食べる
6. 湯豆腐, 食べる	6. 庭, 見る	6. 八つ橋入りどら焼き, 買う
7. 蕎麦, 食べる	7. ゆばプリン, 食べる	7. ジオラマ, 見る
8. パンケーキ, 食べる	8. ジオラマ, 見る	8. 鮎の塩焼き, 食べる
9. 天龍寺パフェ, 食べる	9. 松花堂弁当, 食べる	9. 流れ星, 見る
10. うどん, 食べる	10. 八つ橋入りどら焼き, 買う	10. 嵯峨野湯カレー, 食べる

おいて人々が特徴的にしている行動発見” と “人々に認知されている行動と都市で実際に人々がしている行動の差異発見” につながるアソシエーションルールを興味深い知識として抽出する手法を示した．2 つの提案手法に共通する点は，ほかのルール，ほかのメディアにおける傾向と比較分析を行い，ある知識（アソシエーションルール）の価値を大局的に判定する点にある．評価実験においては，ブログデータや Twitter データから実際に知識抽出を行い，体験情報の利用者にとって興味深い知識を効率的に選択できることを示した．

今後は，情報提供者側の観点と消費者側の観点の差異分析を深めていく．本研究では，主に，都市の生活者が，いつ，どこで，何をしたかの情報（時間，空間，行動属性）に焦点を当てたが，人間の主観情報を含めて消費者側の観点をより精緻に分析していく必要がある．既存の評判情報抽出技術を利用して評価属性値の抽出に取り組むと同時に，評価属性値を用いた感情属性値推定，成功/失敗属性値の付与を試みる．前述の通り，体験が成功したか失敗したかの判断は，動作主の主観（“評価” と動作主が抱いた “感情”），特に感情に因るところが大きい．しかし，動作主が明示的に自らの感情をテキストに述べていない，つまり，感情属性値をテキストから抽出不可能な場合や，成功/失敗以外とした感情カテゴリが “驚

き” に含まれる語が存在する場合に対処していくには，感情属性同様に動作主の主観を示す表現である評価属性の値から，感情属性や成功/失敗属性の値を推定する処理が必要である．また，本論文においては，評価属性を “動作をともなう対象に対する評価” と定義した．この定義によれば，“昨日，清水寺に紅葉を見に行きました．確かに紅葉はきれいだったのですが，観光客で大混雑でした．がっかりです．” という例文から，動作 “見る” をともなう対象である “紅葉” に対する評価 “きれい” は表現可能だが，空間属性値である “清水寺” の状態を述べている “大混雑” は表現できないことになる．“大混雑” という表現は，この例が失敗に終わった要因となり得る情報である．本論文は，評価属性抽出にフォーカスしたものではなく扱ってはいないが，今後，このような種類の表現にも対応していく必要がある．情報提供者側の観点で発信された情報についても，Web 検索エンジンの検索結果だけでなく，そこに表示されるオンライン広告や，実世界に存在する各種広告情報などを広く分析し，情報の認知度推定の精度向上を試みる．

第5章　写真共有サイトのジオタグ情報を利用したトラベルルート推薦

5.1　緒言

本章では，過去に都市を訪れた人々の体験情報を，新たに都市を訪れたユーザ行動の自動拡張のために利活用する仕組みを構築する．具体的には，写真共有サイトのジオタグ情報を利用したトラベルルート推薦手法を提案する．小型のデジタルカメラやカメラ付きの携帯電話の普及により，Flickr[1] や Google Picasa[2] などの写真共有サイトに関心が集まっている．これらのサイトは，写真にタグ情報を付与してアップロードする機能を提供しており，タグ情報付きの写真を大量に集めることに成功している．ジオタグ（緯度・経度）は，写真に付与可能なメタデータの 1 つであり，GPS を搭載したデジタルカメラや携帯電話により自動的に，もしくは，写真共有サイトの投稿機能の中でユーザが指定することで付与される．空間に関連づけられたコンテンツのナビゲーションにおいて，ジオタグは重要なメタデータとして利用されてきた．ローカル画像検索 [37, 38]，写真撮影位置の推定 [44, 45, 46]，ローカルコンテンツのブラウジング [39, 40, 41, 42, 43] など，ジオタグを利用した研究もさかんに行われている．

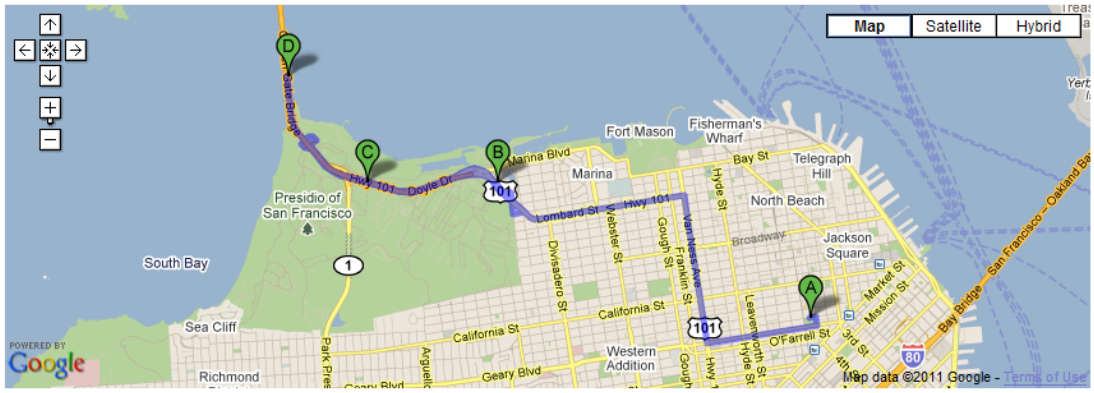
本研究では，主に Web 上のコンテンツナビゲーションに用いられてきたジオタグ情報の新たな利用方法として，実世界におけるトラベルルート推薦に活用する

[1] http://www.flickr.com/
[2] http://picasa.google.com/

手法を提案する．各ユーザの写真集合を時間情報でソートすると，ジオタグ付き写真集合は，個人の移動履歴とみなすことができる．このジオタグ情報に基づく移動履歴の重要性は，“質” と “量” の観点で論じることができる．まず，“写真を撮る” 行為は実世界のある場所に対する投票とみなすことができる．つまり，訪れる価値のある場所を効率的に発見することができる質の高い情報源である．さらに，我々の調査によれば，2010 年の時点で，400,000 人以上が投稿した 40,000,000 以上のジオタグ付き写真が Flickr には投稿されている．これは，広く利用可能な移動履歴データの中でも最大規模のものであり，観光客から居住者まで，様々な背景知識を持つ人々の興味に合致する地域情報が高い確率で含まれていると考える．前章までの試みにおいては，ブログ情報に記述された体験情報に焦点を当ててきた．ブログ情報には，行動内容が記述されているメリットがある反面，時間属性，空間属性情報の粒度が荒い問題があった．そのため，ソーシャルメディアの中でも，旅行者の時間，空間属性が細粒度で得られ，かつ，大規模なものとして，写真共有サイトのジオタグ情報を利用した．

提案手法は，旅行者の移動履歴と，旅行に費やすことができる空き時間が与えられた条件のもと，旅行者の現在地，空き時間，興味に合致したトラベルルートを推薦する．具体的には，移動履歴からのユーザ行動モデルの学習と，生成した行動モデルに基づく推薦ルート生成を本研究で扱う技術課題とする．ユーザ行動のモデル化においては，場所間の移動しやすさを考慮するマルコフモデルと，旅行者の興味を考慮するトピックモデルとを，確率論的枠組みの中で結合することで，旅行者が次に行く場所を予測する *Photographer Behavior Model* を提案する．推薦ルートの生成においては，グラフの最良優先探索アルゴリズムをベースとした手法により，効率的に現在地からの推薦ルートを生成する手法を述べる．また，移動履歴からのユーザ興味推定の高度化を目的として従来のトピックモデルを改良したジオトピックモデルについても述べる．

提案手法に基づいてユーザの旅行計画を支援するトラベルルート推薦システムを開発した．図 5.1 にトラベルルート推薦システムのユーザインタフェースを示す．ユーザが，現在地情報を含む移動履歴，旅行に費やすことができる空き時間，移



1, 2, 3, 4, 5, hours travel in sanfrancisco

Your spare time: 3 hours

Your travel history: -

Your present location: unionsquare

A mode of transportation: public transportation

1: →[63 min]→exploratorium→[29 min]→crissyfield→[28 min]→goldengatebridge

2: →[33 min]→pier39→[42 min]→alcatraz→[55 min]→lombardstreet

3: →[35 min]→ferrybuilding→[50 min]→lombardstreet→[44 min]→goldengatebridge

4: →[43 min]→lombardstreet→[41 min]→exploratorium→[29 min]→crissyfield→[28 min]→goldengatebridge

5: →[40 min]→attpark→[50 min]→pier39→[37 min]→lombardstreet

図 5.1: トラベルルート推薦システムのユーザインタフェース

動手段などを指定すると，トラベルルート推薦システムは，ユーザの興味に合致したトラベルルートを自動提示する．各トラベルルートは，ランドマークの系列と，ランドマーク間の移動時間を含む情報とから構成される．徒歩，公共交通機関，自動車などのユーザが指定した移動手段に応じて，ランドマーク間の移動時間は自動推定される．ユーザが選択する可能性が高い順にトラベルルートはソートされており，ユーザの選択確率は提案するユーザ行動モデルに基づいて推定されたものである．

表 5.1: トラベルルート推薦手法の説明に用いる記号

Symbol	Description
$\boldsymbol{U}$	ユーザ集合
u	ユーザ, $u \in \boldsymbol{U}$
l_i^u	ユーザ u の i 番目の写真の位置情報（緯度・経度）
t_i^u	ユーザ u の i 番目の写真の時間情報
v_i^u	ユーザ u の i 番目の写真に付与されたタグ集合
τ^u	ユーザ u の移動履歴内の位置情報数
h^u	ユーザ u の移動履歴．移動履歴内の位置情報は
	日付情報でソートされており，$h^u = \langle l_1^u, \cdots, l_{\tau^u}^u \rangle$
K	推薦ルート数
T^{uk}	k 番目の推薦ルート内の位置情報数

5.2　トラベルルート推薦手法

5.2.1　問題定義

写真共有サイトのジオタグ情報は，実世界のランドマーク及び，ランドマーク間のルートを発見する上で重要な情報源である．この情報源から行動モデルを学習し，将来的にその場所を訪れる旅行者への情報推薦に利活用する．我々は，旅行者が次にどこを訪れるかを決定するプロセスをモデル化することでトラベルルート推薦を実現する．

本論文内で用いる表記を表 5.1 にまとめた．ユーザ集合 $\boldsymbol{U}$ 内のユーザ u に関する，位置情報と時間情報が付与された写真集合を $\{(l_i^u, t_i^u, v_i^u)\}_{i=1}^{\tau^u}$ とする．取り組む研究課題は，ユーザ u に関する移動履歴 h^u と，ユーザが旅行に費やすことが可能な空き時間 d が与えられたときに，位置情報から成るシーケンスであるトラベルルート $\{\langle l_{\tau^u+1}^{uk}, \cdots, l_{\tau^u+T^{uk}}^{uk} \rangle\}_{k=1}^{K}$ を推薦することである．推薦するトラベルルートの所要時間は，空き時間 d を満たす必要がある．l_i^u はユーザ u の i 番目の写真の

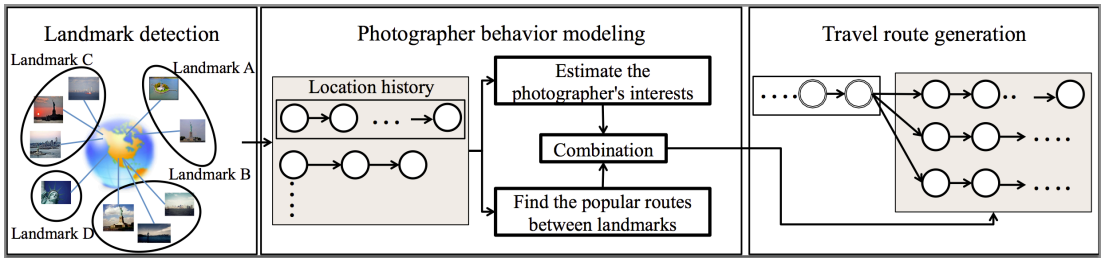


図 5.2: トラベルルート推薦の処理

位置情報を示すが，被写体の位置情報を示すものでもある．また，写真を撮影したデバイスの時計が正しく設定されていない可能性があるが，写真撮影時間の間隔は信頼できる情報として利用する．

提案法の処理フローを図 5.2 に示す．最初に，入力として与えられた全ユーザに関する位置情報集合から，多くのユーザが撮影したランドマークを抽出する．ランドマークとは，建物，寺社，店舗など，旅行の目的地となる実世界の対象であると定義する．このランドマーク情報の中から，最終的にユーザに推薦/提示するものが選択される．前述の通り，写真を撮る行為は，実世界に対する投票とみなすことができるため，投票によって選ばれたランドマークは，推薦候補として相応しいと考える．次に，ユーザが将来的に各ランドマークを訪れる確率を推定するためにユーザ行動をモデル化する．提案モデルは，次に訪れるランドマークが

[1] 現在地や最近訪れた場所とによって決まる
[2] ユーザ固有の興味によって決まる

という，2 つの仮説に基づいている．たとえば，現在，ニューヨークのタイムズスクエアにいる人は，サンフランシスコのゴールデンゲートブリッジよりも，同じニューヨークにある自由の女神を訪れやすい．また，アートに興味がある人はメトロポリタンミュージアムを，スポーツに興味がある人は，ヤンキースタジアムを訪れやすいと考えられる．最後に，学習した確率的行動モデルの予測値に基づいて，ユーザが選択する可能性が高いトラベルルートを自動生成する．以降，各ステップの詳細について述べる．

5.2.2　Step 1: Mean-shift 法によるランドマーク抽出

全ユーザに関して得られた，緯度・経度から成る位置情報集合から，多くのユーザが撮影したランドマーク情報を抽出する．観光地，店舗，寺社，橋脚など，都市を代表する特定の場所をランドマークと定義する．たとえば，ニューヨークにおいては，ユニオンスクエア，自由の女神，セントラルパークが，京都においては，京都駅，清水寺，三条大橋がランドマークに該当する．

本研究においては，Mean-shift 法を用いてランドマーク抽出を行う．Mean-shift 法は，与えられた二次元空間（緯度・経度）のデータ点集合で定義されるカーネル密度関数の最頻値（mode）を自動推定する手法である．ここでいう最頻値は，密度関数の極大値（局所的な最大値）として定義され，対応する極大値点は，空間上で点の密度が局所的に最も高いところを示す．つまり，推定した最頻値（の集合）が，多くのフォトグラファーが撮影したランドマーク（の集合）に対応する．Mean-shift 法は，主に，画像領域分割や物体追跡などで広く用いられているが，Crandall らが，緯度・経度から成る空間データに対しても適用可能性が高いことを示している [41].

位置情報 l が与えられときに，Mean-shift ベクトルを，

$$m_{w,g}(l) = \frac{\sum_u \sum_i l_i^u g\left(\left\| \left(l - l_i^u\right)/w \right\|^2\right)}{\sum_u \sum_i g\left(\left\| \left(l - l_i^u\right)/w \right\|^2\right)} - l \tag{5.1}$$

で求める．g は，採用したカーネル関数によるデータ点に対する重みである．本研究ではカーネル関数としてガウシアンカーネルを用いた．w はガウシアンカーネルの平滑化パラメータ（bandwidth）である．各データ点 $l^{(1)}$ の位置情報を，Mean shift ベクトルによって，

$$l^{(c+1)} = l^{(c)} + m_{w,g}(l^{(c)}) \tag{5.2}$$

で繰り返し更新する．Mean-shift ベクトルの値がゼロに近づくと同時に，各データ点は最頻値に収束する．各ランドマーク（最頻値）は，周辺に存在する複数のデータ点を代表する点である．また，各データ点が収束したランドマーク（最頻値）を，そのデータ点が所属するクラスタ（の中心）とするクラスタリング法を

Mean-shift クラスタリングという．Mean-shift クラスタリングは，既存のクラスタリング手法とは異なり，あらかじめクラスタ数を設定する必要がない．また，母集団の分布について仮定を設けないノンパラメトリックな手法であり，指定するパラメータは bandwidth のみという実用上のメリットがある．

次に，抽出したランドマーク集合と各データ点のクラスタリング結果に基づき，位置情報から成る各移動履歴 h^u を，以下のように変換する．

[1] Mean-shift クラスタリングの結果に基づき，移動履歴内の各位置情報を，所属するランドマーク（クラスタ）へと変換する（以降，位置情報と同様に，l_i^u を，ユーザ u が i 番目に訪れたランドマークを表す表記として用いる）

[2] 上記 1 を適用後，移動履歴内の連続するランドマークは 1 つに集約する．連続するランドマークは，ユーザが同一のランドマーク付近で連続して写真を撮影したことを意味する．

[3] 上記 2 を適用した場合，ユーザがランドマークを訪れた時間と，出発した時間の中央値を，集約したランドマークの新たな時間情報とする

以上より，Step 1 の結果は，ランドマーク集合と，変換後の移動履歴集合となる．

5.2.3 Step 2: Photographer Behavior Model

時間 $t-1$ に，ランドマーク l_{t-1} にいるユーザ u が，時間 t にランドマーク l_t を訪れる確率 $P(l_t|l_{t-1}, h^u)$ を推定する．$P(l_t|l_{t-1}, h^u)$ はユーザが場所を選択する原理を表現する数理モデルである．ユーザが次に訪れる場所は，現在地からのアクセスしやすさと，個人の興味とによって決まると仮定し，それぞれの関係を，マルコフモデルとトピックモデルを用いてモデル化する．さらに，マルコフモデルとトピックモデルを確率論的枠組みの中で組合せることでユーザ行動をモデル化する．

マルコフモデル

マルコフモデルは時系列情報を扱う確率モデルとして広く用いられる．1 次マルコフモデルの場合，次に訪れるランドマークは，直前に訪れたランドマークに依存し，

$$P(l_t|l_{t-1}, l_{t-2}, \cdots, l_1) = P(l_t|l_{t-1}) \tag{5.3}$$

で求まる．$P(l_t|l_{t-1})$ は最尤推定式,

$$P(l_t|l_{t-1}) = \frac{N(l_{t-1}, l_t)}{N(l_{t-1})} \tag{5.4}$$

で求まる．$N(l_{t-1}, l_t)$ は，データ中において l_{t-1} の直後に l_t を訪れた観測数であり，$N(l_{t-1})$ は，l_{t-1} を訪れた観測数である．なお，説明の簡略化のため，1 次マルコフモデルを用いて説明を行ったが，ほかの次数のマルコフモデルを用いてもよい．

トピックモデル

トピックモデルは，各ユーザ（文書）はユーザ（文書）固有のトピック比率に基づいて潜在トピックを選択し，各潜在トピックは潜在トピック固有の選択確率によってランドマーク（単語）を選択する，とした確率モデルである．トピックモデルはこれまで，情報検索分野 [61, 62]，購買ログ分析におけるユーザの興味分析 [63, 64] などに利用されている．実世界の移動履歴に対して適用することで，“アート” や “自然” などの，実世界における人々の興味対象が潜在トピックとして推定されることが期待できる．

Z をトピック数，$\boldsymbol{Z}$ をトピック集合とする．潜在変数 $z \in \boldsymbol{Z}$ が与えられたもとで，h^u と l_t が独立であると仮定した場合，トピックモデルにおいては，移動履歴 h^u のユーザがランドマーク l_t を訪れる確率を

$$p(l_t|h^u) = \sum_{z \in \boldsymbol{Z}} P(z|h^u) P(l_t|z) \tag{5.5}$$

で求める．ユーザ u がトピック z に興味を持つ確率 $P(z|h^u)$ はユーザの “興味” を表し，トピック z がランドマーク l を選択する確率 $P(l_t|z)$ はトピックの “トレンド” を表す．

Latent Semantic Analysis（LSA）, *Probabilistic Latent Semantic Analysis*（PLSA）[62] や，*Latent Dirichlet Allocation*（LDA）[61] など，これまでいくつかの潜在トピック分析手法が提案されてきた．本手法においては，PLSA を用いてトピック推定を行う．

EM アルゴリズムを用いて，$z \in \boldsymbol{Z}$ と $u \in \boldsymbol{U}$ に関するトピック比率 $P(z|h^u)$ と，$l \in \boldsymbol{L}$ と $z \in \boldsymbol{Z}$ に関するランドマーク出現確率 $P(l|z)$ とを推定する．$\boldsymbol{L}$ はランドマーク集合である．E ステップでは，ベイズ則に従い，トピック事後確率を，

$$P(z|l_t, h^u) = \frac{P(z|h^u)P(l_t|z)}{\sum_{z' \in \boldsymbol{Z}} P(z'|h^u)P(l_t|z')} \tag{5.6}$$

で計算する．M ステップでは，尤度関数を最大化するために以下のようにパラメータを更新する．

$$P(z|h^u) \propto \sum_{l \in \boldsymbol{L}} N(l, h^u)P(z|l, h^u) \tag{5.7}$$

$$P(l|z) \propto \sum_{u \in \boldsymbol{U}} N(l, h^u)P(z|l, h^u) \tag{5.8}$$

$N(l, h^u)$ は移動履歴 h^u でランドマーク l が出現する回数であり，U はユーザ集合である．収束するまで E ステップと M ステップを交互に繰り返すことにより，観測データの尤度を最大化する最適解を得ることができる．

マルコフモデルとトピックモデルの融合

マルコフモデルとトピックモデルを確率的な枠組みの中で組合せることで，現在地と興味を考慮した *Photographer Behavior Model*（以降，PBM）を生成する．l_{t-1} が与えられたとき，h^u と l_{t-1} が条件付き独立という仮定のもとで，以下の式で 2 つの確率モデルを融合する．

$$P(l_t|l_{t-1}, h^u) = \frac{P(l_t|l_{t-1})}{C(l_{t-1}, h^u)} \frac{P(l_t|h^u)}{P(l_t)} \tag{5.9}$$

$P(l_t|l_{t-1})$ と，$P(l_t|h^u)$ は，それぞれマルコフモデルとトピックモデルによって導き出された確率値であり，$C(l_{t-1}, h^u)$ は正規化項である．この結合式はユニグラム

リスケーリングと呼ばれ，確率的言語モデルの 1 つである [65]．$P(l_t)$ は，

$$P(l_t) = \frac{N(l_t)}{N} \tag{5.10}$$

で求まる．N は全写真数である．

5.2.4　Step 3: トラベルルート生成

前節において，現在地と興味情報とから，ユーザが次に訪れるランドマークを予測する PBM について述べた．本節においては，現在地からのトラベルルートを効率的に求める手法を述べる．トラベルルートは，ランドマークのシーケンス（及び，ランドマーク間の移動時間とから構成される情報）である．

ユーザ u が与えられたもとで，

$$P(\langle l^{ux}_{\tau^u+1}, \cdots, l^{ux}_{\tau^u+T^{ux}} \rangle | l^u_{\tau^u}, h^u) \tag{5.11}$$

の値が高い X 個のトラベルルート $\{\langle l^{ux}_{\tau^u+1}, \cdots, l^{ux}_{\tau^u+T^{ux}} \rangle\}_{x=1}^{X}$ を生成する．推薦されるルートは，ユーザの現在地と興味に合致したものである必要がある．最も単純には，ランドマーク間のすべての組合せを考慮し，取り得る全ルートに関する選択確率を求めた後，確率の高い X 個のトラベルルートを選ぶ方法が考えられる．しかし，この方法は計算量が膨大となるため，インタラクティブな推薦システムで用いるのは現実的ではない．そこで，最良優先探索アルゴリズム [66] のアイディアをもとに，効率的にトラベルルートを生成する．

トラベルルートを生成するアルゴリズムをアルゴリズム 1 に示す．入力は，ユーザ u の旅行履歴 h^u，旅行に費やすことができる空き時間 d とその許容範囲 ϵ，推薦して欲しいルートの個数 X である．出力は，X 個のトラベルルートを格納した配列である．s はランドマークの系列，d^s は s の旅行時間，p^s は s が選択される確率，s_{last} は s 内で最後に訪れたランドマーク，s_{+l} は，ランドマーク l を新たに訪れた場合に更新されたランドマーク系列である．

最初に，現在地 $l^u_{\tau^u}$ を優先度付きキュー Q に挿入する（4 行目）．ポップ操作により，優先度付きキューは最も優先度の高い単一の要素を返す．ここでは，各要

素の優先度が確率値 p^s で与えられた特別な優先度付きキューを用いる．なお，優先度付きキューは *max heap* を用いて実装した．次に，Q から最も高い確率値を持つ系列 s を取得し，その s が指定の時間制約条件を満たすかどうかをチェックする（6-10 行目）．もし，条件を満たさなかった場合，s の現在地（s 中で最後に訪れたランドマーク）からほかのランドマークへのルートを新たに展開し，系列 s を更新する (11-17 行目)．更新後の系列が選択される確率値は PBM から（13 行目），更新後の系列の旅行時間は旅行時間の推定結果から（14 行目），それぞれ新たに計算する．$TravelTime$ は，任意の 2 つのランドマーク間の旅行時間を返す関数であり，その推定方法に関しては 5.2.5 項で説明する．以上のプロセスを，出力とするトラベルルート数が X に達するまで繰り返す（18 行目）．

$P(l|s_{\text{last}}, h^u)$ は常に 1 以下であるため，$p^{s+l} \leq p^s \times P(l|s_{\text{last}}, h^u)$ である．したがって，本アルゴリズムが発見するランドマーク系列は，制約条件を満たすすべてのシーケンスの中で，確率の高いものから X 個のシーケンスであることを保証する．

5.2.5 移動時間の推定

あるランドマークからほかのランドマークへ移動する際にかかる移動時間は，徒歩，公共交通機関，自転車や自家用車の利用など，移動手段によって異なる．たとえば，徒歩で旅行をしたいと考えるユーザに対しては，徒歩圏内に存在するランドマークを推薦する，また，多くのランドマークを含むトラベルルートを推薦しないなど，ユーザの移動手段に応じて推薦する情報を変化させることが望ましい．そこで，写真共有サイト上の移動履歴の時間情報をもとに，ランドマーク間の典型的な移動手段ごとの期待移動時間を推定する．写真共有サイト上には，徒歩，バスや電車の利用，自転車や自動車の利用など，様々な種類の移動手段を利用した場合の移動履歴が混在すると考えられる．本研究では，ランドマーク間には K 個の異なる移動手段が存在すると仮定し，K-means クラスタリングを用いて K 個の代表的な移動手段に関する移動時間を推定する．

あるランドマーク l^{from} からほかのランドマーク l^{to} へのトラベルルートの移動

Algorithm 1 Generate travel routes

Require: $X > 0$, $d > 0$ and $\epsilon > 0$

Set an array $\boldsymbol{A} \leftarrow \boldsymbol{\Phi}$
Set $x \leftarrow 0$
Set a priority queue $\boldsymbol{Q} \leftarrow \boldsymbol{\Phi}$
Insert $l^u_{\tau^u}$ into $\boldsymbol{Q}$
repeat
  $s \leftarrow$ get the highest-probability one from $\boldsymbol{Q}$
  if $d - \epsilon \leq d^s \leq d + \epsilon$ **then**
    Push s into A
    $x \leftarrow x + 1$
  end if
  if $d^s < d + \epsilon$ **then**
    for $l \in \boldsymbol{L}$ **do**
      Set $p^{s+l} \leftarrow p^s \times P(l|s_{\text{last}}, h^u)$
      Set $d^{s+l} \leftarrow d^s + TravelTime_{s_{\text{last}},l}$
      Insert s_{+l} into $\boldsymbol{Q}$
    end for
  end if
until $x = X$
Output $\boldsymbol{A}$

時間を推定したいとする．システムは，l^{from} の次に l^{to} を訪れた移動履歴を抽出した後，各ランドマークに付与された時間情報をもとに，その時間間隔を求める．観測データに含まれる l^{from} から l^{to} への時間間隔データをすべて抽出した後，各観測値を 1 つのデータポイントとみなし，K-means クラスタリングを適用する．得られた K 個のクラスタの代表点を，あるランドマーク l^{from} からほかのランドマーク l^{to} へ移動する際の代表的な移動時間とみなす．得られた移動時間の推定値はアルゴリズム 1 の 14 行目において，トラベルルートの総旅行時間を計算する際に用いる．

5.3 トラベルルート推薦手法の評価実験

写真共有サイト Flickr のジオタグ情報を用いて，提案法の精度評価を行った．実験 1 では，提案法のパラメータと行動予測精度との関係を分析する．実験 2 では，ユーザが次に訪れる単一のランドマークを予測するタスクに関して，提案法と比較手法との精度比較を行う．実験 3 では，ユーザが次に訪れるランドマークの系列を予測するタスクに関して，同様の評価を行う．実験 4 では，旅行時間推定によって得られた結果を考察する．最後に，提案法を実装したシステムが推薦するトラベルルートの例を示す．

5.3.1 データセット

Flickr API を利用し，2006 年 1 月 1 日から 2009 年 6 月 31 日までにアメリカ西海岸，及び東海岸で撮影された写真のジオタグ情報を収集した．収集データは，71,718 ユーザに関する 696,394 枚の写真のジオタグデータである．

以下，データ収集方法の詳細について述べる．まず，東海岸，西海岸の各地域における主要都市であるワシントン D.C.，ニューヨーク，フィラデルフィア，ボストン（以上，東海岸地域），ロサンゼルス，サンフランシスコ，ラスベガス（以上，西海岸地域）を選定し，各都市の中心部から 20km 以内で撮影された写真のジ

表 5.2: 単一ランドマークの予測実験に用いた移動履歴情報

Dataset	Region-scale	The number of sequences	The number of landmarks	Average length of sequences
(a)	East - 50m	9,267	414	14.02
(b)	West - 50m	6,450	316	14.05
(c)	East - 10m	11,354	1,419	15.82
(d)	West - 10m	7,913	1,119	15.65

オタグ情報を収集した．Flickr API は，緯度・経度と半径を指定するとその領域内のジオタグ情報を返す関数を提供している．これらの都市は，アメリカ国内において，特にジオタグ付き画像が数多く投稿されている都市として報告されている [41]．これらの都市を中心とした領域を集中的にクローリングすることで，効率的に東海岸，及び西海岸地域のジオタグ情報を収集した．

次に，Mean-shift 法により，東海岸，及び西海岸地域のランドマーク抽出を行った．Mean-shift 法の bandwidth パラメータ w としては 2 つの値（0.0001（10m），0.0005（50m））を設定した．w を小さい値に設定するほど，細かい粒度のランドマークを検出することができ，その結果，ランドマークの種類数も多くなる．最後に，地域と bandwidth パラメータの組合せごとに，ユーザの移動履歴を生成した．生成したデータセットにおけるユーザ数（移動履歴数），ランドマーク数の詳細を表 5.2 に示す．ただし，ランドマーク数が 3 未満の移動履歴，及び，訪れたユーザ数が 3 未満のランドマークは除いた．

5.3.2　実験 1: パラメータの影響

実験 1 では，トピック数が提案法の精度に与える影響について述べる．提案法はトピックモデルを取り入れた手法であるため，トピックモデルのパラメータである潜在トピック数に影響を受ける．

表 5.3: 実験によって得られた最適パラメータの値（潜在トピック数）

Dataset	Region-scale	Topic model	Markov-topic model
(a)	East - 50m	9	9
(b)	West - 50m	8	10
(c)	East - 10m	9	15
(d)	West - 10m	15	15

提案する PBM はユーザが次に訪れるランドマークを予測するモデルである．そのため，各ユーザが最後に訪れたランドマークをそれ以前の移動履歴から予測するタスクにおいて精度評価を行った．評価指標として用いたのは適合率であり，全テストデータ中で予測が成功したユーザの割合で定義される．情報推薦システムのサーベイによると，予測タスクにおける適合率は情報推薦技術の最も一般的な評価指標である [47]．なお，今回は，5 分割交差検定を実施した．つまり，5 分の 1 のユーザの移動履歴をテストデータとし，残りのユーザの移動履歴を学習データとして用いる試行を，各サブセットごとに 5 回，繰り返し行った．最終的な評価値は，5 回の試行における適合率の平均である．

本実験では，表 5.2 の各データセットごとに，最適なパラメータ（潜在トピック数）を決定した．その評価結果を 5.3 に示す．図の X 軸は潜在トピック数，Y 軸は 5 回の試行における適合率の平均である．Topic がトピックモデルの，Markov-topic（PBM）が提案法の結果である．潜在トピック数は 3 から 9 まで，10 から 50 まで 5 刻みで変化させた．トピックモデルについても，まったく同じ条件で適合率を求めている．提案法の適合率はトピックモデルと同様の形状をとる．10 前後の潜在トピック数で最大値をとり，それ以上になると適合率は減少する傾向にある．この結果は，提案法の精度が，トピックモデルの精度に大きく依存することを示する．また，適切なパラメータの選定が精度改善に重要であることを示している．本実験で得られた最適パラメータの値を表 5.3 に示す．以降の実験においては，提案法，トピックモデル共に，本実験で得られた最適パラメータを用いている．

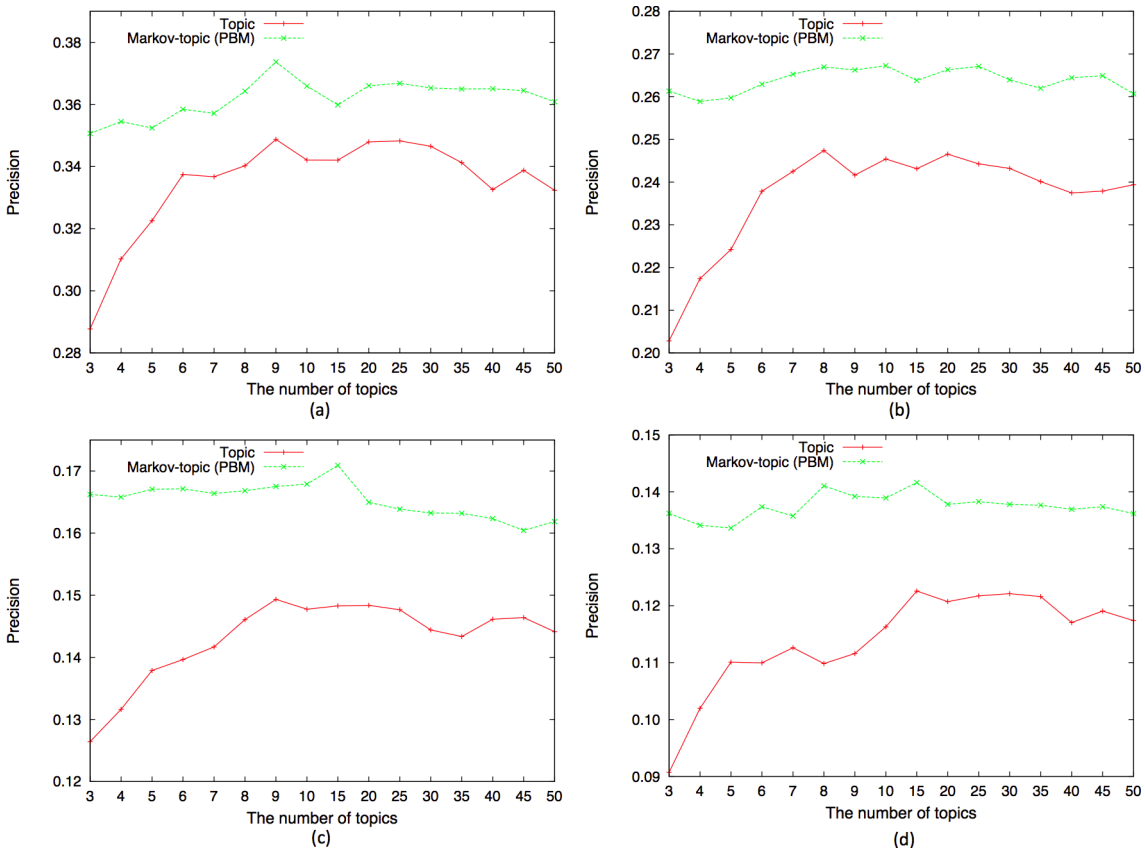


図 5.3: 潜在トピック数と適合率の関係

5.3.3　実験 2: 単一ランドマークの予測

ユーザが次に訪れる単一のランドマークを予測するタスク（One-step prediction）に関して，以下の確率モデルの精度比較を行う．

多項分布モデル ：ランドマークに関する多項分布 $P(l)$ による一般性に基づく推薦．領域内で最も訪れる人が多いランドマーク（現在地は除く）を推薦する．

マルコフモデル: 現在地を考慮した推薦．マルコフモデル $P(l_t|l_{t-1})$ に基づき，現在地と同じ場所にいる人々が，次に行く傾向が強いランドマークを推薦する．

トピックモデル: 興味を考慮した推薦．PLSA で求めた $P(l|h^u)$ に基づき，ユーザの興味に合致したランドマークを推薦する．

PBM（提案法）： 現在地と興味の両方を考慮した推薦．ユニグラムリスケーリングでマルコフモデルとトピックモデルを組合せたモデルである．

全ユーザに関して最後に訪れたランドマークを除いた移動履歴の集合から 1 つのモデルを学習する．そのモデルを用いて，各ユーザが最後に訪れたランドマークを予測するテストを行う．したがって，学習データの総数は，表 5.2 に示した移動履歴数（ユーザ数）に等しく，また，テスト回数も，表 5.2 に示した移動履歴数（ユーザ数）に等しい．評価指標は全ユーザ中で予測が正解した人の割合（適合率）である．その評価結果を図 5.4 に示す．X 軸はデータセット，Y 軸は適合率である．さらに，提案法とベースライン手法の適合率に統計的な優位差があることを符号検定によって確認した（両側検定: p 値 < 0.01）．この結果は，提案モデルがベースライン手法よりも適切に行動予測が可能であることを示すものである．なお，本実験で算出したのは，単一のランドマークをユーザに提示した場合の予測精度である．提案法の精度自体が高い値を示しているわけではないが，推薦システム上では複数の情報を推薦することを想定しており，より多くのユーザの情報要求に応えることができると考える．

5.3.4 実験 3: ランドマーク系列の予測

ユーザが空き時間内で辿るランドマークの系列を予測するタスクに関して精度比較を行う．実験 2 においては，各ユーザが最後に訪れた単一のランドマークを予測したが，本実験においては，各ユーザが最後の s 時間に辿ったランドマーク系列を予測する．学習データとして用いたのは，最後の s 時間に辿ったランドマーク系列（テストデータ）を除いた全ユーザの移動履歴の集合である．経過時間が s 時間より短い移動履歴は学習データ，及びテストデータから除いたため，学習データの総数，テスト回数は実験 2 とは異なる．本実験に用いた移動履歴と，ランドマーク数の詳細を 5.4 に示す．ただし，ランドマーク数が 5 未満の移動履歴，及び，訪れたユーザ数が 3 未満のランドマークは除いた．

提案手法によって生成される系列と，正解データとなる系列の “近さ” を測る評

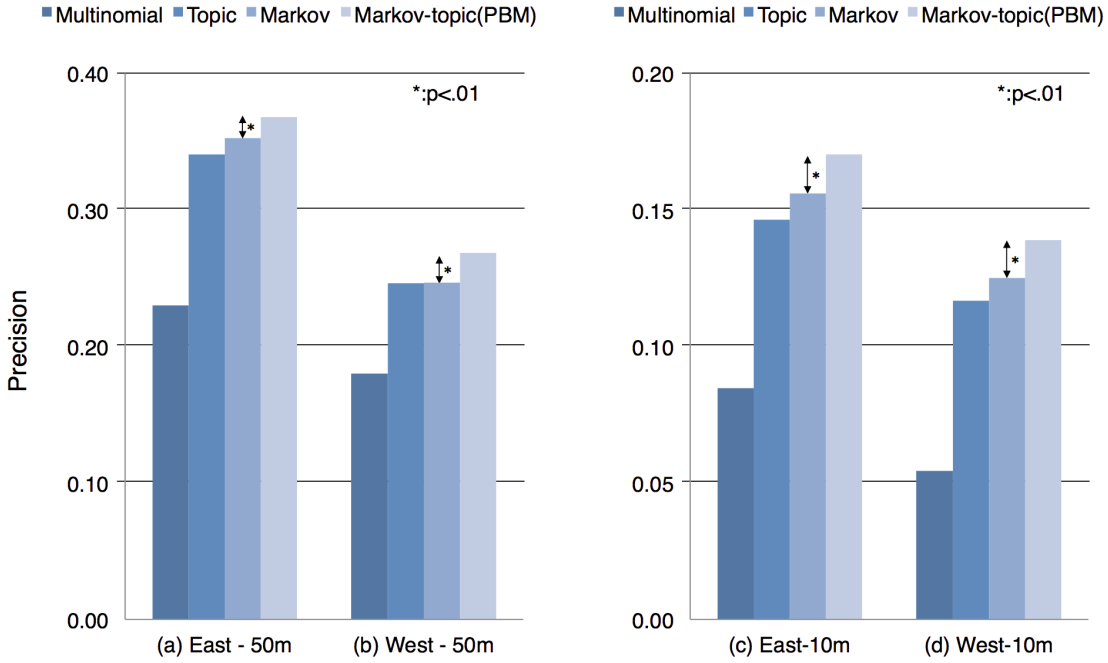


図 5.4: 単一ランドマークの予測タスクにおける適合率の比較

価指標として，編集距離（edit distance）を用いた．編集距離は，2 つの系列が与えられた場合に，一方の系列をもう一方の系列に変換する編集操作（追加，削除，置換）の最小ステップ数で定義され，複数の系列間の類似性を測るための指標である [67]．図 5.5 に，空き時間 d ごとの編集距離の全ユーザに関する平均値を示す．空き時間 d は 2 時間から 5 時間まで，1 時間刻みで設定した．実験 2 と同様，すべての空き時間，データセットにおいて，提案法の予測精度が最も高くなった（提案法の編集距離が最も低くなった）．また，提案法とベースライン手法の適合率に統計的な優位差があることを符号検定によって確認した（両側検定: p 値 < 0.01）．この結果は，提案法がユーザの現在地と興味の両者を考慮することで，ランドマークの系列（トラベルルート）を正確に予測できることを示すものである．

表 5.4: ランドマーク系列の予測実験に用いた移動履歴情報

Dataset	Region-scale	Time period (hours)	The number of sequences	The number of landmarks
(a)	East - 50m	2	9,142	414
		3	9,124	414
		4	9,115	414
		5	9,102	414
(b)	West - 50m	2	6,349	316
		3	6,346	316
		4	6,333	316
		5	6,323	316
(c)	East - 10m	2	11,139	1,419
		3	11,102	1,418
		4	11,074	1,418
		5	11,049	1,418
(d)	West - 10m	2	7,763	1,117
		3	7,741	1,117
		4	7,723	1,117
		5	7,704	1,117

5.3.5 実験 4: 移動手段に応じた移動時間の推定

提案法は，あるランドマークからほかのランドマークへ移動するための時間を，移動手段ごとに推定する．本実験では，既存の地域情報ナビゲーションサービスで得られる移動時間と比較することで，提案手法の特徴を述べる．

本実験に用いたのは，2 つのデータセット（アメリカ西海岸-平滑化パラメータ=50m，アメリカ東海岸-平滑化パラメータ=50m）である．各データセットにおいて，2 つのランドマークをつなぐトラベルルートを観測数の順にソートし，その上位 15 件を人気のトラベルルートとして抽出した．次に，各トラベルルートについて，K 個の代表的な移動時間を抽出した．なお，移動時間の間隔が 1 時間 30 分以上の観測値は，寄り道をした可能性が高いためノイズとして除去している．また，

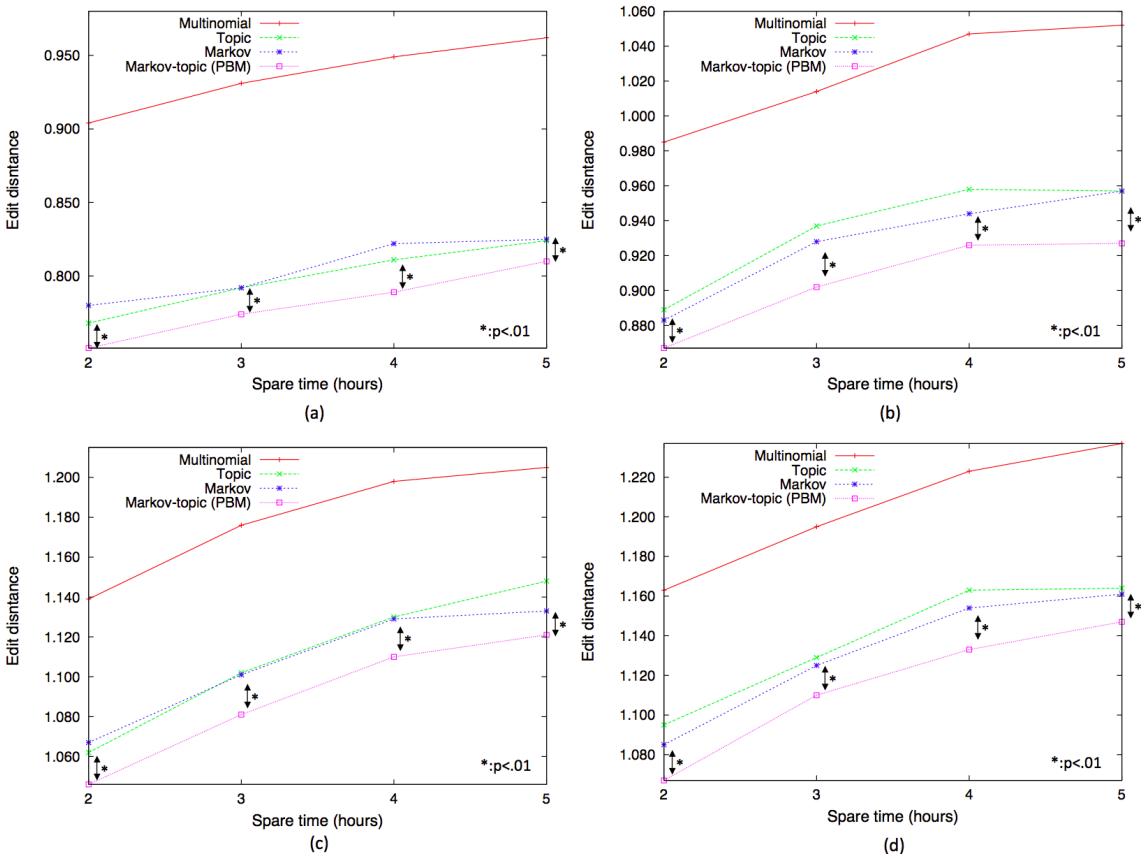


図 5.5: 空き時間と平均編集距離の関係

徒歩，公共交通機関，自動車を主要な移動手段と考え，K の値を 3 としている．図 5.6 に，観測数の上位 3 件のトラベルルートに関する移動時間の頻度分布を示すが，実際に，3 つのピークが存在することが確認できる．さらには，Google Maps のような既存の地域情報ナビゲーションサービスにおいても，これら 3 つの移動手段の結果を提示していることから，K の値を 3 とするのが妥当だと考えた．多くの場合，徒歩よりも公共交通機関，公共交通機関よりも自動車での移動のほうが移動時間が短くなる．したがって，本研究においては，推定値が小さい（移動時間が短い）ものから，徒歩の場合の移動時間，公共交通機関を利用した場合の移動時間，自動車を利用した場合の移動時間の推定値とみなした．比較対象としたのは Google Maps のルート・乗り換え検索機能である．

図 5.7 に実験結果を示す．Y 軸は，移動時間の推定値を 15 個のトラベルルート

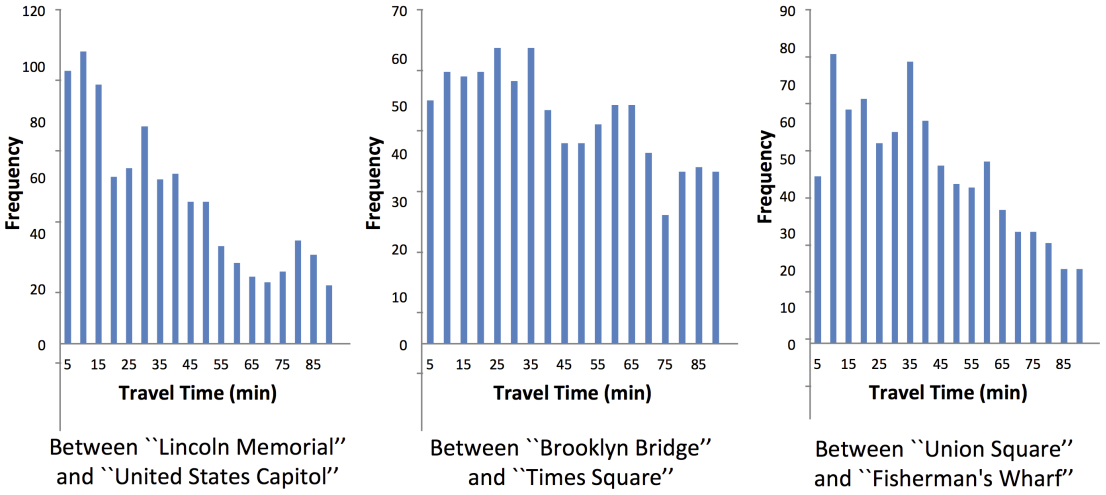


図 5.6: 移動時間の観測値の頻度分布

について平均した値である．(a)，(b) はデータセットである．各移動手段ごとに，Google Maps の結果と比較した．すべての移動手段において Google Maps で得られる値よりも，提案手法のほうが移動時間を長い値として推定していることが分かる．図 5.7 に各トラベルルートについて提案法と比較手法の推定移動時間の差分を計算し，全 15 個のトラベルルートについて平均したものを示すが，こちらのデータにおいても同様の結論が得られる．提案手法は，実際に都市を訪れた人々の移動履歴に基づいて移動時間を推定する．単純に移動するための時間だけでなく，各ランドマークへの滞在時間や，交通機関の混雑混雑などを総合的に加味した値であるため，現実的な旅行計画を立てる際に適切な情報を提示可能であるといえる．

5.3.6 トラベルルートの例

提案するトラベルルート生成手法を実装したシステムの実際の出力例を示し，ユーザからの入力（移動履歴と空き時間）と推薦されるトラベルルートとの関係性について述べる．図 5.1 に示した通り，提案システムは地図インタフェースをベースとしたシステムとして実装されている．地図インタフェース上で空間領域を指

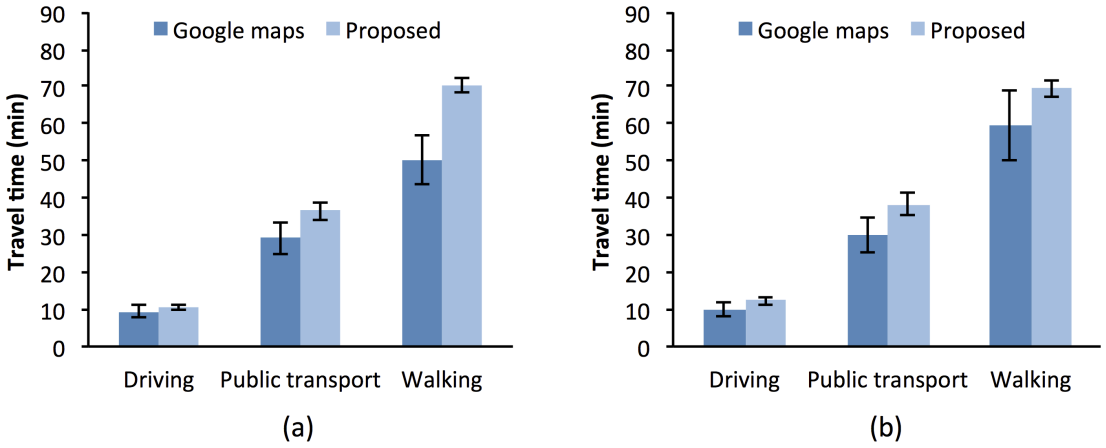


図 5.7: 全 15 個のトラベルルートにおける推定移動時間の平均値

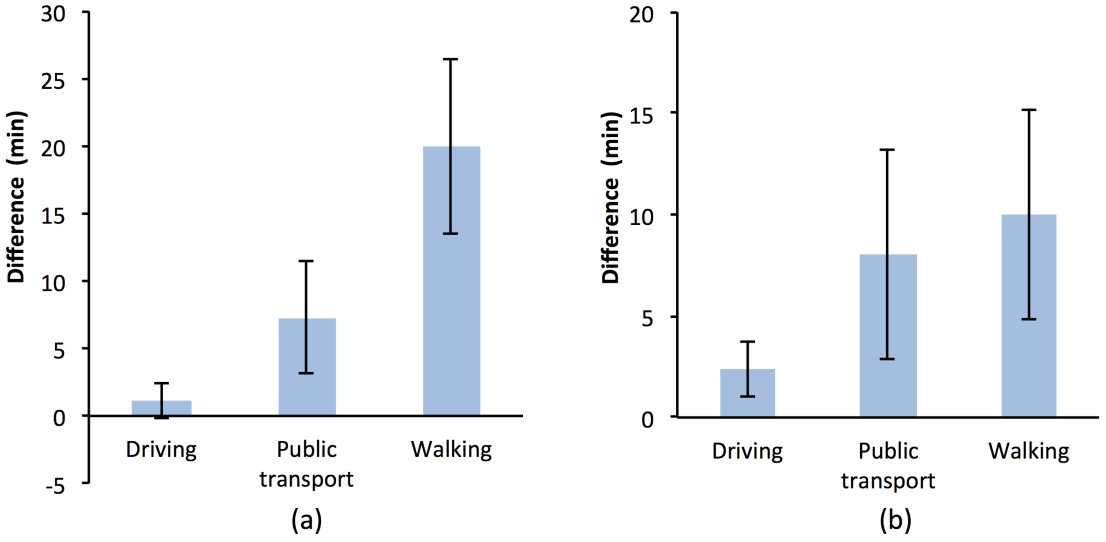


図 5.8: 提案手法と Google Maps の推定移動時間の差の平均値

定すると，システムは領域内で抽出されたランドマーク情報を緯度，経度情報に基づいて地図インタフェース上にプロットする．地図上の各アイコンはランドマークを示しており，それらを選択することで，過去の移動履歴や現在地情報をシステムに伝えることができる．ここで，“pier35” や “goldengatebridge” などの各ランドマークを表現する代表的なテキストタグも，写真に付与されたテキストタグ情報から自動抽出している．ランドマークを代表するテキストタグは，Crandall らが示した手法に基づいて抽出した [41]．ランドマーク l おける，タグ v のスコアは，

$$TagScore_l(v) = P(l|v) = \frac{N(v,l)}{N(v)} \tag{5.12}$$

であり，$n(v,l)$ は，ランドマーク l にクラスタリングされた写真群の中で，タグ v が付与された写真数，$n(v)$ は，全データセット中でタグ v が付与された写真数である．なお，$n(v,l)$ が，ランドマークにクラスタリングされた写真群の 5%に満たないタグについてはノイズとして除去した．

状況に応じた推薦

図 5.9 に，ワシントン D.C. のスミソニアン博物館にいるユーザへの推薦例を示す．空き時間に余裕がない 2 時間の場合は，リンカーン記念館など，スミソニアン博物館に隣接し，都市の中心部に位置するランドマークが推薦される．一方，空き時間に余裕がある 4 時間の場合はリンカーン記念館やアーリントン国立墓地などを周遊するトラベルルートが推薦される．これまで，旅行者はガイドブックや地域ポータルサイトなどを利用して地域情報を収集していた．これらのメディアに掲載された地域を代表するトラベルルートは，最低でも半日を要するものであった．本システムは，ユーザの状況（現在地と空き時間）に応じ，より柔軟にルートを推薦することができる．また，提案システムは徒歩，公共交通機関，自動車など，ユーザの移動手段に応じて推薦情報を変化させることもできる．図 5.10 に，サンフランシスコのピア 39 にいるユーザが，複数の移動手段（徒歩，公共交通機関，自動車）を指定した場合の推薦結果を示す．提案システムは，移動手段に応じて推薦されるトラベルルートに含まれるランドマーク数を適切に変化させる．

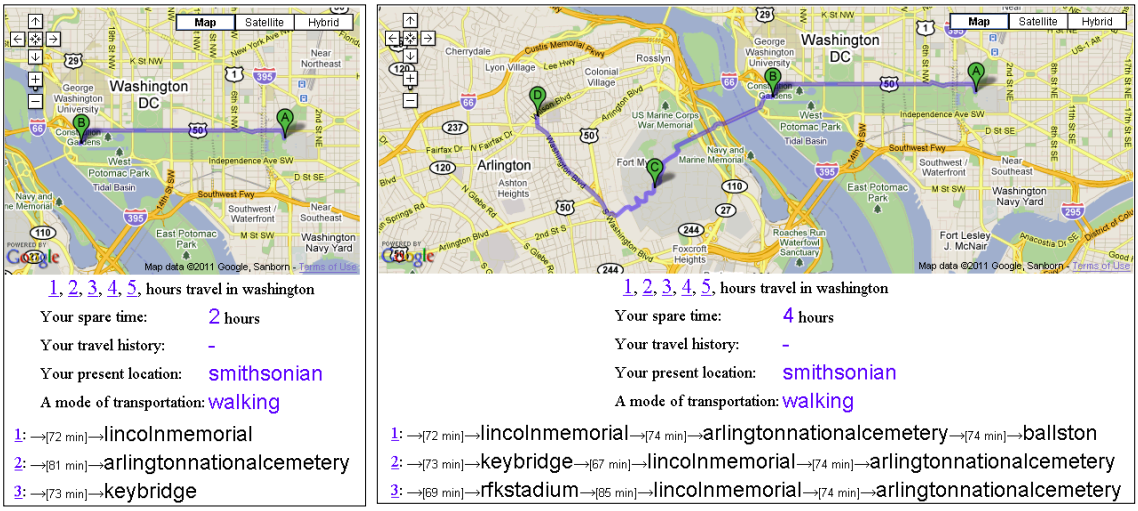
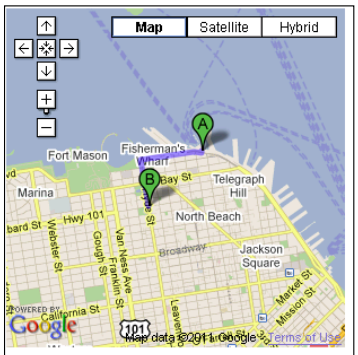


図 5.9: 空き時間に応じたトラベルルートの推薦例（現在地=“スミソニアン博物館”, 平滑化パラメータ=“10m”）

興味と状況に応じた推薦

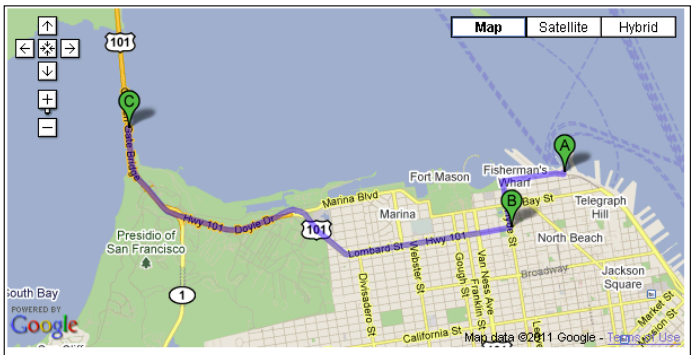
次に，アートに興味を持つユーザに対する推薦例を示す．この例においては，ニューヨークにあり，アートに由来のある 5 つのランドマーク（Chelsea, SoHo, DUMBO, Brooklyn Museum, the Lower East Side）をシステムユーザの移動履歴とした．ユーザの空き時間は 3 時間，現在地はタイムズスクエアとしている．タイムズスクエアは，収集したデータセットにおいて，ニューヨーク近郊で最も多くの人が撮影したランドマークである．図 5.11 に，マルコフモデルと提案法の推薦結果を示す．マルコフモデルが，グランドゼロやブルックリンブリッジなどの都市において有名なランドマークを含むトラベルルートを推薦しているのに対し，提案法は，アメリカ自然史博物館やブロードウェイといったアートに関連するランドマークを推薦している．また，提案法はトピックモデルに加えてマルコフモデル（アクセスのしやすさ）をも考慮したモデルである．そのため，アメリカ自然史博物館やブロードウェイなどは，タイムズスクエアからのアクセスしやすさ，という観点でも上位に提示されていると考えられる．



Your spare time: 2 hours
Your travel history: -
Your present location: pier39
A mode of transportation: walking

1: →[68 min]→lombardstreet
2: →[65 min]→fishermanswharf
3: →[76 min]→alcatraz
4: →[74 min]→goldengatebridge
5: →[63 min]→exploratorium

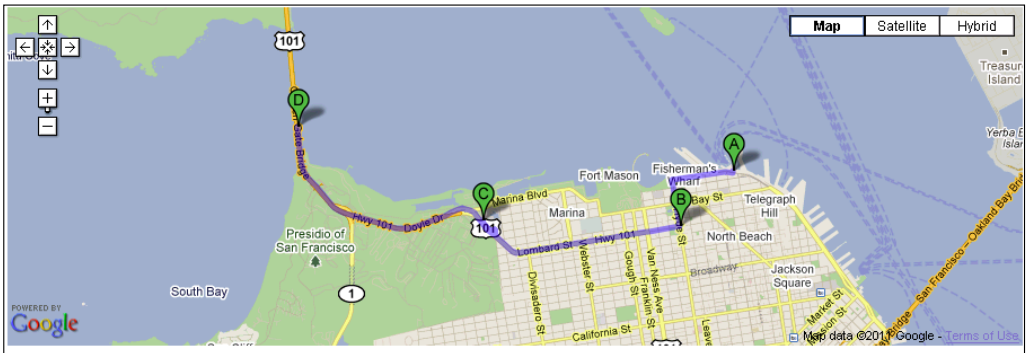
Walking



Your spare time: 2 hours
Your travel history: -
Your present location: pier39
A mode of transportation: public transportation

1: →[37 min]→lombardstreet→[44 min]→goldengatebridge A
2: →[32 min]→fishermanswharf→[35 min]→lombardstreet
3: →[45 min]→goldengatebridge A→[31 min]→goldengatebridge B
4: →[42 min]→alcatraz→[55 min]→lombardstreet
5: →[33 min]→exploratorium→[29 min]→crissyfield

Public transportation

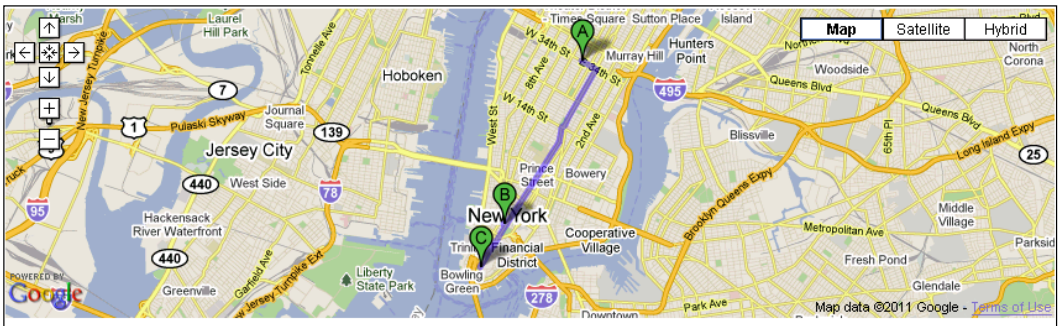


Your spare time: 2 hours
Your travel history: -
Your present location: pier39
A mode of transportation: driving

1: →[20 min]→lombardstreet→[21 min]→exploratorium→[20 min]→goldengatebridge A
2: →[32 min]→crissyfield→[16 min]→goldengatebridge A→[16 min]→goldengatebridge B
3: →[20 min]→alcatraz→[21 min]→lombardstreet→[24 min]→goldengatebridge A
4: →[16 min]→exploratorium→[18 min]→crissyfield→[16 min]→goldengatebridge A→[16 min]→goldengatebridge B
5: →[23 min]→fortmason→[19 min]→exploratorium→[18 min]→crissyfield

Driving

図 5.10: 移動手段に応じたトラベルルートの推薦例（現在地=“ピア 39”，空き時間=“2 時間”，平滑化パラメータ=“10m”）



Your spare time: 3 hours
Your travel history: chelsea,soho,dumbo,brooklynmuseum,lowereastside
Your present location: timessquare
A mode of transportation: -

1: →[79 min]→groundzero→[50 min]→batterypark
2: →[74 min]→westvillage→[56 min]→groundzero
3: →[82 min]→brooklynbridge→[50 min]→groundzero
4: →[62 min]→eastvillage→[67 min]→groundzero
5: →[70 min]→metropolitanmuseumofart→[55 min]→americanmuseumofnaturalhistory

Markov model

Your spare time: 3 hours
Your travel history: chelsea,soho,dumbo,brooklynmuseum,lowereastside
Your present location: timessquare
A mode of transportation: -

1: →[78 min]→americanmuseumofnaturalhistory→[57 min]→metropolitanmuseumofart
2: →[51 min]→broadway→[74 min]→americanmuseumofnaturalhistory
3: →[79 min]→groundzero→[67 min]→americanmuseumofnaturalhistory
4: →[62 min]→eastvillage→[66 min]→metropolitanmuseumofart
5: →[118 min]→unionhall→[33 min]→gowanus

Markov-topic model (proposed model)

図 5.11: マルコフモデルと提案法の推薦結果の比較（現在地=“タイムズスクエア”, 空き時間=“3 時間”, 平滑化パラメータ=“10m”）

5.4 ユーザ興味推定のためのトピックモデル

本節では，トピックモデルによるユーザ興味のモデリングを高度化する手法を提案する．5.2.3 項で述べた PBM では，トピックモデルを用いてユーザの興味をモデル化した．しかし，ある場所を訪れるか決定するうえで，移動コストは考慮すべき重要な要素の 1 つであり，ユーザは物理的に近い場所ほど訪れやすい傾向がある．近い場所を選択する傾向が色濃く反映された移動履歴集合に対してトピックモデルを適用すると，位置が近い場所どうしをまとめるものとして潜在トピックが抽出され，同時に，各ユーザの行動範囲を表すものとして，ユーザ固有のトピック比率が学習されてしまう問題があった．

本節では，その問題に対処し，ユーザの純粋な興味を説明する潜在トピックを推定するためのジオトピックモデルを述べる．提案するジオトピックモデルは，1) 普段の行動範囲からの近さと，2) 個人的な興味とに基づいてユーザは行き先を決定するという人間行動に関する仮定に基づく確率的行動モデルである．移動履歴からユーザの興味情報を推定するためのキーとなるアイディアは，ユーザの行動範囲が場所の選択に与える影響を明示的にモデリングすることにある．これにより，ユーザの興味対象を表現するものとして潜在トピックを学習することが可能となる．より直感的には，ユーザ自身の行動範囲から近い場所の選択は “アクセスのしやすさ” が理由で引き起こされたものとして説明し，逆に，ユーザ自身の行動範囲から離れた場所の選択は “ユーザ固有の興味” によって引き起こされたものとして潜在トピックを用いて説明するモデルである．さらに，各場所のメタデータ（テキストタグ）を用いて，潜在トピックに意味付け（タギング）を行うことで，“言語的解釈を与えた潜在トピックに対する確率的重み” としてユーザの興味を表現できる．

5.4.1 ジオトピックモデル

N 人のユーザ集合を $\boldsymbol{U} = \{u\}_{u=1}^{N}$，$I$ 個の訪問対象となる場所を $\boldsymbol{I} = \{i\}_{i=1}^{I}$ と表現する．また，各場所は位置座標 r_i(緯度，経度情報) を持ち，各ユーザは移動履歴

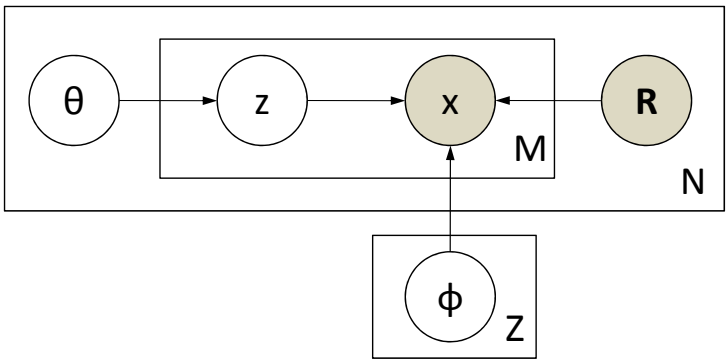


図 5.12: ジオトピックモデルのグラフィカルモデル

$\boldsymbol{x}_u = \{x_{u1}, ..., x_{uM_u}\}$ を持つ．ここで，x_{um} はユーザ u が m 番目に訪れた場所であり，$x_{um} \in \boldsymbol{I}$ である．ジオトピックモデルの説明に用いる記号を表 5.5 にまとめた．

ユーザの場所選択のモデル化において，我々は以下の仮定を設けた．

[1] ユーザが訪れる場所はユーザ自身の行動範囲によって決まる．たとえば，勤務先が東京駅にある人は東京駅近辺の場所を訪れやすい．

[2] ユーザが訪れる場所はユーザ個人の興味によって決まる．たとえば，アートに興味がある人は東京ドームよりも東京国立近代美術館を訪れやすい．

これらの仮定を場所生成プロセスに反映させる．各ユーザは，ユーザ自身の興味を表すユーザ依存の潜在トピック分布 $\boldsymbol{\theta}_u = \{\theta_{uz}\}_{z=1}^{Z}$ を持つ．$\sum_z \theta_{uz} = 1$，$\theta_{uz} \geq 0$ であり，Z は潜在トピック数である．ユーザは $\boldsymbol{\theta}_u$ によってある潜在トピックを選択した後，潜在トピック依存のパラメータ $\boldsymbol{\phi}_z = \{\phi_{zi}\}_{i=1}^{I}$ と，過去に訪れた場所の位置座標集合 $\boldsymbol{R}_u = \{r_{x_{um}}\}_{m=1}^{M_u}$ とに基づいてある場所を選択する．ここで，ϕ_{zi} は場所 i の潜在トピック z からの選ばれやすさを表すパラメータである．提案モデルのグラフィカルモデルを図 5.12 に示す．塗潰し円が観測変数を，中抜き円が潜在変数を表している．

位置座標集合 $\boldsymbol{R}_u$ を持つユーザ u が場所 i を選択する確率は，

$$P(i|u, \boldsymbol{R}_u, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \sum_{z=1}^{Z} P(z|u, \boldsymbol{\Theta}) P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi}) \tag{5.13}$$

表 5.5: ジオトピックモデルの説明に用いる記号

記号	説明
$\boldsymbol{U}$	ユーザ集合
u	ユーザ，$u \in \boldsymbol{U}$
$\boldsymbol{I}$	場所集合
i	場所，$i \in \boldsymbol{I}$
z	潜在トピック
M_u	ユーザ u の移動履歴に含まれる場所数
x_{um}	ユーザ u の m 番目の場所，$x_{um} \in \boldsymbol{I}$
$\boldsymbol{x}_u$	ユーザ u の移動履歴，$\boldsymbol{x}_u = \{x_{u1}, ..., x_{uM_u}\}$
r_i	場所 i の位置座標 (緯度，経度)
$\boldsymbol{R}_u$	ユーザ u の場所集合の位置座標集合，$\boldsymbol{R}_u = \{r_{x_{um}}\}_{m=1}^{M_u}$
θ_{uz}	ユーザ u の潜在トピック z の選択確率
$\boldsymbol{\theta}_u$	ユーザ u の潜在トピック選択確率集合，$\boldsymbol{\theta}_u = \{\theta_{uz}\}_{z=1}^{Z}$
ϕ_{zi}	ランドマーク i が潜在トピック z から選択される確率
$\boldsymbol{\phi}_z$	潜在トピック z の場所選択確率集合，$\boldsymbol{\phi}_z = \{\phi_{zi}\}_{i=1}^{I}$
N	ユーザ数
I	場所数
Z	潜在トピック数
β	行動範囲を調整する bandwidth パラメータ

となる．なお，$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_u\}_{u=1}^{N}, \boldsymbol{\Phi} = \{\boldsymbol{\phi}_z\}_{z=1}^{Z}, P(z|u, \boldsymbol{\Theta}) = \theta_{uz}$ である．$P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi})$ は，ユーザの位置座標集合 $\boldsymbol{R}_u$ を加味したうえで，場所 i が潜在トピック z から選択される確率である．我々は，過去に訪れた場所に地理的に近いほかの場所もまた選ばれやすいと仮定している．$P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi})$ は，

$$P(i|z, \boldsymbol{R}_u, \boldsymbol{\Phi}) = \frac{1}{C} \exp(\phi_{zi}) \sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2} \|r_i - r\|^2) \tag{5.14}$$

となる．なお，$C = \sum_{i' \in \boldsymbol{I}} \exp(\phi_{zi'}) \sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2} \|r_{i'} - r\|^2)$ は正規化項，$\|\cdot\|$ は位置座標空間におけるユークリッド距離である．β は行動範囲の広さを決定する bandwidth パラメータであり，β の値を小さく設定するほど，行動範囲をより広くとらえることができる．$\exp(\phi_{zi})$ がユーザの興味が場所の選択に及ぼす影響を決

定する項である．第 1 項 ϕ_{zi} は，場所 i が潜在トピック z から選択される確率であり，PLSA などの既存のトピックモデルも同様のパラメータを含んでいる．もし，潜在トピック z がアートを表現するものとして推定されれば，場所 i として，ミュージアムや美術館などのアートに関する場所が選択されやすくなる．第 2 項 $\sum_{r\in\boldsymbol{R}_u}\exp(-\frac{\beta}{2}\|r_i-r\|^2)$ はユーザの行動範囲が場所の選択に及ぼす影響を表現しており，ユーザの移動履歴 $\boldsymbol{x}_u$ に含まれる場所の近くに存在するほかの場所は選択されやすくなる．たとえば，ユーザがある駅の周辺に存在するレストランを良く利用する場合，その駅の周辺に存在するほかのレストランも選択されやすくなる．場所（とユーザ）に関する地理的な特徴を明示的にモデリングすることにより，地理的な特徴以外の潜在的特徴（とそれに対するユーザの興味）を扱うものとして潜在トピックを推定する．より直感的に提案モデルを述べると，ユーザ自身の行動範囲から近い場所の選択はアクセスのしやすさが理由で引き起こされたものとして説明し，逆に，ユーザ自身の行動範囲から離れた場所の選択は，ユーザ固有の興味によって引き起こされたものとして潜在トピックを用いて説明する．

5.4.2　パラメータ推定

最尤法を用いて提案モデルのパラメータを推定する．未知パラメータは潜在トピック選択確率集合 $\boldsymbol{\Theta}$ と，場所選択確率集合 $\boldsymbol{\Phi}$ であり，全未知パラメータ集合を $\boldsymbol{\Psi}=\{\boldsymbol{\Theta},\boldsymbol{\Phi}\}$ で表現する．ユーザの移動履歴集合 $X=\{\boldsymbol{x}_u\}_{u=1}^N$ が与えられたときのパラメータ集合 $\boldsymbol{\Psi}$ の対数尤度は，

$$L(\boldsymbol{\Psi}|\boldsymbol{X})=\sum_{u=1}^{N}\sum_{m=1}^{M_u}\log\sum_{z=1}^{Z}\theta_{uz}P(x_{um}|z,\boldsymbol{R}_u,\boldsymbol{\Phi}) \tag{5.15}$$

となる．事後確率を EM アルゴリズム [68] を用いて最大化することにより，未知パラメータ $\boldsymbol{\Psi}$ を推定する．事前確率を含む完全対数尤度は，

$$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})=\sum_{u=1}^{N}\sum_{m=1}^{M_u}\sum_{z=1}^{Z}P(z|u,m;\hat{\boldsymbol{\Psi}})\log\theta_{uz}P(x_{um}|z,\boldsymbol{R}_u,\boldsymbol{\Phi}) \tag{5.16}$$

となる．ここで，$\hat{\boldsymbol{\Psi}}$ は現在の推定値，$P(z|u,m;\hat{\boldsymbol{\Psi}})$ は現在の推定値が与えられたときの，ユーザ u の m 番目の場所のトピック事後確率を表す. E ステップでは，ベ

イズ則に従い，トピック事後確率を，

$$P(z|u,m;\hat{\boldsymbol{\Psi}}) = \frac{P(z|u,\boldsymbol{\Theta})P(x_{um}|z,\boldsymbol{R}_u,\boldsymbol{\Phi})}{\sum_{z'=1}^{Z} P(z'|u,\boldsymbol{\Theta})P(x_{um}|z',\boldsymbol{R}_u,\boldsymbol{\Phi})} \tag{5.17}$$

で計算する．ここで，$P(x_{nm}|z,\boldsymbol{R}_u,\boldsymbol{\Phi})$ は式 (2) により計算する．M ステップでは，$\sum_{z=1}^{Z}\theta_{uz}=1$ という制約のもと，$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ を θ_{uz} に関して最大化することにより，トピック選択確率の推定値 $\hat{\theta}_{uz}$ を，

$$\hat{\theta}_{uz} = \frac{\sum_{m=1}^{M_u} P(z|u,m;\hat{\boldsymbol{\Psi}})}{\sum_{z'=1}^{Z}\sum_{m=1}^{M_u} P(z'|u,m;\hat{\boldsymbol{\Psi}})} \tag{5.18}$$

で求める．$\boldsymbol{\phi}_z$ の推定値は閉形式で書くことができない．そのため，準ニュートン法 [69] などの最適化法を用いて $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ を最大化することにより推定する．準ニュートン法で必要となる $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ の $\boldsymbol{\phi}_z$ に関する勾配ベクトルは，

$$\frac{\partial Q}{\partial \boldsymbol{\phi_z}} = \sum_{u=1}^{N}\sum_{m=1}^{M_u} P(z|u,m;\hat{\boldsymbol{\Psi}}) - \sum_{u=1}^{N}\sum_{m=1}^{M_u} P(z|u,m;\hat{\boldsymbol{\Psi}})P(x_{um}|z,\boldsymbol{R}_u,\boldsymbol{\Phi}) \tag{5.19}$$

となる．収束するまで E ステップと M ステップを交互に繰り返すことにより，$\boldsymbol{\Psi}$ の最適解を推定することができる．なお，潜在トピック数と行動範囲の bandwidth パラメータ β は交差検定を用いて推定する．

5.5 ジオトピックモデルの評価実験

ランドマークとレストランの訪問履歴を用いて提案法の有効性を示す．まず，ユーザが次に訪れる場所を予測するタスクにおける予測精度を評価することでモデルの妥当性を検証する．その後，場所に付与されたメタデータを用いて提案法が推定する潜在トピックに対して意味付けを行い，既存のトピックモデルと異なる観点で場所の特徴 (とそれに対するユーザの興味情報) を推定可能なことを示す

表 5.6: 潜在トピック抽出実験に用いた移動履歴

地域	ユーザ数	場所数	移動履歴の平均長
東京	6,257	910	14.872
大阪	2,304	763	15.995
東日本	7,326	1,067	14.871
西日本	3,531	956	15.245
ニューヨーク (NY)	3,363	967	9.778
サンフランシスコ (SF)	4,433	1,030	12.231
ロサンジェルス (LA)	2,848	943	11.686

5.5.1 データセット

食べログサービス[3] に投稿されたレビュー情報をもとに 4 種類の移動履歴データを作成した．東京都，大阪府，東日本（東京都，神奈川県，埼玉県），西日本（大阪府，京都府，兵庫県）に存在する店舗の訪問履歴（レビュー履歴）である．同時に，各地域に存在する店舗の位置座標も収集した．また，写真共有サイト Flickr に投稿された写真に付与されたジオタグ（緯度，経度）情報を Flickr API を利用して収集した．ユーザ単位にジオタグ情報を集約することで，個人の移動履歴を再現することができる．収集した Flickr データから，ニューヨーク，サンフランシスコ，そしてロサンジェルスに関する 3 種類の移動履歴データを作成した．最初に，Flickr API を利用して各都市の中心部から 50km 以内で撮影されたジオタグ情報を取得した．次に，前節の実験と同様の方法でジオタグ集合から多くの人々が撮影したランドマークを抽出し，ランドマーク訪問履歴を生成した．実験に用いる計 7 種類の移動履歴について表 5.6 にまとめた．なお，観測データの少ないユーザ（訪問場所数が 5 未満のユーザ）は削除している．また，推薦対象となる場所数を各データセットで均一となるように，訪問ユーザ数が 100，40，100，50，10 未満の場所を東京，大阪，東日本，西日本，Flickr データセット（ニューヨーク，サンフランシスコ，ロサンジェルス）からそれぞれ削除した．

[3] http://tabelog.com/

5.5.2 モデル比較

予測精度で提案モデルの妥当性を示す．我々は提案モデルを次の 4 つの確率モデルと比較した．

多項分布モデル: 場所に関する多項分布 $P(i)$ による一般性に基づく予測．

カーネルモデル: ユーザの行動範囲に基づく予測．過去に訪れた場所の近辺に存在するほかの場所を出力する．

$$P_K(i|u) = \frac{\sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2}\|r_i - r\|^2)}{\sum_{i' \in \boldsymbol{I}} \sum_{r \in \boldsymbol{R}_u} \exp(-\frac{\beta}{2}\|r_{i'} - r\|^2)} \tag{5.20}$$

トピックモデル: PLSA が推定したユーザの興味情報に基づく予測．

$$P_T(i|u) = \sum_{z \in \boldsymbol{Z}} P(z|u)P(i|z) \tag{5.21}$$

カーネル - トピックモデル: ユーザの行動範囲と（PLSA が推定した）興味情報とに基づく予測．カーネルモデルとトピックモデルの線形結合，

$$P_{KT}(i|u) = \alpha P_K(i|u) + (1-\alpha)P_T(i|u) \tag{5.22}$$

で計算する．α は 2 つのモデルの影響をコントロールするパラメータ．

各ユーザが最後に訪れた場所を予測する．最初に，全ユーザの移動履歴から最後に訪れた場所を除いたデータを学習データとしてモデルを学習する．その後，各ユーザが最後に訪れた場所の予測テストを，全ユーザに対して繰り返し行う．したがって，テスト回数は表 5.6 に示したユーザ数と等しくなる．評価指標は 5-best prediction accuracy を用いた．各ユーザに対して 5 つの場所を出力し，正解データが含まれた場合に 1，含まれない場合に 0 とする指標である．最終的に，1 を出力したユーザの全テストユーザに対する割合で各モデルを比較する．その比較結果を図 5.13 に示す．X 軸はデータセット，Y 軸は 5-best prediction accuracy をプロットしている．すべてのデータセットにおいて提案法が最も高精度であり，符号検

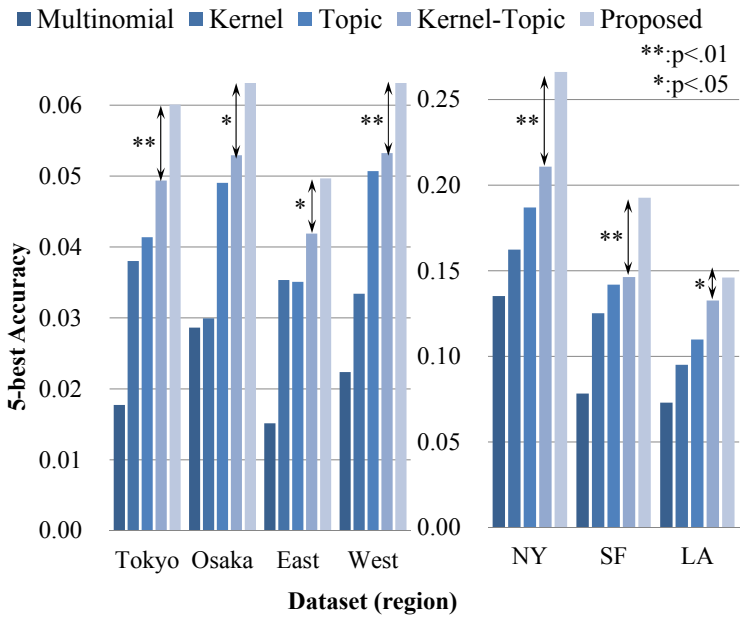


図 5.13: ユーザの行動予測精度（5-best prediction accuracy）の比較結果

定の結果，その差が有意であることが認められた（両側検定: p 値 < 0.05）．提案法とカーネル - トピックモデルは，ユーザの行動範囲と興味情報とに基づく予測を行うモデルである．それぞれ，潜在トピックとして扱う情報が異なることが，予測精度における優位性につながったといえる．これにより，提案モデルの妥当性を示せた．次節では，推定された潜在トピックについての定性的な考察を行う．

5.5.3　潜在トピック表現

既存のトピックモデルである PLSA との比較により，提案モデルで推定可能な潜在トピックの特徴を定性的に示す．まず，提案法，PLSA がそれぞれ推定した潜在トピックにおいて選択確率が高かった場所を地図上にマッピングした結果を図 5.14 から図 5.17 に示す．データセットは Flickr のジオタグ情報に基づくニューヨークとサンフランシスコのランドマーク訪問履歴である．また，アイコンの色が潜在トピックの種類を表し，各潜在トピックから選択確率が高かった上位 30 件のランドマークをマッピングしている．提案法と比較し，同色のアイコンが近い位置

に配置されていることから，既存のトピックモデル（PLSA）は位置が近い場所をまとめるものとして潜在トピックが推定されやすいことが分かる．一方で，提案法は，同色のアイコンが比較的離れた位置に配置されている．位置とは異なる何らかの場所特徴を扱うものとして潜在トピックが推定できていることが分かる．

それでは，提案法の潜在トピックはどのような場所特徴を扱っているのだろうか．この疑問に答えるために，Flickr にアップロードされた写真に付与されたテキストタグ情報を用いて潜在トピックに意味付けを行った．テキストタグは，地名，被写体，行動内容など，様々な観点でユーザによって付与されたものであるが，この中から各潜在トピックを代表するタグ集合を自動で取得する．

代表タグ抽出方法は以下の通りである．タグを g，タグ集合を G とする．各潜在トピックに特徴的に出現するタグを取得するために，潜在トピック z が与えられたときのタグ g のリフト値 (lift value) を，

$$Lift_z(g) = \frac{P(g|z)}{\sum_{z' \in \boldsymbol{Z}} P(g|z')P(z')} \tag{5.23}$$

で計算する．リフトはアソシエーションルール抽出において用いられる指標であり，$P(g|z)$ と，全トピックにおけるタグ g の出現確率との比で表される．全トピックに共通して出現する “nyc” や “sf” などのタグに対して低い値を与え，潜在トピック z に相対的に多く出現するタグ g に対して高い値を与える．潜在トピック z が与えられた場合のタグ g の出現確率 $P(g|z)$ は，

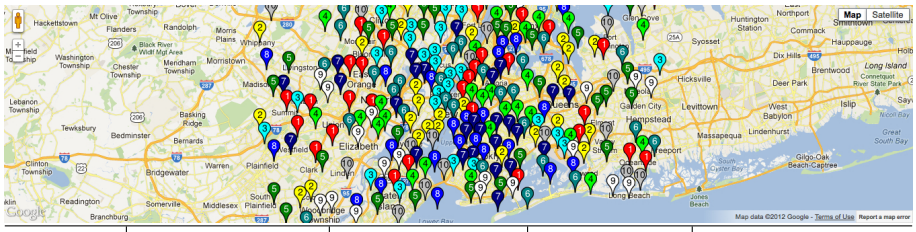
$$P(g|z) = \frac{1}{C} \sum_{i \in \boldsymbol{I}} \exp(\phi_{zi}) P(g|i) \tag{5.24}$$

で求めることができる．$C = \sum_{i' \in \boldsymbol{I}} \exp(\phi_{zi'})$ は正規化項である．$P(g|i)$ はランドマーク i が与えられた場合のタグ g の出現確率であり，

$$P(g|i) = \frac{H(g,i)}{H(i)} \tag{5.25}$$

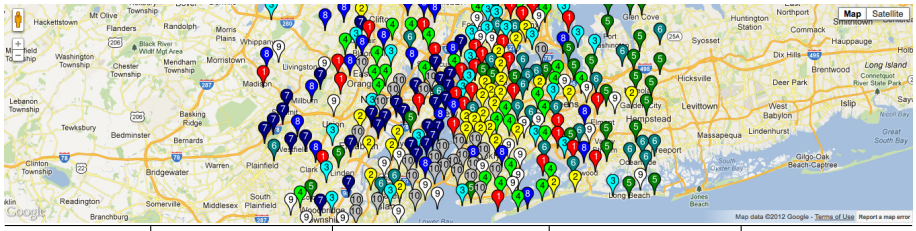
で求める．$H(i)$ はランドマーク i を訪問したユーザ数であり，$H(g,i)$ はランドマーク i にタグ g を付与したユーザ数である．

提案法，PLSA の各潜在トピックをテキストでラベリングした結果を，図 5.14 から図 5.17 中の各地図の下に示す．各モデルの $P(z) = \sum_{u \in \boldsymbol{U}} P(z|u)P(u)$ の降順で



topic 1	topic 2	topic 3	topic 4	topic 5
centralpark	ferry	construction	art	yankeestadium
guggenheim	bird	architecture	bar	coneyisland
met	sunset	chelsea	streetart	baseball
fifthavenue	lake	show	soho	summer
rockefeller	boat	urban	dumbo	mets
topic 6	topic 7	topic 8	topic 9	topic 10
cityscape	garden	livemusic	beach	spring
liberty	flowers	graffiti	zoo	bike
batterypark	prospectpark	music	vacation	bicycle
panorama	green	party	animals	color
skyline	nature	waterfront	garden	library

図 5.14: 提案モデルが推定した潜在トピックの代表タグ抽出結果（NY）



topic 1	topic 2	topic 3	topic 4	topic 5
met	parkslope	greenwichvillage	newark	coneyisland
guggenheim	prospectpark	soho	ice	neon
columbuscircle	astoria	eastvillage	jersey	beach
centralpark	williamsburg	jfk	nature	island
fifthavenue	eastvillage	lowereastside	tree	parade
topic 6	topic 7	topic 8	topic 9	topic 10
yankees	jerseycity	libertyisland	harlem	jerseycity
mets	airport	ellisisland	upperwestside	bridge
bronx	statenisland	statueofliberty	library	statenisland
queens	ferry	statenisland	broadway	astoria
eastriver	newark	batterypark	centralpark	trees

図 5.15: 従来モデルが推定した潜在トピックの代表タグ抽出結果（NY）

topic 1	topic 2	topic 3	topic 4	topic 5
goldengate	woods	ship	store	landscape
palaceoffinearts	goldengatepark	alcatraz	hotel	marin
downtown	trees	sunrise	house	sunrise
presidio	muirwoods	sea	dinner	sea
chinatown	park	bridge	marketstreet	coast
topic 6	topic 7	topic 8	topic 9	topic 10
graffiti	concert	sausalito	bike	coast
art	live	urban	bicycle	sailing
color	fireworks	cafe	bus	summer
mural	show	bar	ship	bakerbeach
haightashbury	livemusic	restaurant	birds	bike

図 5.16: 提案モデルが推定した潜在トピックの代表タグ抽出結果（SF）

topic 1	topic 2	topic 3	topic 4	topic 5
prison	palaceoffinearts	alameda	giants	goldengatepark
alcatraz	presidio	oakland	embarcadero	castro
island	marina	eastbay	soma	museum
pier39	fishermanswharf	berkeley	baseball	oceanbeach
presidio	pier39	train	baybridge	garden
topic 6	topic 7	topic 8	topic 9	topic 10
airport	muirwoods	oceanbeach	marinheadlands	mission
sfo	sausalito	pacificocean	coast	mural
berkley	marincounty	museum	marin	hill
baybridge	marin	soma	marincounty	house
baseball	marinheadlands	alameda	pacificocean	bart

図 5.17: 従来モデルが推定した潜在トピックの代表タグ抽出結果（SF）

潜在トピック 1 から 10 とした．2 つの都市において，提案法が抽出した最も影響力の大きい潜在トピックは有名な観光地をまとめるものであった（トピック 1）．このトピックに含まれる場所集合は，観光客にとって興味深いものである．トピック 2 から 10 は，自然，アート，エンターテイメント，スポーツ，眺めの良さなどの特徴を扱い，これらの特徴を持つ場所をまとめる役割を果たしていることが分かる．一方で，PLSA は地名を多く含んでおり，場所の位置的特徴を主に扱っていることが分かる．このように，提案手法は場所の潜在的な特徴を潜在トピックとして抽出し，潜在トピックに対して言語的な意味解釈を与えることができる．さらに，各潜在トピックは，自動で推定した $\boldsymbol{\theta}_u$ に基づき個人ごとに並び替え，ユーザプロファイルとして提示することもできる．トラベルルート推薦システムの中で，推薦の文脈でジオトピックモデルを用いた場合，各潜在トピックは，自動で推定した $\boldsymbol{\theta}_u$ に基づき個人ごとに並び替え，推薦する場所情報とともに，ユーザに提示することもできる．提示情報はユーザインタラクションを発生させ，推薦の満足度を全体として向上させることが期待できる．たとえば，最初に推薦された場所情報にユーザが満足しなかった場合，ユーザはシステムが自動で推定した自身のユーザプロファイルを確認し，明示的に自らの興味に合致した特定の潜在トピックを選択する．システムは，選択された潜在トピック z 依存の場所選択確率 ϕ_z に基づき，推薦リストを再提示する．このように，ユーザは最終的に満足のいく推薦結果をインタラクティブに引き出すことができる．

5.6　結言

本章では，人々の過去体験情報を利用して，自動でユーザの行動拡張をする仕組みの実現を目指し，ユーザの現在地と興味を考慮して，ユーザが将来的に訪れる確率の高い場所を予測し，推薦する手法を提案した．提案モデルは，個々のユーザの過去の移動履歴に反映された興味情報を扱うトピックモデルと，現在地からのアクセシビリティを扱うマルコフモデルを，確率的な枠組みに基づいて統合する．また，ユーザ自身の空き時間を入力として受け付けることで，単一の場所と

してではなく，より具体的な旅行計画としてユーザに情報提示を行うことができる．さらに，ユーザの過去の移動履歴からのユーザ興味推定を高度化するためのジオトピックモデルも提案した．評価実験においては，各ユーザの場所選択をどの程度予測できるか，を評価指標として，提案モデルの妥当性を確認した．

今後の課題は，前章までに述べた体験情報抽出技術との融合である．提案手法は，ユーザの行き先を予測し推薦するが，その推薦理由までは提示しない．予測が外れた場合や，普段とは異なる選択をしたい場合など，多様な背景を持つユーザに提案システムを使ってもらうためには，ユーザがインタラクティブに提示情報を修正する仕組みへと発展させる必要がある．“ホタルを見る”や“紅葉を見る”などの行動情報はある場所を訪れる理由や目的の説明となる場合も多いため，推薦情報の一部として提示するようにしたい．また，ランドマークの開店時間，利用料金，外観など，緯度/経度以外の場所特徴を扱うことも視野に入れている．

予測精度向上の観点からすると，写真共有サイト上のソーシャルネットワークや，各ユーザのプロファイル情報（性別，居住地など）を考慮することが必要と考える．同様の興味を持つユーザどうしは友人関係になりやすいと考えられるし，お互いがお互いに影響を与えやすいため，行き先の選択も類似する．また，観光客か，居住者かで場所の選択は異なると考えられる．実際に，情報推薦に友人関係を考慮する研究も提案されており [70]，これらの研究を参考にして，誰にどのような情報を提示すべきかを整理し，深めていきたい．

また，提案システムの被験者実験も必要である．過去の移動履歴を用いた予測精度による評価は，提案モデルの妥当性を測るうえで必須のステップだと考える．しかし，推薦システムとして考えると，最初の提示情報をインタラクティブに修正していく機能も含め，まだ必要な機能は多い．被験者実験を通じて問題点の洗い出しを行いブラッシュアップさせながら，より，推薦システムとしての完成度を高めていきたい．

第6章　結論

本論文では，ソーシャルメディアに存在する人々の体験を構造化し，さらに，蓄積した体験データに埋もれている有用な傾向を抽出し，個人や企業の意志決定に役立てるまでの処理プロセスを支援する体験マイニング技術を提案し，その有効性の検証を行った．

第 1 章では，ソーシャルメディアに存在する体験情報の利活用に関する社会的背景および学術的背景について述べ，本研究の目的を明らかにした．さらに，個人の行動計画や，企業活動で生じる意志決定に，ソーシャルメディア上の人々の体験情報を活用するプロセスの現状，及び問題点と，その問題点解決のための研究課題を整理し，それを実現するための大まかなアプローチについて論じた．

第 2 章では，第 3 章〜5 章で提案した内容に関連する研究について整理し，ソーシャルメディアデータからのマイニング研究の中での，また，ソーシャルメディアデータを活用したシステムの中での，本研究の位置づけを述べた．

第 3 章では，自然言語で記述された非構造なブログデータを対象とし，体験情報を表現する最小構成要素として，時間属性，場所属性，行動属性の組合せから成る情報を抽出する手法を述べた．提案手法においては，フィルモアの格文法解析，動詞の意味解析をすることで，行動内容を示す表現を順に取捨選択していき，さらに，全ブログデータ中で時間，空間，行動属性情報の共起しやすい組合せをアソシエーションルールとして抽出し，行動属性情報の抽出処理に反映させることで，精度の高い体験情報抽出を実現した．評価実験においては，係り受け解析により，動詞，名詞句，格助詞の組合せ情報を抽出するベースライン手法と比較して，大きな精度向上を確認した．また，提案技術のアプリケーションとして，体験情報を構成する属性を指定することで，柔軟に人々の体験情報を検索可能な体

験ブログマップ（Blog Map of Experiences）を提案し，その機能の詳細を述べた．

第 4 章では，体験情報抽出手法によって構造化した大量の体験情報集合をもとに，価値ある情報を抽出する仕組み 2 つのシナリオに基づいて検討した．第 1 のシナリオでは，ある時間に，ある場所を人々が訪れる理由や目的に関心を持つ，調査会社やマーケッターをユーザとして想定し，都市における人々の移動や集中の理由を説明する情報として，“ある特定の状況（時間，空間）において，人々が特徴的にしている行動” を表現するルールを，“興味深い知識” として抽出する手法を提案した．提案手法は，人間の体験を構成する属性の中でも特に 5 属性（時間，空間，動作，対象，感情）をブログから抽出する．次に，そのように生成した構造化データから，支持度と確信度を低い値に設定し，少数派なものも含めたアソシエーションルールを幅広く抽出する．さらに，特定の時間や空間条件において，行動の出現傾向がどの程度変化するかを評価することでルールの価値を評価し，ランキングする．評価実験においては，データマイニング分野で提案されている様々な指標との比較において，提案手法の有効性を確認した．

第 2 のシナリオでは，旅行ガイドブックや Web 検索エンジンなどのメディアを用いて地域情報を収集したことのあるユーザを想定し，人々に認知されている行動と，都市で（ソーシャルメディア上で）実際に人々がしている行動との差異発見につながるアソシエーションルールを選択する仕組みを提案した．提案手法においては，メディアの種類，性質によって情報の出現傾向がどの程度変化するか，どの程度異なるかを評価する．評価実験においては，過去に観光したことのある都市に関する新しい情報を探し出すタスクにおいて，提案法の有効性を示した．今後は，人間の主観的な表現を含め，ソーシャルメディア上における都市の消費者の反応をより深く分析していく必要がある．具体的には，感情属性，評価属性，成功/失敗属性など，主観属性の抽出，推定精度向上が課題である．

第 5 章では，過去の他人の体験情報を利用し，自動で実世界の行動拡張を行う取り組みとして，写真共有サイトのジオタグ情報を利用したトラベルルート推薦技術を提案した．提案手法においては，ユーザが過去にどの場所を訪れたか，を示す移動履歴をもとにユーザがどのような特徴を持つ場所を好むかを分析し，さ

らに，ユーザの現在地からのアクセシビリティを考慮しながら，ユーザが次に行く確率の高い場所を予測する．また，ユーザ自身の空き時間を入力とすることで，単一の場所としてではなく，より具体的な旅行計画（トラベルルート）としてユーザに情報提示を行う．評価実験においては，各ユーザの移動をどの程度予測できるか，を評価指標として，提案モデルの妥当性を確認した．さらに，ユーザがどのような特徴を持つ場所を好むかを移動履歴から推定する問題を深掘りし，ジオトピックモデルを提案した．

本研究は，ソーシャルメディアに反映された人々の体験を分析する最初の試みである．人々の体験を分析することで得られる有用な知識とは何かを議論し，その知識を効率的に抽出するための仕組みを構築した点に学術的な貢献があると考える．今後は，技術によって実現するシステムが実用に足るかどうか，という観点で技術の価値を評価していく姿勢を強める必要があると考える．予測精度や抽出精度をどの程度まで上げれば良いか，あるいは，他に必要な機能や要件はないかを明確にするために，本論文で提案したシステムを実際のユーザに使ってもらう取り組みを行っていきたい．

参考文献

[1] 立石健二, 石黒義英, 福島俊一：インターネットからの評判情報検索, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 101, No. 189, pp. 75–82 (2001).

[2] Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K. and Fukushima, T.: Collecting evaluative expressions for opinion extraction, in *Natural Language Processing (IJCNLP 2004)*, pp. 596–605 (2005).

[3] Hu, M. and Liu, B.: Mining opinion features in customer reviews, in *Proceedings of the 19th Conference on Artificial Intelligence (AAAI 2004)*, pp. 755–760 (2004).

[4] Hu, M. and Liu, B.: Mining and summarizing customer reviews, in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pp. 168–177 (2004).

[5] Liu, B., Hu, M. and Cheng, J.: Opinion observer: analyzing and comparing opinions on the web, in *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*, pp. 342–351 (2005).

[6] 安田宜仁, 平尾努, 鈴木潤, 磯崎秀樹：ブログ作者の居住域の推定, 自然言語処理学会第 12 回年次大会 (NLP 2006) (2006).

[7] Jindal, N. and Liu, B.: Mining comparative sentences and relations, in *Proceedings of the 21th Conference on Artificial Intelligence (AAAI 2006)*, Vol. 22, pp. 1331–1336 (2006).

[8] Jindal, N. and Liu, B.: Identifying comparative sentences in text documents, in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pp. 244–251 (2006).

[9] Kurashima, T., Bessho, K., Toda, H., Uchiyama, T. and Kataoka, R.: Ranking entities using comparative relations, in *Proceedings of the 19th International Conference on Database and Expert Systems Applications (DEXA 2008)*, pp. 124–133 (2008).

[10] Plutchik, R.: *Emotion: A psychoevolutionary synthesis*, Harper & Row New York (1980).

[11] Kumamoto, T. and Tanaka, K.: Proposal of impression mining from news articles, in *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, pp. 901–910 (2005).

[12] 福原知宏, 中川裕志, 西田豊明：感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出, 人工知能学会全国大会 (2006).

[13] 乾健太郎, 原一夫：経験マイニング: Web テキストからの個人の経験の抽出と分類, 言語処理学会第 14 回年次大会論文集 (NLP 2008), pp. 1077–1080 (2008).

[14] Inui, K., Abe, S., Hara, K., Morita, H., Sao, C., Eguchi, M., Sumida, A., Murakami, K. and Matsuyoshi, S.: Experience mining: Building a large-scale database of personal experiences and opinions from web documents, in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2008)*, pp. 314–321 (2008).

[15] Park, K. C., Jeong, Y. and Myaeng, S. H.: Detecting experiences from weblogs, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pp. 1464–1472 (2010).

[16] Kleinberg, J.: Bursty and hierarchical structure in streams, *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397 (2003).

[17] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学：document stream における burst の発見, 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 23, pp. 85–92 (2004).

[18] 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕：blog ページの自動収集と監視に基づくテキストマイニング, 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01 (2004).

[19] 南野朋之, 鈴木泰裕, 藤木稔明, 奥村学：blog の自動収集と監視, 人工知能学会論文誌, Vol. 19, No. 6, pp. 511–520 (2004).

[20] Fujimura, K., Toda, H., Inoue, T., Hiroshima, N., Kataoka, R. and Sugizaki, M.: BLOGRANGER —a multi-faceted blog search engine, in *Proceedings of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2006).

[21] 戸田浩之, 藤村考, 井上孝史, 廣嶋伸章, 杉崎正之, 片岡良治, 奥雅博：目的指向型ブログ検索システム BLOGRANGER の提案およびユーザ評価, 情報処理学会論文誌: データベース, Vol. 48, No. 14, pp. 132–151 (2007).

[22] Rodriguez, M. G., Leskovec, J. and Krause, A.: Inferring networks of diffusion and influence, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pp. 1019–1028 (2010).

[23] Rodriguez, M. G., Balduzzi, D. and Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks, in *Proceedings of the International Conference on Machine Learning (ICML 2011)*, pp. 561–568 (2011).

[24] Iwata, T., Shah, A. and Ghahramani, Z.: Discovering latent influence in online social activities via shared cascade poisson processes, in *Proceedings of the 19th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013), pp. 266–274 (2013).

[25] Rodriguez, M. G., Leskovec, J. and Schölkopf, B.: Structure and dynamics of information pathways in online media, in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM 2013)*, pp. 23–32 (2013).

[26] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: Structure and evolution of blogspace, *Communications of the ACM*, Vol. 47, No. 12, pp. 35–39 (2004).

[27] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A.: On the bursty evolution of blogspace, in *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*, pp. 159–178 (2005).

[28] Bar-Ilan, J.: An outsider's view on topic-oriented blogging, in *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, pp. 28–34 (2004).

[29] Fujimura, K. and Tanimoto, N.: The eigenrumor algorithm for calculating contributions in cyberspace communities, in *Trusting Agents for Trusting Electronic Societies*, pp. 59–74 (2005).

[30] Nakajima, S., Tatemura, J., Hara, Y., Tanaka, K. and Uemura, S.: Identifying agitators as important blogger based on analyzing blog threads, in *Proceedings of the 8th Asia-Pacific Web Conference (APWeb 2006)*, pp. 285–296 (2006).

[31] Horozov, T., Narasimhan, N. and Vasudevan, V.: Using location for personalized POI recommendations in mobile environments, in *Proceedings of the International Symposium on Applications and the Internet (SAINT 2006)*, pp. 124–129 (2006).

[32] Ashbrook, D. and Starner, T.: Using GPS to learn significant locations and predict movement across multiple users, *Personal and Ubiquitous Computing (UbiComp 2003)*, Vol. 7, No. 5, pp. 275–286 (2003).

[33] Krumm, J. and Horvitz, E.: Predestination: Inferring destinations from partial trajectories, in *Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp 2006)*, pp. 243–260 (2006).

[34] Zheng, Y., Zhang, L., Xie, X. and Ma, W.-Y.: Mining interesting locations and travel sequences from GPS trajectories, in *Proceedings of the 18th International Conference on World Wide Web (WWW 2009)*, pp. 791–800 (2009).

[35] De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R. and Yu, C.: Constructing travel itineraries from tagged geo-temporal breadcrumbs, in *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pp. 1083–1084 (2010).

[36] De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R. and Yu, C.: Automatic construction of travel itineraries using social breadcrumbs, in *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (ACM HT 2010)*, pp. 35–44 (2010).

[37] Kennedy, L., Naaman, M., Ahern, S., Nair, R. and Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections, in *Proceedings of the 15th ACM International Conference on Multimedia (ACM MM 2007)*, pp. 631–640 (2007).

[38] Kennedy, L. S. and Naaman, M.: Generating diverse and representative image search results for landmarks, in *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pp. 297–306 (2008).

[39] Ahern, S., Naaman, M., Nair, R. and Yang, J. H.-I.: World explorer: visualizing aggregate data from unstructured text in geo-referenced collections, in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2007)*, pp. 1–10 (2007).

[40] Jaffe, A., Naaman, M., Tassa, T. and Davis, M.: Generating summaries and visualization for large collections of geo-referenced photographs, in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 89–98 (2006).

[41] Crandall, D. J., Backstrom, L., Huttenlocher, D. and Kleinberg, J.: Mapping the world's photos, in *Proceedings of the 18th International Conference on World Wide Web (WWW 2009)*, pp. 761–770 (2009).

[42] Snavely, N., Seitz, S. M. and Szeliski, R.: Photo tourism: exploring photo collections in 3D, *ACM Transactions on Graphics (TOG)*, Vol. 25, No. 3, pp. 835–846 (2006).

[43] Snavely, N., Seitz, S. M. and Szeliski, R.: Modeling the world from internet photo collections, *International Journal of Computer Vision (IJCV)*, Vol. 80, No. 2, pp. 189–210 (2008).

[44] Kalogerakis, E., Vesselova, O., Hays, J., Efros, A. A. and Hertzmann, A.: Image sequence geolocation with human travel priors, in *Proceedings of the 12th International Conference on Computer Vision (ICCV 2009)*, pp. 253–260 (2009).

[45] Cao, L., Yu, J., Luo, J. and Huang, T. S.: Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression, in *Proceedings of the 17th ACM International Conference on Multimedia (MM 2009)*, pp. 125–134 (2009).

[46] Hays, J. and Efros, A. A.: IM2GPS: estimating geographic information from a single image, in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8 (2008).

[47] Herlocker, J. L., Konstan, J. A., Terveen, L. G. and Riedl, J. T.: Evaluating collaborative filtering recommender systems, *ACM Transactions on Information Systems (TOIS)*, Vol. 22, No. 1, pp. 5–53 (2004).

[48] Popescu, A., Grefenstette, G. and Moëllic, P. A.: Gazetiki: automatic creation of a geographical gazetteer, in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*, pp. 85–93 (2008).

[49] Backstrom, L., Sun, E. and Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity, in *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pp. 61–70 (2010).

[50] Shaw, B., Shea, J., Sinha, S. and Hogue, A.: Learning to rank for spatiotemporal search, in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM 2013)*, pp. 717–726 (2013).

[51] Agrawal, R., Imieliński, T. and Swami, A.: Mining association rules between sets of items in large databases, *ACM SIGMOD Record*, Vol. 22, No. 2, pp. 207–216 (1993).

[52] Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules, in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994)*, Vol. 1215, pp. 487–499 (1994).

[53] Fillmore, C. J.: 格文法の原理: 言語の意味と構造, 三省堂 (1975).

[54] Matsumoto, Y., Kitauchi, A., Yamashita, T. and Hirano, Y.: Japanese morphological analysis system ChaSen version 2.0 manual, in *NAIST Techinical Report* (1999).

[55] Kudo, T. and Matsumoto, Y.: Japanese dependency analysis using cascaded chunking, in *Proceedings of the 6th Conference on Natural Language Learning-Volume 20*, pp. 1–7 (2002).

[56] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦：日本語語彙大系, 岩波書店 (1997).

[57] Geng, L. and Hamilton, H. J.: Interestingness measures for data mining: A survey, *ACM Computing Surveys (CSUR)*, Vol. 38, No. 3, p. 9 (2006).

[58] Smyth, P. and Goodman, R. M.: An information theoretic approach to rule induction from databases, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, Vol. 4, No. 4, pp. 301–316 (1992).

[59] Fuchi, T. and Takagi, S.: Japanese morphological analyzer using word co-occurrence: JTAG, in *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1 (COLING 1998)*, pp. 409–413 (1998).

[60] 齋藤邦子, 鈴木潤, 今村賢治：CRF を用いたブログからの固有表現抽出, 言語処理学会第 13 回年次大会 (NLP2007) (2007).

[61] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research (JMLR)*, Vol. 3, pp. 993–1022 (2003).

[62] Hofmann, T.: Probabilistic latent semantic indexing, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pp. 50–57 (1999).

[63] Hofmann, T.: Collaborative filtering via gaussian probabilistic latent semantic analysis, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR 2003)*, pp. 259–266 (2003).

[64] Iwata, T., Watanabe, S., Yamada, T. and Ueda, N.: Topic tracking model for analyzing consumer purchase behavior., in *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, Vol. 9, pp. 1427–1432 (2009).

[65] Gildea, D. and Hofmann, T.: Topic-based language models using EM, in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECIt 1999)*, pp. 2167–2170 (1999).

[66] Dijkstra, E. W.: A note on two problems in connexion with graphs, *Numerische Mathematik*, Vol. 1, No. 1, pp. 269–271 (1959).

[67] Navarro, G.: A guided tour to approximate string matching, *ACM Computing Surveys (CSUR)*, Vol. 33, No. 1, pp. 31–88 (2001).

[68] Dempster, A. P., Laird, N. M., Rubin, D. B., et al.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Vol. 39, No. 1, pp. 1–38 (1977).

[69] Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, Vol. 45, No. 1-3, pp. 503–528 (1989).

[70] Yang, S.-H., Long, B., Smola, A., Sadagopan, N., Zheng, Z. and Zha, H.: Like like alike: joint friendship and interest propagation in social networks, in *Proceedings of the 20th International Conference on World Wide Web (WWW 2011)*, pp. 537–546 (2011).

謝辞

本研究の遂行ならびに論文の作成にあたり，懇切なる御指導を賜りました京都大学大学院情報学研究科教授 田中克己先生に謹んで感謝の意を表します．主体性を重んじつつも豊かな発想力で私を牽引してくださった田中先生のおかげで，体験マイニングという私にとってとても大切な研究に出会うことができました．今後も御指導のほどよろしくお願い致します．

本論文をまとめるにあたり，有益な御助言と御教示を賜りました京都大学大学院情報学研究科教授 石田亨先生，西田豊明先生に心より謝意を申し上げます．石田先生には“考えること”の大切さを学びました．石田先生の言葉ひとつひとつがとても重く，社会の未来を考える研究者としての私の成長を促すものでした．西田先生は研究の発展についてとても親身になって議論してくださいました．西田先生からご提案頂いた数多くの斬新なアイディアを今後の研究に活かします．

大妻女子大学社会情報学部の藤村考先生には，NTT サイバーソリューション研究所にて大変お世話になりました．2005 年に始めた体験マイニング研究をここまで続けてこれたのも藤村先生が支えてくださったおかげです．また，企業研究者としての姿勢，楽しさを藤村先生から学びました．厚く謝意を申し上げます．

本研究の遂行ならびに論文の作成にあたり御協力いただいた，京都大学大学院情報学研究科田中克己研究室の皆様に感謝致します．また，諸般の事務手続きを行っていただいた京都大学大学院情報学研究科田中克己研究室秘書の佐藤香織さん，白石真弓さん，足羽美恵さんに感謝致します．

最後に，私の選択を尊重し支えてくれた家族に感謝します．

2014 年 9 月 倉島 健

研究業績

主要論文

[1] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka
"Blog map of experiences: Extracting and geographically mapping visitor experiences from urban blogs"
Proceedings of the 6th International Conference on Web Information Systems Engineering (WISE 2005), pp.496-503, November 2005.

[2] Taro Tezuka, Takeshi Kurashima, and Katsumi Tanaka
"Toward tighter integration of web search with a geographic information system"
Proceedings of the 15th International Conference on World Wide Web (WWW 2006), pp.277-286, May 2006.

[3] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka
"Mining and visualizing local experiences from blog entries"
Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006), pp.213-222, September 2006.

[4] Takeshi Kurashima, Ko Fujimura, and Hidenori Okuda
"Discovering association rules on experiences from large-scale blog entries"
Proceedings of the 30th European Conference on Information Retrieval (ECIR 2009), pp.546-553, April 2009.

[5] 倉島 健, 藤村 考, 奥田 英範
“大規模テキストからの経験マイニング”
電子情報通信学会論文誌: D, Vol.J92-D, No.3, pp.301-310, 2009 年 3 月.

[6] Takeshi Kurashima, Tomoharu Iwata, Go Irie and Ko Fujimura
“Travel route recommendation using geotagged photos”
Knowledge and Information Systems, Vol.37, No.1, pp.37-60, October 2012.

[7] Takeshi Kurashima, Tomoharu Iwata, Go Irie, and Ko Fujimura
“Travel route recommendation using geotags in photo sharing sites”
Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), pp.579-588, October 2010.

[8] Takeshi Kurashima, Tomoharu Iwata, Takahide Hoshide, Noriko Takaya, and Ko Fujimura
“Geo topic model: Joint modeling of user’s activity area and interests for location recommendation”
Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM 2013), pp.375-384, February 2013.

[9] 倉島 健, 岩田 具治, 星出 高秀, 高屋 典子, 藤村 考
“行動範囲と興味の同時推定モデルによる地域情報推薦”
情報処理学会論文誌: データベース (TOD), Vol.6, No.2, pp.30-41, 2013 年 3 月.

[10] Takeshi Kurashima, Tomoharu Iwata, Noriko Takaya, and Hiroshi Sawada
“A probabilistic behavior model for discovering unrecognized knowledge”
Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013), pp.1097-1102, December 2013.

[11] Takeshi Kurashima, Tomoharu Iwata, Noriko Takaya, and Hiroshi Sawada
“Probabilistic latent network visualization: Inferring and embedding diffusion networks”
Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2014), August 2014.

[12] 倉島 健, 別所 克人, 戸田 浩之, 内山 俊郎, 片岡 良治, 奥 雅博
“比較評価情報に基づくランキング手法”
日本データベース学会 Letters, Vol.6, No.1, pp.5-8, 2007 年 6 月.

[13] Takeshi Kurashima, Katsuji Bessho, Hiroyuki Toda, Toshio Uchiyama, and Ryoji Kataoka
“Ranking entities using comparative relations”
Proceedings of the 19th International Conference on Database and Expert Systems Applications (DEXA 2008), pp.124-133, September 2008.

学術報告

[1] 倉島 健, 手塚 太郎, 田中 克己
“Blog からの街の話題抽出手法の提案”
電子情報通信学会第 16 回データ工学ワークショップ (DEWS 2005).

[2] 倉島 健, 手塚 太郎, 田中 克己
“街 Blog からの体験表現抽出とその空間的提示手法の提案”
情報処理学会 DBS 研究会/電子情報通信学会 DE 研究会 合同ワークショップ (DBWS 2004).

[3] 倉島 健, 藤村 考, 奥田 英範
“大規模テキストからの経験マイニング”
電子情報通信学会第 19 回データ工学ワークショップ (DEWS 2008).

[4] 倉島 健
“人々の経験を活かすための経験マイニング”
情報処理学会 情報学基礎研究会報告 (jDB ワークショップ).

[5] 倉島 健, 岩田 具治, 入江 豪, 藤村 考
“写真共有サイトにおけるジオタグ情報を利用したトラベルルート推薦”
電子情報通信学会技術研究報告 ライフインテリジェンスとオフィス情報システム研究会 (LOIS 研究会).