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

A Study on Experience Mining from Social Media Data

倉島 健・

Takeshi KURASHIMA

1. はじめに

現在のところ，ブログ，TwitterやFlickr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のが現状である．

・本会員　京都大学大学院情報学研究科博士後期課程
kurashima.takeshi@lab.ntt.co.jp

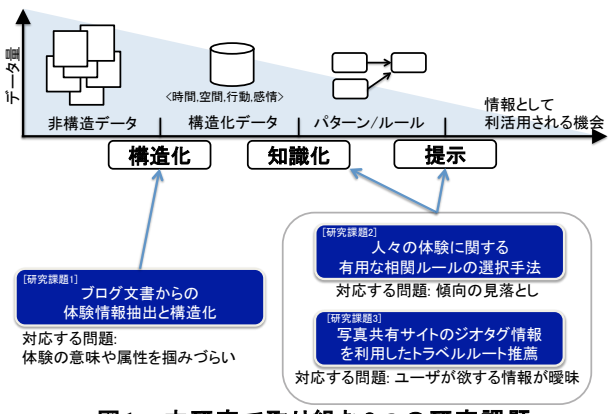

図1　本研究で取り組む3つの研究課題

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どであり，“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．この目的を達成するために3つの研究課題に取り組んだ(図1)．まず，2章で体験情報について定義したのち，3章から5章で各取り組みについて述べる．

2. 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．本研究では主に，動作主属性と評価属性を除く，5属性について考えるが，5章で述べる第3の研究テーマに関しては，体験の順序関係，移動経路を分析するために，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用いる．

3．ブログ文書からの体験情報抽出と構造化

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，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（図2）．たとえば“南禅寺で11月にする行動”を知りたい場合，空間属性値として“南禅寺”を，時間属性値として“11月”を指定して検索ボタンを押すと，“湯豆腐を食べる”などの行動情報がマップインタフェース上の吹き出しの中に提示される．つまり，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．

図2　体験ブログマップのユーザインタフェース

4．人々の体験に関する有用な相関ルールの選択手法

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，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，2つの利用シナリオを想定し，それぞれで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検討する．

第1の利用シナリオ：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）をブログデータから抽出する．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．

第2の利用シナリオ：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を提案する．提案手法は，Web検索エンジン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評価し，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，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．

5．写真共有サイトのジオタグ情報を用いたトラベルルート推薦

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，ユーザの現在地からのアクセシビリティを考慮しながら，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（図3）．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．

図3　トラベルルート推薦システムのユーザインタフェース

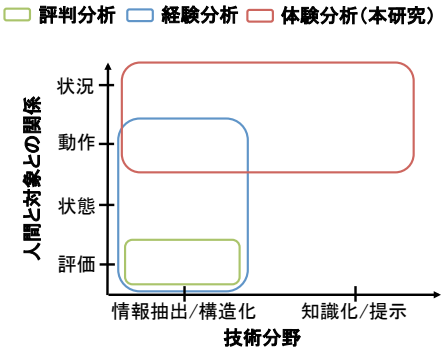

図4　体験マイニングが扱う技術分野とその周辺

6．関連研究

ソーシャルメディアデータのマイニング技術は，コンテンツ分析技術と，リンク構造分析技術とに大別できる．本研究はコンテンツ分析技術に関するものであり，その中でも特に関連が深い評判情報抽出，経験情報抽出の既存研究について述べる．また，蓄積した体験情報を利活用する仕組み，システムに関する既存研究についても述べる．

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[1,2]．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．

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[3]．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．乾らの経験マイニング技術と比較した場合の本研究（体験マイニング）の特徴は以下の2点である．

1. 時空的な背景（状況）も含めた都市行動の分析
2. 蓄積した（抽出した）情報からの知識発見/情報提示までを扱う技術領域

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である．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り，それを自然な形で表現するものとして体験情報を定義している．具体的には，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こうとしている．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．

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ユーザに対して提示する処理までの領域を広く扱う．図4は，本研究と評判情報抽出技術，経験情報抽出技術が扱う領域とを比較したものである．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．

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存在する人々の体験情報を用いて，情報推薦，情報検索，コンテンツブラウジング，自動アノテーションなどに関連して様々な試みが存在する[4, 5, 6]．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（3章），第2（4章）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，既存研究の多くが単一の場所をユーザに推薦することを研究課題としていたのに対して，本研究の第3の研究テーマ（5章）はトラベルルートとしての推薦を実現し，ユーザの旅行計画の自動化を実現した．

6．おわりに

本稿では，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の概要を紹介した．ソーシャルメディアに反映された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した点に本研究の学術的な貢献がある．

［文献］

[1] Liu, B., Hu, M. and Cheng, J.: “Opinion observer: analyzing and comparing opinions on the web”, Proceedings of the 14th International Conference on World Wide Web (WWW 2005), pp. 342–351 (2005).
[2] Hu, M. and Liu, B.: “Mining and summarizing customer reviews”, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 168–177 (2004).
[3] 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”, Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2008), pp. 314–321 (2008).
[4] Horozov, T., Narasimhan, N. and Vasudevan, V.: “Using location for personalized POI recommendations in mobile environments”, Proceedings of the International Symposium on Applications and the Internet (SAINT 2006), pp. 124–129 (2006).
[5] Kennedy, L. S. and Naaman, M.: “Generating diverse and representative image search results for landmarks”, Proceedings of the 17th International Conference on World Wide Web (WWW 2008), pp. 297–306 (2008).
[6] Cao, L., Yu, J., Luo, J. and Huang, T. S.: “Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression”, Proceedings of the 17th ACM International Conference on Multimedia (ACM MM 2009), pp. 125–134 (2009).

倉島 健 Takeshi KURASHIMA
2006 年京都大学大学院情報学研究科博士前期課程修了．同年，NTT サイバーソリューション研究所に入社．2014 年京都大学大学院情報学研究科博士後期課程修了．現在，NTT サービスエボリューション研究所にて，データマイニング，テキストマイニング，情報推薦等の研究に従事．博士（情報学）．情報処理学会正会員．電子情報通信学会正会員．日本データベース学会正会員．