LBSN を用いた都市利用者の特徴単語と都市の機能との関係抽出

○菊地　弘祐[*1]　木原　己人[*2]　遠田　敦[*3]　高柳　英明[*4]
木村　謙[*5]　林田　和人[*6]　渡辺　仁史[*7]

キーワード：　Location Based Social Media, Hadoop, Mahout, , TFIDF, クラスタリング , 都市利用動態

1. 研究背景

1.1 現在までの都市分析手法

今までに様々な都市分析手法が提案されてきた。これらを大きく分類すると、エスノグラフィーと数学的モデルに基づくものであろう。これらは一定の成果を上げたものの、様々な問題を抱えている。例えば、前者では追跡調査やインタビューといった方法では、その対象の特性はわかるが、サンプル数を増やすことが非常に困難である。逆に後者では、数学的モデルを用いる際には、研究者らは特定のパラメータに着目しているために、他の定性的なパラメータが削除されてしまう危険性がある。

これに対して、Web による情報爆発が起きるにつれ、データ量が多ければ質を上回るという主張がなされるようになってきた。Halevy ら [1] は、データ量が多くなれば全体の量に対するノイズの割合が小さくなるために、少量のデータにノイズ処理とモデル化という複雑な処理を施すより、大量のデータをシンプルにモデル化する方が良いと主張した。また、矢野ら [2] は一見役に立たないかのように見える情報も、臨界量を超えると価値のあるものになるとの見解を示した。

もちろん、情報量が増えるのは一概に良いことではなく、その分データ処理にかかる時間が増え、データを扱いにくくなることが欠点である。実際に、情報処理や統計処理にかかる時間が長くなれば、データマイニングが前提とするヒューリスティックな分析が難しい。データ量の違いによる特性を表 1 に示す。

1.2 本研究におけるアプローチ

筆者らは先述の主張を元に、ソーシャルメディア上にあるデータから知見を得るための方法論を研究し、都市への知見を抽出してきた。そのデータソースで注目したのは Twitter[3] と foursquare[4] である。まず、Twitter は API を通じて容易にユーザー名、投稿時間、投稿内容、場所情報が取得できる。140 文字に限定されているため、利用者は気楽に投稿することができ、ここから利用者の特性が抽出可能である。また、foursquare には場所の機能とその利用回数が抽出可能であると思われる。foursquare は Venue と呼ばれる特定の場所でチェックインすることで利用者が得点を稼ぎ、他のユーザーと競い合うというソーシャルメディアである。これらの情報は API を通じて容易にアクセス可能である。

表 1 データ量ごとの利点と欠点

少ないデータ量	多いデータ量
ノイズ処理を含めた複雑なモデル	単純なモデル
データを扱いやすい	データを扱いにくい
処理速度が比較的早い	処理速度が非常に遅い
ヒューリスティックな手法が使える	ヒューリスティックな手法が使いにくい

また技術的な問題として、データ量が多くなるにつれて処理速度の遅さとアプリケーション上での統計分析が難しくなってきた。これらの解決のため、並列分散処理のプラットフォームとして Hadoop[5]、機械学習ライブラリとして Mahout[6] を用いて、統計処理には R[7] を用いた。これらにより、情報処理速度を高めてヒューリスティックな手法の実現を試みた。

2. 研究目的

Twitter と Foursquare から得られたデータから利用者の特性と街の店舗利用の特性を明らかにすることを目的とする。

3. 研究方法

3.1 データ収集方法（図 2）

本研究では Twitter と foursquare から公開されているデータを取得した。Twitter では、日本近辺でジオタグを付けているユーザーに対し、Streaming API にてリストアップした。その個々のユーザーに対して、Rest API にて可能な限り履歴を取得し、ユーザーごとに、TweetID、ユーザー ID、ユーザー名、投稿日時、テキスト、緯度、経度

を CSV 形式でユーザーごとに書き出した。Twitter から得られたデータは、テキストでユーザー群の特性を抽出し、緯度経度で行った場所を追跡するために用いることが狙いである。foursquare では、日本のすべての駅[8]に対して半径 400 m以内にある Venue の情報を API で取得し、Venue 名、緯度、経度、チェックイン数、ユーザー数、カテゴリ名を駅ごとにファイルに書き出した。ここで得られる駅圏内における foursquare のチェックイン数は、Twitter のどのクラスタが関連しているのか導くために用いる。

3.2 テキストデータ前処理

Twitter 利用者をクラスタリングするために lucene-gosen[9]にてテキストデータに対して形態素解析を行い、名詞のみを抽出した。この形態素解析ライブラリは Java で書かれていて、ユーザー辞書の構築が比較的容易に行えることが特徴である。そこで、テキストデータの前処理をするにあたり、”Wikipedia”、” はてな ” の主要単語を加えたユーザー辞書を作成した[10) 11)]。このユーザー辞書を用いて、形態素解析時に名詞のみを抽出し、それらを半角スペースで区切り、ユーザーごとにテキスト形式で出力した。

3.3 Hadoop と Mahout によるテキスト処理

3.2 で得られたすべてのファイルを Hadoop に読み込ませ、HDD のシークエンスアクセスに適した Sequence 形式のフォーマットに変換した。次に、Mahout のパーサーから登録された文字ベクトルデータから単語リスト、すべての文章データから得られる絶対頻度（Term Frequency）、ユーザーごとに現れる回数（文章頻度、Document Frequency）、絶対頻度を文章頻度の対数でかけたノイズ処理後の相対頻度（TFIDF, Term Frequency Inverse Document Frequency）を算出した。このときの単語リストには 92164 個の単語が記された。

3.4 ユーザーのクラスタリング

3.3 の結果のうち、92164 の単語とその TFIDF の結果を用いてクラスタリングを行った。まず、クラスタリング時の初期値を定めるために、キャノピークラスタリングを用いた。これは二つの閾値から、特性の近いものをまとめ、比較的高速にクラスタリングを行うことができる手法である。この際に、テキストマイニングに適したコサイン類似度を距離計算として用いた。また、解が収束せずにエラーが発生しない範囲で閾値が最小になるように閾値を設定した。その結果、閾値は 0.87 と 0.91 を設定し、1058 個のクラスタが得られた。

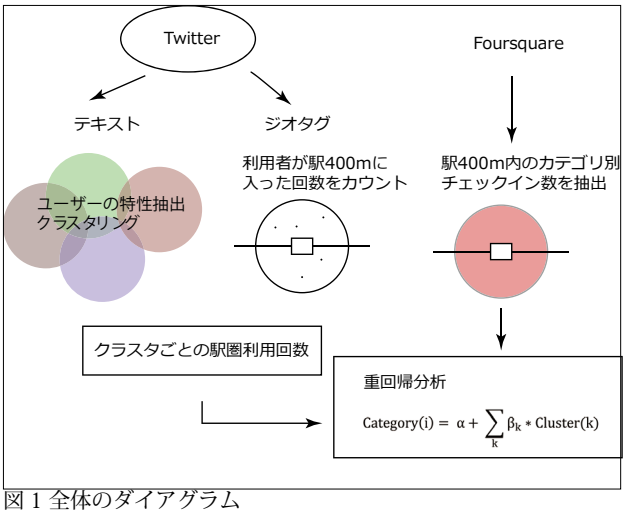

図 1 全体のダイアグラム

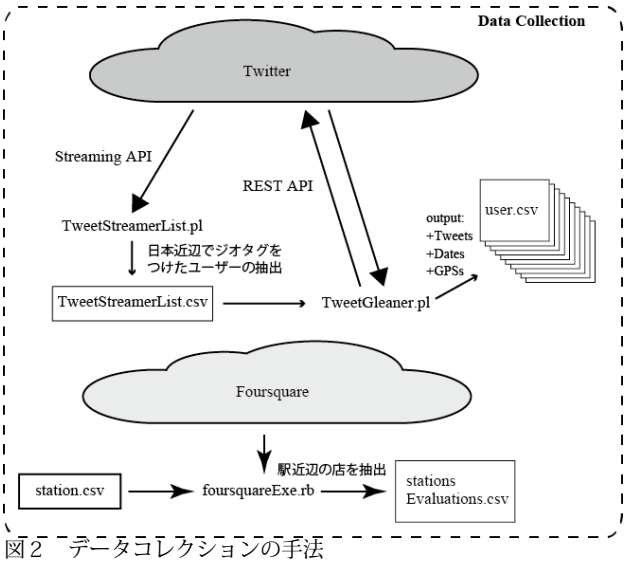

図 2　データコレクションの手法

次に、KMeans 法をキャノピークラスタリングの結果を初期値に設定し、クラスタリングを行った。この結果を Mahout の clusterdump コマンドを用いてテキスト形式と CSV 形式を出力した。テキスト形式のファイルにはクラスタごとの単語ベクトルが記されており、また CSV 形式のファイルには、クラスタに属するユーザー名が記されている。

3.5 クラスタごとの駅圏利用回数と駅圏のチェックイン数

前項にて得られた CSV 形式のファイルからクラスタごとに所属するユーザーを抜き出した。その後、その緯度経度を抽出した。最後に、駅の 400 m範囲内に入る回数をファイルに書き出した。また、3.1 にて得られた Foursquare のデータをカテゴリ別に分類し、チェックイン数を駅圏ごとに集計した。本研究で得られたカテゴリ

表２　Foursquare のカテゴリ別チェックイン数上位 30 件

	チェックイン数	カテゴリ
1	3327408	Mall
2	1622263	Café
3	1612316	BookStore
4	800664	Park
5	512760	Bar
6	461473	Restaurant
7	442883	Plaza
8	337163	Bakery
9	304248	Multiplex
10	230543	Diner
11	216575	Pub
12	215189	Neighborhood
13	123540	Playground
14	121364	Nightclub
15	120154	Steakhouse

	チェックイン数	カテゴリ
16	110882	Theater
17	94722	Boutique
18	74225	Stadium
19	64852	Museum
20	63239	Garden
21	32955	Food
22	32662	River
23	28773	Lake
24	28646	Brewery
25	23634	Casino
26	23479	Beach
27	20155	Lounge
28	18865	Butcher
29	18596	Castle
30	18575	Pool

数は 265 件であった。

4. 分析結果

4.1 カテゴリの集計

foursquare のデータ全体で得られたカテゴリとそのチェックイン数を計算した。そのうち上位 30 件であるカテゴリを表２にて示す。この結果、Mall が一番多く、そのチェックイン数は 3,324,708 件にも達した。次に Café と Bookstore が位置し、これらは Mall のチェックイン数の約半分であった。

4.2 カテゴリ別の重回帰分析

前項で得られた上位カテゴリのうち 30 件を目的変数にし、得られたクラスタ 1058 個を説明変数にし、重回帰分析を行った。この結果から p 値が 0.05 以下水準で優位な説明変数のみを残し、再び重回帰分析を行った。この作業を 0.05 以上の説明変数がなくなるまで行った。ここで得られた決定係数を表３に示す。全体的にあてはまりが良くなく、この中でも比較的あてはまりがよかったものは Bookstore でその自由度調整済み決定係数は 0.395 であった。また決定係数が 0.2 以上のものは 4 つに限られた。この決定係数が全体的に低い点に関しては 4.4 にて言及する。

4.3 クラスタごとの回帰式の係数

重回帰分析の結果、比較的決定係数が高いものについて、表 4 に Bookstore において説明変数にしたクラスタの回帰式の係数を示す。この係数はあるクラスタがチェックイン数に対して影響を及す。例えば、Bookstore では 496 番目のクラスタが一番大きくチェックイン数に正の影響がある。逆に 587 番目のクラスタはチェックイン数に負の影響がある。従って、これらをクラスタのジオタグを用いて、あるのクラスタが利用しやすい場所を特定することで、潜在的な需要を導きだすことが可能である。

表３　上位カテゴリ 30 位の重回帰分析後の p 値と決定係数

チェックイン数	カテゴリ	p値	決定係数
1612316	BookStore	2.20E-16	0.395
216575	Pub	2.20E-16	0.2539
28646	Brewery	2.20E-16	0.2131
1622263	Café	2.20E-16	0.201
461473	Restaurant	2.20E-16	0.193
18865	Butcher	2.20E-16	0.1763
121364	Nightclub	2.20E-16	0.1455
215189	Neighborhood	2.20E-16	0.1148
337163	Bakery	2.20E-16	0.09835
3327408	Mall	2.20E-16	0.08086
74225	Stadium	2.20E-16	0.08056
230543	Diner	2.20E-16	0.07239
120154	Steakhouse	2.20E-16	0.06525
23634	Casino	2.20E-16	0.0524
18596	Castle	3.72E-10	0.03599
110882	Theater	4.85E-06	0.02534
442883	Plaza	0.763	-0.003934
32955	Food	0.9487	-0.008783
800664	Park	0.9998	-0.01849
18575	Pool	1.00E+00	-0.02059
304248	Multiplex	1.00E+00	-0.0208
123540	Playground	1.00E+00	-0.0441
23479	Beach	1.00E+00	-0.0473
63239	Garden	1.00E+00	-0.07215
64852	Museum	1.00E+00	-0.07891
32662	River	1.00E+00	-0.08407
94722	Boutique	1.00E+00	-0.08945
512760	Bar	1.00E+00	-0.09325
28773	Lake	1.00E+00	-0.1034
20155	Lounge	1.00E+00	-0.1169

表４　Bookstore における各クラスタの回帰式の係数

クラスタ番号	ユーザー数	BookStore
496	7	976.753818
1055	2	573.256741
359	8	351.834879
592	2	345.311751
138	4	305.631628
754	1	299.461169
210	3	86.822216
55	1	39.48611
47	9	28.615237
481	15	19.832384
625	37	9.034952
423	23	8.397439
640	10	1.242562
430	11	-30.409635
882	3	-143.815675
817	5	-186.789789
349	8	-298.896758
587	1	-1027.712701

4.4 決定係数の低さ

Brewery のカテゴリに対して分位数―分位数プロットを行った結果を図３にて示す。分位数―分位数プロットは分析時に扱ったデータが正規分布に従うかどうか確認する図である。重回帰分析は正規分布を前提としているために、データが正規分布に従うかどうかを確認した。図３の分位数―分位数プロットは Y 軸が残差で X 軸が残差

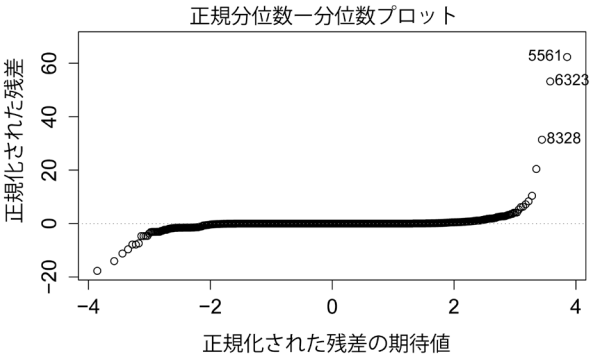

図3　Brewery での分位数―分位数プロット

表5　回帰分析の残差の 5 数要約

	最小値	下側ヒンジ	中央値	上側ヒンジ	最大値
Bookstore	−2848	−156	−156	−156	35626
Pub	−1440.9	−23.3	−23.3	−22.1	5887.7
Brewery	−1394.6	−1.1	−1.1	−1.1	4475.7
Café	−6133.2	−162.2	−160.3	−127.2	26469.7

の期待値である。仮に、ここで用いた foursquare のチェックイン数や駅圏に入った回数が正規分布に近ければ、分位数―分位数 プロットでは残差と残差の期待値が y=x の直線を描く。しかしながら、本研究では残差の期待値が大きくなるにつれ、想定される残差より大きくなる傾向になった。そのため、これらのデータは正規分布ではない可能性が高い。この傾向は残差の 5 数要約にも表れている。

これに対して、分位数―分位数 plot の結果や表 3 から、foursquare のチェックイン数はパレート分布に従う可能性がある。そのために、これらの数字を正規化する必要があったと思われる。

5. まとめ

本研究では、Twitter にて利用者をクラスタに分類し、駅 400m に入った回数を計算した。foursquare のチェックイン数のデータを目的変数にし、利用者のすべてのクラスタを説明変数にし、重回帰分析を行った。その結果を箇条書きにて示す。

- 重回帰分析から得られる説明変数の係数から、どのクラスタが foursquare のチェックイン数に寄与しているのかがわかった。
- あるカテゴリにおいてチェックイン数に寄与しているクラスタを追跡することで、チェックイン数が低くともポテンシャルがある場所を抜き出すことが可能であろう。
- 重回帰分析後の説明変数の調整済み決定係数が小さい。説明変数自体の分布が重回帰分析の前提となる正規分布に従っていないためである。
- 順位と回数は経験的にパレート分布に従うとされていて、foursquare のチェックイン数もその傾向がうかがえるため、これらを正規化する必要がある。

謝辞

本研究は、科学研究費補助金挑戦的研究（11538100）「ソーシャルネットワーキングサービスに投稿された記事に基づく都市・建築空間像の解析」の助成を受けたものである。

参考文献

1) Haleby A., Norvig P. and Pereira F: The Unreasonable Effectiveness of Data, IEEE Intelligent Systems, Vol. 24(2), 2009

2) 矢野和男 , 栗山裕之 :「人間 X センサ」 センサ情報が変える人・組織・社会 , 日立評論 , Vol. 89(7), 2007

3) Twitter Developers: Twitter 参照日 2013.10.11: https://dev.twitter.com

4) foursquare for Developer: foursquare 参照日 2013/10/11: https://developer.foursquare.com

5) Welcome to Hadoop: The Apache Software Foundation 参照日 2013/10/11: http://hadoop.apache.org

6) Apache Mahout: Scalable machine learning and datamining: The Apache Software Foundation 参照日 2013/10/11: http://mahout.apache.org

7) The R Project for Statistical Computing 参照日 2013/10/11: http://www.r-project.org

8) 駅データ : CodePlus 参照日 2013/10/11: https://www.ekidata.jp

9) Lucene-gosen – Japanese analysis for Apache Lucene/Solr 3.6 and 4.5 参照日 2013/10/11: http://code.google.com/p/lucene-gosen/

10) Java 製形態素解析ライブラリ「lucene-gosen」を試してみる : mwSoft 参照日 2013/10/11: http://www.mwsoft.jp/programming/munou/lucene_gosen.html

11) mecab のユーザ辞書で wikipedia と hatena キーワードを利用する 参照日 2013/10/11: http://tmp.blogdns.org/archives/2009/12/mecabwikipediah.html

＊1　早稲田大学総合研究所　客員研究員

＊2　早稲田大学大学院建築学専攻　博士課程

＊3　東京理科大学建築学科　助教　博士（建築学）

＊4　滋賀県立大学環境科学部　准教授　博士（工学）

＊5　A&A Co. Inc.　研究員　博士（建築学）

＊6　早稲田大学総合研究所　客員准教授　博士（工学）

＊7　早稲田大学建築学科　教授　博士（工学）