ブログマイニングからの行動データの抽出・分析可能性とアンケート調査との比較

佐々木　邦明[1]・紀藤　舞華[2]・山崎　慧太[3]

[1]正会員　山梨大学教授　医学工学総合研究部（〒400-8511 山梨県甲府市武田4-3-11）
E-mail: sasaki@yamanashi.ac.jp

[2]正会員　中日本ハイウェイ・メンテナンス中央株式会社（〒192-0083　東京都八王子市旭町12-4）

[3]正会員　東海旅客鉄道株式会社　社員研修センター（静岡県三島市文教町１-4-19）

本研究は交通行動データの新たなソースとしてインターネット上のブログの活用を検討するものであり，これからの交通計画における価値のあるデータの抽出可能性を検証する．そのために同じ地域で実施されたアンケート調査結果との比較を通じてその特性を明らかにする．事例研究として八ヶ岳南麓地域を対象としたブログデータを収集し，個人属性等の推定可能性と，ブログに基づいた行動特性の抽出を行い，既存データとの比較を行った．その結果ブログにはアンケートでは見られない特性を抽出することができ，魅力ある観光地形成のための有益なデータが得られる可能性を示した．

Key Words : *travel behavior survey, blog, text mining, tourism*

１．はじめに

よりよい観光地をつくるためには，観光地の特色を活かして常に新しい価値を創造する取り組みがなされる必要があるといわれる．観光地における価値の創造には，観光客のニーズに応えるだけでなく，観光地での新たな観光形態や資源の発掘を行うことが求められる．しかし，そのような新たな価値は既存のものとまったくかけ離れたものとは限らず，既存の観光客の行動の中にも隠れていると考えられる．それを活かすためには，平均的な観光客像を描き出すのではなく，広いバリエーションのある行動から，特定の観光行動に関連する要素を拾い上げていくものになる．そのような観点に立って観光客の調査を考えたとき，従来型の正規分布的母集団の仮定からのランダムサンプリングを前提とした調査とは異なり，日々の生活を含め，観光だけでない時系列的な調査が必要になってくると考えられる．しかし，そのような調査は，被験者の自発的な協力無くして実行するのは困難である．また，元来観光行動は希な行動であるため，居住地ベースの調査は困難なため，主に観光地での配布によることが多い．このような調査は配布場所の選定にサンプルの特性が依存し，そもそもランダムサンプリング的なアプローチ自体が困難である．

本研究では，そのような背景のもと，インターネット空間上に多数存在するブログ（blog）サービスに着目した．ブログとは，書き手の個人的な体験や日記，または関心をもった製品やニュースについて書き手なりの意見や感想などのコメントを書いた記事を Web 上に掲載するサイトである．ブログは使い方が比較的容易であるため，一般の多くの人々からの情報発信が盛んになった．その発信された情報の中には，リアルタイム性のある新鮮な情報や「生の声」などといわれる率直な評価や意見が含まれている．また，個人の時系列的な生活の一部や意見・態度などが記録されているため，先に述べた価値創造型の観光情報抽出において有効な要素が存在する．そこで本研究は観光のエントリーをもつブログをデータソースとして，観光行動情報の抽出と活用を検討するものである．

２．ブログの特性とデータ抽出

ブログから企業の活動等に有益な情報を抽出する研究がこれまで行われており，それらの手法を総称してブログマイニング手法という．現在企業では製品開発や企業活動にブログマイニング手法が活用されていると言われている [1)]．ブログマイニングの基本的な方法論は，ブログの文章データを単語や文節で区切り，それらの同時出現頻度などを解析することで，文章から有用な情報を取り出すいわゆるテキストマイニング法である．観光ブ

ログをブログマイニング手法で分析するにあたって，アンケート調査とブログマイニング手法にはどのような違いがあるかについて簡単に表－１に示す．

表－１　アンケート調査とブログマイニングの特性比較

	入込み調査	ブログマイニング
サンプリング	■観光施設等でアンケートを配布	■ネット上にあるブログより抽出
対象者(観光地の訪問者)	■アンケート配布場所へ訪れた人 ■アンケートに回答した人	■インターネットが使える環境がある人 ■観光行動に関するブログを書いた人
調査期間	■予算に応じた期間	■収集数に応じる
調査項目	■アンケートの設計に依存 ■個人の時系列的な要素は得づらい	■必ずしも行動総てが記録されない ■個人属性はエントリー等からの推定

ある特定の観光地についてアンケートを用いて調査を実施する場合には，一般に観光施設等でアンケートを配布することになる．結果として周遊経路上に配布場所が含まれない観光客は調査の対象外になる．その意味では，配布地点を通過したという条件付の部分集合からのサンプリングになる．ランダムサンプリング化するためには，各施設での入込みなどの総数に基づいて補正係数を算定することなどが必要となる．一方ブログマイニングでは，Web 上にあるブログの中から抽出なので，配布場所等には依存しないが，もともと「ブログを書く」という行為自体に，インターネットを使える環境にあり，ブログを開設しているという前提条件がある．そのためこれらの条件を満たした部分集合からのサンプリングであり，母集団代表性の確保は困難である．また補足的な総数調査などから補正することも，ブログには必ずしも行動全てが記述されているとは限らないため困難である．それ以前に，ブログオーナーからの抽出もランダム化は非常に困難であり，母集団からのランダムサンプリングによる代表的な行動の抽出は一般的には不可能であると考えられる．

３．ブログデータの収集と観光行動調査

今回事例研究の対象として，図－1 に示すような山梨県と長野県にまたがる八ヶ岳南麓地域を選定した．この地域は 2010 年には観光圏の指定を受け，圏域でのネットワーク的な観光施設の整備が進展している地域である．この地域では 2009 年度に山梨大学交通工学講座が北杜市観光課の協力のもとに観光客対象のアンケート調査を行っているため，そのデータも比較対象として利用可能である．

（1）ブログの検索と検索エンジンによる差

この地域に関連する観光ブログを収集するために，本研究では検索エンジンを利用した．検索エンジンを利用した理由は，それ以外でブログを収集することが非常に困難であることが理由であるが，スパムブログ[1]の排除に役に立つことも重要な要素である．検索エンジンはキーワードからの適切な検索結果導出が検索エンジンの評価を決めるため，各社が独自のアルゴリズムに基づいたフィルタリングを行っている．そのフィルターは系統的な選別を行っていると考えられるため，上位にリストされるブログのランダム性はまったく保証されない．しかし，その選別はスパムブログ等を排除することや，信頼性の高いサイトを上位に表示することを目的としているため，その上位から利用することは，ブログの信憑性の低さをカバーすると考えられる．

ブログを収集するために，八ヶ岳南麓地域を代表であろうキーワードを複数選択し，それらの組み合わせによって複数のエンジンを用いて検索を行った．観光旅行のエントリーがあるブログ抽出の効率の高さから，最終的には「八ヶ岳」と「行った」をキーワードとして Google および Yahoo! Japan を用いて検索を行い[2]，192 ブログを収集した．上記キーワードで抽出されたブログのうち両エンジンともに上位でリスとされ，重複して検索されたものは 80 であった．

図－１　八ヶ岳南麓地域

（2）観光行動調査

ブログからの行動データの抽出と比較するために，既存の観光行動調査データを用いた．この調査データは山梨大学交通工学講座が北杜市観光課と協力して実施したものである．その概要を表－２に示す．調査日は平成 21 年 10 月であり，当日は三連休の中日で紅葉が始まっ

た季節であって天候にも恵まれ大勢の観光客が訪れた．

表－２　観光客周遊調査の概要[2]

調査日	平成 21 年 10 月 11 日（日）
調査方法	街頭配布・郵便回収
配布部数	1181 部
回収部数	501 部（回収率 42.4%）

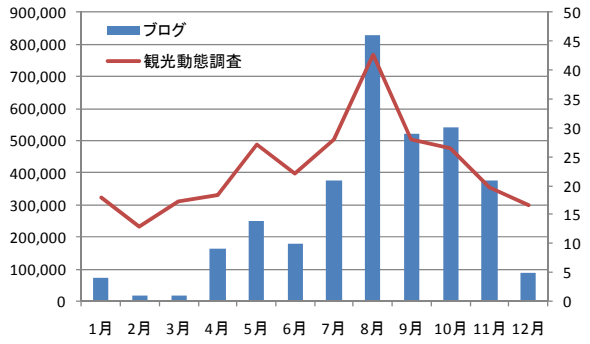
この調査の配布は観光課および地域観光業者等にアドバイスを求め，東沢大橋，清泉寮，スパティオ小淵沢，道の駅はくしゅうの４カ所で行った．配布地点は，より代表性を増すために，多くの観光客が最低１カ所は来訪するであろうとのアドバイスの元に選択された．調査の内容は個人属性，旅行形態およびその日の周遊行動であった．

４．観光ブログの解析とアンケートとの比較

ブログの解析は KHCoder[3]と茶筅 [4]を用いてブログを構成する文章を単語単位に分解し，そこからキーワード的にブログの特性を抽出した．

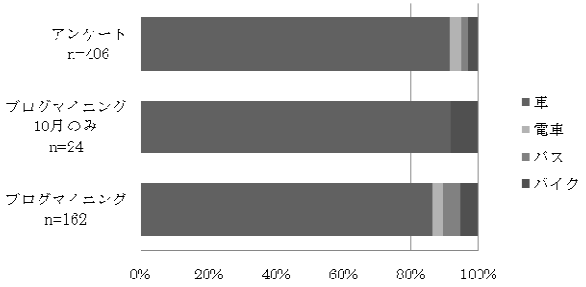
（1）ブログから推測される観光行動の基礎的特性

ブログエントリーの日付から，八ヶ岳南麓への来訪日時を推定した．その月別の分布を示したのが図－２である．また，山梨県が実施している観光客動態調査の八ヶ岳地域の H21 年度月別観光客数結果をあわせて示した．ブログ数が少ないため統計的な同一性は示す事ができないが，ブログは８月が突出している一方，冬にほとんど見られない．これは動態調査では冬の観光客にはウィンタースポーツが含まれるが，我々の行った検索条件では，八ヶ岳を冠したスキー場等が無いため含まれなかった事が考えられる．



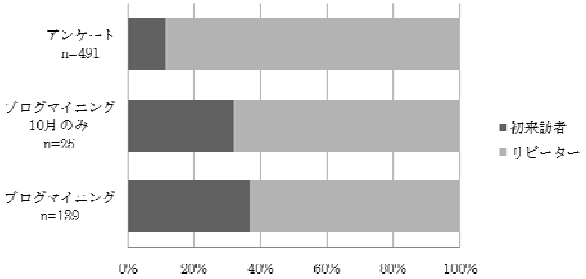
図－２　推定来訪月と山梨県観光動態調査

来訪時の利用交通手段について分析した結果が図－３である．交通手段に関連する「自動車」，「駐車場」，「高速道路」「カーナビ」「バイク」等の単語の有無で交通手段の判定を行ったが，その判定が困難なブログが 15%程度存在した．この交通手段分担率を比較するとバイクの割合がアンケートと比較して高くなった．ツーリングは一つのカテゴリーを形成するもので，ブログに取りあげられる可能性が高かったと考えられる．



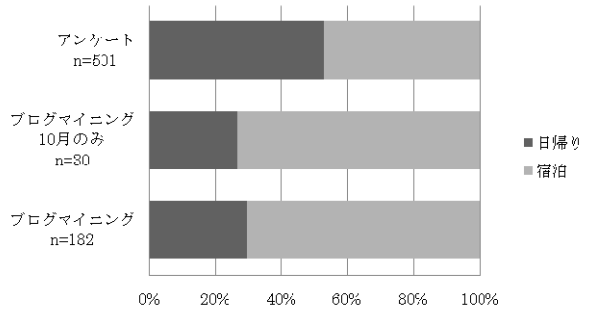
図－３　交通手段の推定結果

観光行動で重視される特性の一つにリピーターの存在がある．繰返し訪れる観光客を増加させることが一つの目的になっているが，ブログから観光客のリピート割合の推定を行った．またそれをアンケートと比較したものが図－４である．「年ぶり」「前回」「初めて」などをキーワードとしてリピーターであるかどうかを判定したが，30%程度が判定できなかった．ブログでは何度も繰り返して来る場合などは，そのようなキーワードが用いられるが，リピートというよりは以前来たことがあるというレベルの場合にはそのようなワードが使われず，リピートと判定されなかった可能性が高い．しかし，現在その特性が注目されるリピーターとは，数年に一度レベルよりはより頻繁に来る層と考えられるため，上記のようなケースはそれ程問題とならない可能性もある．また，今回は「初めて」など初来訪を推測される単語が用いられ，リピート行動と関連する単語が用いられなかった場合には初来訪と判定したが，実際には異なる施設等に以前行ったことがある可能性が高いことも指摘できる．これが来訪者の割合が高くなった理由の一部とも考えられる．また繰返し何度も来ている別荘地の住民などでは高齢者が多くブログというツールと関係が薄いことなど，リピート行動の正確性については検討の余地が大きい．



図－４　リピーターの割合の推定

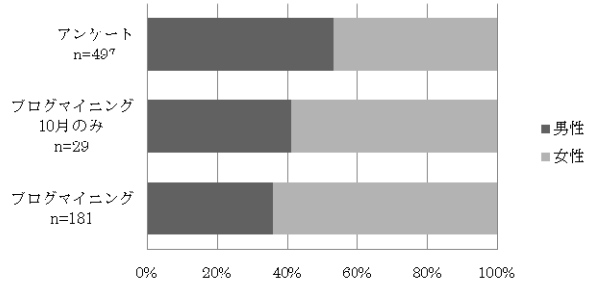
宿泊と日帰りの分類は，連続した日で旅行のエントリーがある，ブログ中に「ホテル」「宿泊」等のキーワードに基づいて判定を行った．その結果５％程度のブログは判定が困難であったが，その他は判定可能であった．その比率を図－５に示すが，アンケートと比較して宿泊の割合が高い．アンケートでは三連休の中日であったにもかかわらず，半数近くが日帰りであったのに対して，ブログでは７０％以上が宿泊であった．これは宿泊旅行はより印象が強いことや，8月の来訪が多く，まとまった休みを取りやすいことが，このような結果の原因と考えられる．



図－５　宿泊・日帰り別の推定

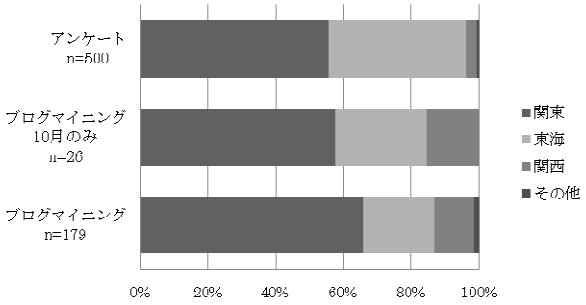
（2）個人属性の推定

個人属性については，旅行に関連するエントリーだけでは判定が困難であるが，その人の１年程度のエントリーを用いることで個人の属性が判定できる．その例としてブログ中に現れるキーワードに基づいて，性別について判定を行った結果を図－６に示す．約 5%程度は判定が困難であった．またアンケート調査との比較を行うと，10 月のみを取り出したとしても女性の割合が高かった．一般にブログのオーナーは女性の比率が高いとされている事がこのような結果になったと考えられる．



図－６　性別の推定結果とアンケートとの比較

居住地については，ブログ中に出現する地名を抽出し，前後の文章から居住地情報を判定し，県レベルで推定を行った．その結果５％程度を除いてほとんどのブログで判定が可能であった．またその地域別の構成比率を示したのが図－７である．これからは，アンケートと比較して，関西方面の比率が高く，東海(中部)の割合が低くなった．これは主に地元である山梨が関東，長野が東海(中部)に割り当てられたため，日常的なレジャーでブログに記載される可能性が低い観光客が減少したためと考えられる．

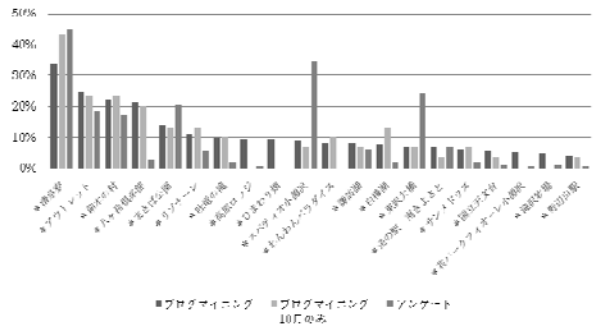


図－７　居住地域別の構成比率

（3）訪問先の分析

訪問先の分布については，形態素解析から出てきた単語群について，同一の施設を指すものを分類し，その上で施設別の来訪を判定した．例えば「萌木の村」という観光施設にはレストラン等が存在し，それらの名前は「ハットウォールデン」「Rock」「ルシャデボワ」などであるため，これらは全て萌木の村と分類するような論理演算子を作成し訪問先を集計した．訪問先の総数は130になったが，その上位20カ所を図－８に示した．あわせてアンケートでの訪問比率を示した．

ア結果としてはアンケートの比率と変わらない施設も多かったが，極端に両者で訪問比率が異なる施設があった．図中でアンケートで比率が高い施設が３カ所あるが，これはいずれもアンケート配布を行った施設であり，代表的な立寄りスポットであると考えられた道の駅などの共通施設には立ち寄らない観光がブログには多く示されていることがわかる．具体的には，アンケートでほとんど見られない訪問先には「八ヶ岳倶楽部」「吐竜の滝」があった．「八ヶ岳倶楽部」「吐竜の滝」はいずれもアウトドア・ガーデニングなど趣味性の強い場所であり，このような場所を訪れる観光客は道の駅・清泉寮等のメジャー観光施設には立ち寄ら無いことがわかる．

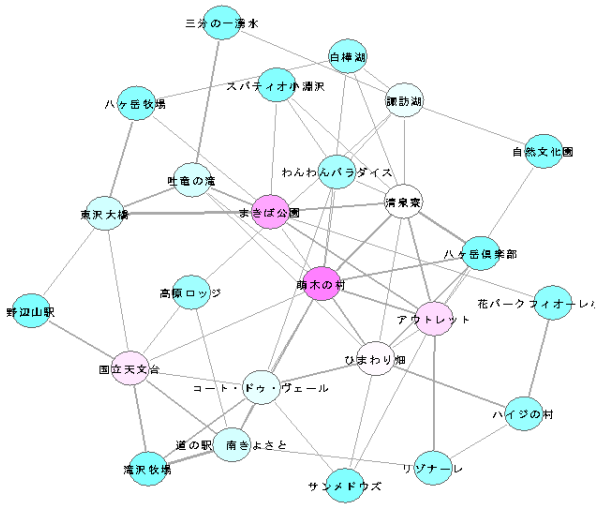


図－８　訪問先上位２０位とアンケートでの訪問比率

（4）観光周遊特性の分析

本章では，ブログエントリーに書かれた内容をもとに観光周遊行動に関するデータの抽出を行う．特に行動データの中から目的地データに着目し，同時に文の中で用いられる比率を計算するJaccard係数を用いて，共起ネットワークを作成した[6]．共起ネットワークでは同時に文章に出てくる割合が高いほど強くネットワークで結ばれるため，どういった観光施設が同時に訪問されるか，また，媒介中心性という概念によって，観光周遊のキーとなっている観光施設が明らかになる．ブログデータを用いて共起ネットワークしたものが図－９である．図では媒介中心性が高いノードは赤く，低いものは水色になっている．また，リンクの太さは共起関係の強さを表している．

このネットワークを見ると，道の駅南きよさとと滝沢牧場の共起関係が最も強く約30%に達していた．その他にも20%を超えるものが4つの組み合わせで確認された．共起性の強いものから上位10を示し，同じようにアンケートで得られた周遊行動から観光施設間の共起性を計算して表－３にまとめた．



図－９　観光目的地の共起ネットワーク

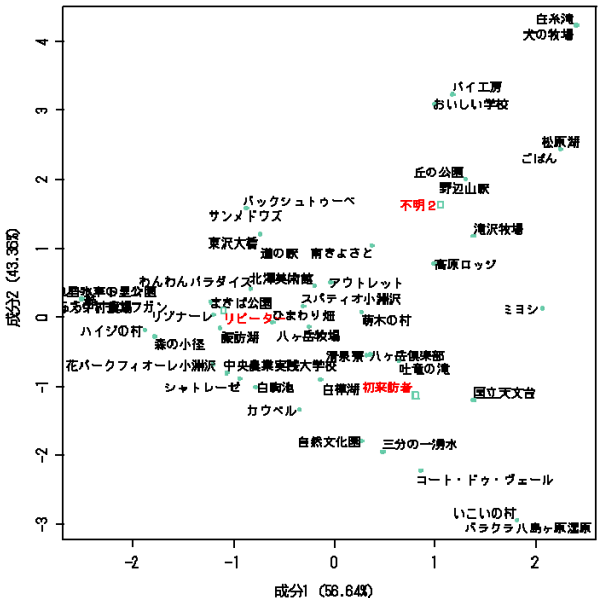
表－３　観光施設間の共起性

観光施設の組み合わせ	ﾌﾞﾛｸﾞ	ｱﾝｹｰﾄ
道の駅南きよさと－滝沢牧場	29.4%	2.4%
東沢大橋－まきば公園	29.0%	16.4%
萌木の村－清泉寮	24.4%	21.2%
八ヶ岳倶楽部－清泉寮	20.5%	3.0%
国立天文台－野辺山駅	18.8%	11.1%
まきば公園－清泉寮	17.9%	25.8%
ひまわり畑－萌木の村	17.6%	0%
国立天文台－滝沢牧場	17.6%	9.1%
萌木の村－八ヶ岳倶楽部	16.9%	6.3%
東沢大橋－八ヶ岳県営牧場	16.7%	1.6%

この結果からは，非常に異なる結果が見られた．全般的にブログから得られた共起性の高い組み合わせはアンケートでは必ずしも高いとは限らない．ブログは全ての行動を記載していない可能性が高く，強く印象づけられたものだけが記載されると仮定すると，この強い共起性が説明可能である．つまり共起性は組み合わせの数が増加すると低下するため，印象に残った魅力度の高い観光施設だけが記載されるならば，共起式の分母が小さくなるため，結果として共起性が高く示されることになる．また，媒介中心性に着目すると，萌木の村が最も高い媒介中心性をもたらし，まきば公園，八ヶ岳アウトレット，国立天文台などが続いている．この地域では清泉寮が中心的な観光施設であるが，ブログデータでは萌木の村がこの地域の核施設であることが明らかになった．

続いてコレスポンデンス分析を用いてリピーターと初来訪者間での訪問地の違いを分析した．今回用いる２次元のコレスポンデンス分析では，総ての単語に着目し，それらが２次元平面上で近隣に存在する単語を示す事で，単語間の類似性が示される．特に訪問地およびリピーター・初来訪のみを取りあげて図に示すことで，初来訪者やリピーターの行動特性が示されることになる．

図には初来訪者，リピーターというフラグを単語として図中に示した．このとき初来訪者の周りには，「吐竜の滝」「国立天文台」などがあるが，それほど近くはない．一方リピーターには「まきば公園」「リゾナーレ」「諏訪湖」「ワンワンパラダイス」などが非常に近接して存在している．つまりこれらの目的地はリピーターとフラグをつけられたブログの中で非常に良く現れていることを示し，リピーターが初来訪者吐比較して高い比率で訪れていることがわかる．その一方「国立天文台」や「吐竜の滝」はリピーターの比率が低いことから，一度訪れた場合には再訪する割合が低いことがわかる．



図－10　ｺﾚｽﾎﾟﾝﾃﾞﾝｽ分析による目的地間の関係

この中でアンケート調査でリピーターが訪れる上位の観光地を示すと，「清泉寮」「道の駅こぶちざわ」「道の駅はくしゅう」「まきば公園」「東沢大橋」「アウトレット」「萌木の村」などである．これらの位置を図－10の中で確認するとリピーターの単語と比較的近いところに位置している．このことからリピーターの目的地に関してはアンケートとブログマイニングは比較的似た結果を出した．一般にブログは一部の目的地について強調して書かれている可能性があるが，リピーターは地域事情を熟知し，重要な目的地しか訪れなくなっていることが予想され，結果としてブログにおいて訪問先の未記入が少ないと考えられる．

５．時系列的なブログを利用した個人の行動特性把握

ブログは一般には同一個人の書いたエントリーが時系列的に連続していることから，個人の意見や経験などについての情報を得ることができる．今回は特にブログに書かれたことから個人の嗜好を推定し，その嗜好別の訪問地等について分析を行う．

注目した情報は各エントリーにつけられたタグもしくはカテゴリーである．各エントリーは何らかのタグがオーナーによって付加されていることが多い．そこでそのタグの分布を元にどういったエントリーを書くことが多い人物であるかを求める．表―４にブログにつけられたタグを集約したものを示す．

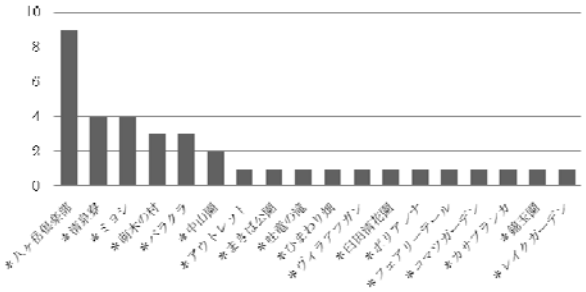
表－４　ブログのタグの集約

ガーデニング	イベント	オフ会
ｲﾝﾃﾘｱ・雑貨	旅行	グルメ
日記	ﾄﾞﾗｲﾌﾞ・ﾂｰﾘﾝｸﾞ	環境
パソコン	買い物	写真
スポーツ	政治・経済	ペット
製品	家族	散歩
映画・TV・本	健康	料理

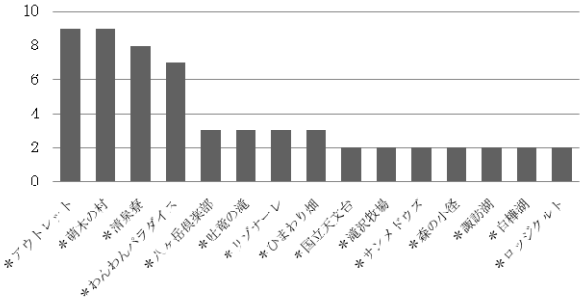
このようにタグを集約して21にまとめ，それぞれのタグがつけられてエントリーがブログ全体に対してどの程度の比率で存在しているかを個人のブログ嗜好指標とした．つまり21種類の0-1の指標が各個人について計算されたことになる．そこから階層的クラスター分析を行い，各個人を７つのグループに分類した．分類された各グループとグループ別の代表的なタグを表－５に示す．得られたクラスターのうち，クラスター２は雑多なクラスターであり，様々な事が書き綴られるグループである．それ以外は比較的明確に特定の話題が多く書かれていた．つまり各クラスターは特定の嗜好を持った内容が書かれていることが明らかであるため，その嗜好と訪問先の関係性を確認するため，クラスター別の訪問地の分布状況の一部を図－11，図－12に示した．

表－５　クラスター分析の結果と代表的タグ

クラスター	カテゴリー
クラスター1（n=4）	スポーツ
クラスター2（n=61）	映画・TV・本，写真，インテリア・雑貨，家族，料理，グルメ，パソコン，家電
クラスター3（n=8）	旅行
クラスター4（n=8）	ドライブ・ツーリング
クラスター5（n=12）	ガーデニング
クラスター6（n=22）	ペット，ペット+日記
クラスター7（n=45）	日記，TV・本，



図－11　クラスター5の訪問地上位



図－12　クラスター６の訪問地上位

図－11はガーデニングに関するエントリーを主体とするブログの訪問先分布を示しているが，先に示した全体の訪問先と比較して「八ヶ岳倶楽部」「ミヨシ」「バラクラ」が上位進出している．これらはいずれも植物を主体としており，ガーデニングに関連した訪問地といえる．また，図－12はペットを主体として書かれているブログの訪問地上位であるが，「八ヶ岳リゾートアウトレット」「萌木の村」「清泉寮」「わんわんパラダイス」などペット同伴入場が可能な施設が上位に集中しており，ペットを趣味とした個人の嗜好と行動の明確なリンクが見られる．また，図には示さなかったが，クラスター7は日記を主体とした日常生活を綴ったグループであり，その訪問地上位は「清泉寮」「八ヶ岳リゾートアウトレ

ット」などブログ全体の訪問地上位と変わりなく，平均的な行動を示していることも明らかになった．

６．おわりに

本研究はインターネット上で広がりを見せていたブログにおける行動記録を抽出し，ランダムサンプリングによる代表的行動の把握とは異なる視点での行動抽出の検討を行ってきた．特に観光行動では代表的観光者という視点よりも，ある特定のグループが特有の行動をしている等の発見しそこから魅力的な観光地づくりへのヒントを得ることを目的とした．

ブログには明確な個人の属性等は示されないことが多いが，多くの属性については推測が可能であり，入込み調査との特性比較からは，女性が多い，より印象深いと考えられる行動の記述が多いなどの予想された特性が示された．また，テキストマイニングの様々な方法論を適用して分析を行った結果，観光施設の結びつきの強さでは，入込み調査では明確でないが，ブログにおいてのみ強く表れる関係性などが示された．ブログは印象等が強かった施設の記入が多くなると考えられることから，総ての訪問先を同等に扱った入込み調査による関係性分析とは異なり，印象等の要素が加えられた結果であると言える．また，コレスポンデンスぶんせきからは，初来訪やリピーターとつながりのある訪問先を明らかにすることも可能であり，リピーターをうまく引き寄せている施設群やリピーターが訪れにくい施設群なども明らかになった．またブログの特徴である個人のエントリー記述の時系列データを用いたクラスタリングからは，ブログに書かれている内容と観光地における周遊行動には強い関係があることも示された．

以上，本研究はブログによる行動データ収集可能性に着目し，テキスト分析の様々な手法を適用して，その可能性を検証した．しかし，今回は適用したレベルにとどまっているため，今後は必要なデータはどのように抽出されるのか，という視点からの検討が必要である．

【補注】

(1) スパムブログとはアフィリエイトの表示や特定のサイトへ誘導することなどを目的として開設されたブログであり，内容は意味をなさないことが多い．

(2) ブログ収集当時はYahoo!Japanは独自の検索エンジンを持っていたが，現在はGoogleエンジンを利用している．

参考文献

1) 奥村学：blogマイニング―インターネット上のトレンド，意見分析を目指して―，人口知能学会誌21巻4号，2006
2) 鈴木大輝：周遊圏域の広がりと地域連関に着目した観光客の行動分析，山梨大学卒業論文，2010
3) 樋口耕一：KHCoder，http://khc.sourceforge.net/，2010年7月
4) 奈良先端科学技術大学院大学情報科学研究科自然言語処理講座：茶筌，http://chasen-legacy.sourceforge.jp/，2010年7月
5) 金明哲：統計的テキスト解析(6)，ESTRELA，No.173，pp.58-63，2008

A STUDY ON FEASIBILITY OF BLOG-MINING FOR TRAVEL BEHAVIOR INFORMATION THROUGH COMPARISON WITH QUESTIONNAIRE SURVEY

Kuniaki SASAKI, Maika KITO and Keita YAMAZAKI

This research tries to validate practical use of the blog on the Internet as a new data sauce of travel behavior. That is, we show the possibility of blog for valuable information mining. To validate the characteristic of blogs, we compare them with the questionnaire survey conducted in the same area. As the empirical study, we collected blogs on the tour to Mount Yatsugatake area. Blog is different from ordinary survey and one important difference is there is almost no information about individual. Thus we tried to estimate individual attribute and some behavioral trait from all the entries of blog. As a result, some featuring characteristic which is not seen in questionnaire survey was extracted from blogs. This means that blogs can be useful data source for attractive tourist resort formation.