マイクロブログを活用した観光地における情報抽出方法に関する研究—埼玉県川越市を対象として—

法政大学　学生会員　○外村　剛久
法政大学　正会員　宮下　清栄

1.研究背景・目的

平成 20 年に観光庁が設立され，多くの地方自治体が観光に関する施策を実施することで観光産業の活性化を図っている．更に新規の観光統計調査も開始され，今後は多様な情報を適切に用いたマーケティングを行い，戦略的な観光施策を講じる必要がある．しかしマーケティングの際，アンケート調査等の観光情報を高頻度かつ広域に取得する事は多大なコストや時間がかかるため，現実的には困難である．一方，平成 12 年以降に SNS（Social Networking Service）と呼ばれるサービスが普及した．サービスの機能であるマイクロブログは，様々な情報がリアルタイムで更新され続けており，東日本大震災以降ビックデータとして注目されている．中には観光客や観光関連従業者の情報も含まれており，マーケティングの基礎データとして有効である．そこで本研究はマイクロブログを収集し，観光地における情報を抽出する方法を提案するとともに，本研究での情報抽出方法の特徴や限界及び課題を明らかにする事を目的とする．

2.対象地域

対象地域は埼玉県川越市とする．選定理由として，川越市は平成 19 年に年間 600 万人以上の入込客数を有する観光地でありマイクロブログのサンプル数が多いと考えられたこと，観光地域内に歴史・文化資源や飲食・小売店舗が集積していたことが挙げられる．

3.情報抽出方法

3.1 施設データベース作成

図 1 に示す抽出規準に従い，観光ガイド「時薫まち川越」の小江戸川越マップに記載の施設（寺社や建造物群）を抽出した．抽出の結果，74 件であり，川越市観光課ホームページに記載された観光スポット 13 件と比べると多くの施設を抽出できた．抽出後に住所，緯度・経度を入力，溝尾らの研究[1)]を参考に抽出した施設の種類の分類を行い，施設データベースを作成した．

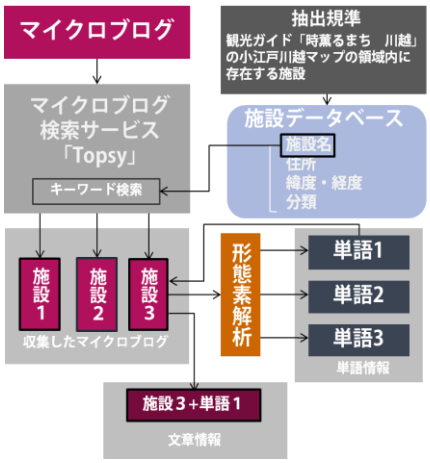

図 1．観光地における情報抽出フロー

表 1．収集方法及び条件

収集方法	(A)同一の施設名なしの場合	A-1:施設データベースに含まれる「施設名」をキーワードとし、検索することで得られたマイクロブログ。 A-2:投稿期間を満たしているマイクロブログ
	(B)同一の施設名ありの場合	B-1:施設データベースに含まれる「施設名」と「川越」をキーワードとし、検索することで得られたマイクロブログ。 B-2:A-1と同様
	(A)・(B)の判別方法	Google Mapを用いて施設名を検索、2件以上の地点が検索された場合は(B)、1件のみで対象地域内に位置する場合(A)
収集条件	収集情報数の条件及び上限	一回の検索で100件が上限、1ヵ月・1日単位で検索しても同様に100件が上限である。1ヵ月の場合は施設1件あたり年間最大で1,200件のマイクロブログが収集できる。
	収集対象の投稿期間及び収集単位	2012年11月-2013年10月、1ヵ月単位で検索
	Twitter(マイクロブログ)検索サービス	Topsy(無料のTwitter検索サービス)

3.2 マイクロブログ収集方法及び収集方法の特徴

本研究でのマイクロブログ収集方法及び条件を表 1 に示す．本研究では無料の Twitter 検索サービス「Topsy」を用いる．Topsy は特定の日，月に投稿されたマイクロブログを収集する事が可能であり，本研究では 2012 年 11 月から 2013 年 10 月まで 1 ヵ月単位で検索した．また，一回の検索で 100 件の収集上限であり，1 ヵ月単位で検索した場合は施設 1 件あたり年間最大で 1，200 件が収集できる．収集方法は(A)と(B)の 2 通りであり，ともに施設データベース中の各施設名を検索キーワードとするが，(B)では各施設名に加えて「川越」を検索キーワードとし，収集を行っている．これは施設名が全国に 2 件以上ある場合，本研究の対象地域外に存在する施設に関するマイクロブログも収集してしまう可能性があるためである．また，(A)・(B)の判別方法は Google Map を用いて各施設に検索を行った．

キーワード　ビッグデータ、マイクロブログ、観光情報、形態素解析、キーワード検索
連絡先:〒162-0843　東京都新宿区市谷田町 2-33　法政大学デザイン工学部都市環境デザイン工学科　miyasita@hosei.ac.jp

従来はプログラミングによる収集や企業が提供する有償の提供サービスでの収集が主流であった．本研究では無料の検索サービスを活用し，施設データベースに基づいたマイクロブログの情報抽出を低コストかつ簡易に行える点で特徴がある．

3.3 形態素解析による単語の抽出

マイクロブログの投稿内容は自由記述欄のテキストデータであるため，観光地の情報特徴を把握しにくいという問題がある．そこで，収集したマイクロブログの投稿内容に対して形態素解析を行い，単語を抽出・集計する事により施設の特徴把握を試みた．しかし，単語から得られる情報には限界があるため，単語を抽出した際に出現回数が多かった 10 個の単語を選定し，収集したマイクロブログの中から選定した単語を含む文章情報も抽出する．

4.結果

4.1 マイクロブログ収集結果

施設別のマイクロブログ収集件数を図 2 に示す．マイクロブログを収集できたのは 74 件の施設の内 61 件であり，13 件の施設についてはマイクロブログを収集できなかった．また，全体では 11853 件のマイクロブログを収集した．1000 件以上抽出できた施設は菓子屋横丁（1200 件)，喜多院（1200 件)，時の鐘（1200 件)，川越スカラ座（1198 件)，川越氷川神社（1196 件)，小江戸蔵里（1157 件）であった．

4.2 形態素解析結果

表 2 は 1000 件以上のマイクロブログを収集できた 6 件の施設を対象に形態素解析により単語を抽出し，施設ごとに抽出できた単語の頻度を降順に 10 件まで示した．しかし，抽出した情報の中には施設名と抽出した単語とのつながりが明確でないものも含まれていることが明らかとなった．表 3 に一例を示す．菓子屋横丁に関するマイクロブログの中から「雰囲気」が含まれる文章情報を抽出したが，文脈から判断して菓子屋横丁の雰囲気について触れていると言えない．

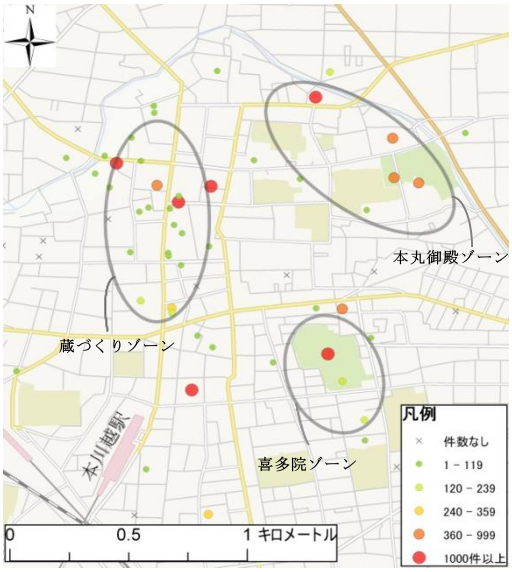

図 2. マイクロブログ収集件数

表 2. 形態素解析結果

菓子屋横丁		喜多院		時の鐘	
施設分類：	歴史的景観	施設分類：	寺社	施設分類：	史跡
単語	頻度	単語	頻度	単語	頻度
茶屋	116	五百羅漢	168	シンボル	48
駄菓子	106	七福神	101	電車	48
味わい	60	節分	84	駅名	45
岩塩	52	名所	84	耐震	45
周辺	41	スポット	83	街並み	40
雰囲気	39	バス	82	距離	37
団子	36	名刹	81	自動	34
街並み	35	境内	78	一興	32
ギャラリー	33	大師	75	七福神	32
パン	33	大黒天	65	大火	32
川越スカラ座		川越氷川神社		小江戸蔵里	
施設分類：	建造物	施設分類：	寺社	施設分類：	建造物
単語	頻度	単語	頻度	単語	頻度
映画館	144	縁結び	240	酒造	248
デジタル	122	結婚式	124	トンカツ	220
ミュージカル	119	一般	105	土産	173
映写機	99	祭礼	105	産業	151
ラジオ	81	祭事	103	イベント	143
プロジェクト	77	呼び名	102	モール	139
男性	59	文化財	102	パン	135
シネマ	54	民俗	102	日本酒	131
ガレキ	52	祭り	84	コーナー	129
イベント	48	大祭	83	商店	129

表 3. 抽出した文章情報の内容

単語	抽出した情報（菓子屋横丁）
雰囲気	あと、友達とよく行く川越の菓子屋横丁外れたあたりにある喫茶店が雰囲気良くてコーヒーも美味しくて好き。

5.結論

本研究で得られた結論を以下に示す。①：施設データベース作成において，川越市観光課ホームページに記載されている観光スポット 13 件の施設に比べ，抽出規準を設けることで 74 件の施設を抽出し，データベースを作成する事ができた．②：無料の Twitter(マイクロブログ)検索サービスを用いて施設に関するマイクロブログの収集する際に，同一の施設名がある場合とない場合の 2 通りの収集方法を設けることで，同一の施設名を持つ施設に関するマイクロブログを含むことなく，74 件中 61 件の観光利用されている施設でマイクロブログを収集する事ができ，11853 件のマイクロブログを収集できた．③：収集したマイクロブログの投稿内容を対象に形態素解析を行い，単語を抽出することで施設の特徴を把握することができた．更に抽出された単語の中から任意で単語を選び，収集したマイクロブログの中からその単語を含む文章情報も抽出することもできた．今後はマイクロブログの発信者や内容の詳細分析等，今後も研究を継続する必要がある．

<参考文献> 1)溝尾良孝：観光資源論-観光対象と資源分類に関する研究，城西国際大学紀要 16(6)，2008-03