日本ソーシャルデータサイエンス学会
2024 年春季研究発表会

Twitter における市町村関心度の数値化手法

武内奎太[†1]　新庄雅斗[†2]　關戸啓人[†2]　岩崎雅史[†1]

キーワード：ソーシャル・ネットワーキング・サービス (SNS)，Twitter，市町村，関心度，Google Cloud Platform（GCP），京都山城地区

1. はじめに

近年，情報技術の急速な発展により，インターネット通信の高速化やモバイル端末の普及が進み，SNS が世の中に広く浸透した．このため，日本の多くの企業は SNS の活用に非常に積極的であり，最近では地方自治体も例外ではない．

SNS の活用には情報の発信者として取り組むものと情報の受信者として取り組むものがあるが，本研究では後者の 1 つとして SNS 上でどのくらい注目を集めているかを市町村ごとに算出する方法を提案する．具体的には，市町村ごとに関連投稿数をカウントし，それを人口で除すことで得られる市町村関心度と称する数値を算出するための手順を示す．これにより，市町村関心度の比較を通じてどの市町村が SNS 上で注目を集めているかが客観的に明らかにできる．

本研究では，対象の SNS として Twitter を選択し，実際に京都府南部の山城地区の 12 市町村を対象とした分析結果を報告する．なお，山城地区を構成する 12 市町村は宇治市，城陽市，八幡市，京田辺市，木津川市，久御山町，井手町，宇治田原町，笠置町，和束町，精華町，南山城村である．

2. ツイートデータの収集

市町村関心度を求める際に Twitter から京都山城地区の 12 市町村に係わると思われるツイートデータを収集した．利用者が非常に多い Twitter を分析対象とする際にはツイートの数は一般的に膨大になるため，一般的なデータ処理ツールで扱うことは現実的ではない．そこで，ツイートデータを収集するために Google が提供するクラウドコンピューティングサービス GCP を用いた. ツイートの収集期間は 2022 年 6 月 1 日から 2022 年 9 月 8 日の 100 日間として，ツイートの本文中に「市，町，村を除いた市町村名」，あるいは「先頭に#が付いた市町村名」の少なくともどちらか 1 つが含まれるものを係わると思われるツイートと判定し，GCP を用いて市町村ごとにツイートを収集した．

3. ツイートの分別

GCP サービスを用いて収集した市町村名を含むツイートの中には係わりのないツイートが含まれる．よって，正確な市町村関心度を数値化するためには，それらのツイートを収集ツイートから取り除く必要がある．収集されたツイートの数は市町村ごとに大きく異なるため，ツイートの分別方法はツイート量に応じて 2 つの方法を考えた. 以降，簡単のために市町村に係わりのあるツイートを関連ツイート，係わりのないツイートを無関連ツイートと呼ぶことにする．

3.1 収集されたツイートが 10 万件未満の場合

収集されたツイートが 10 万件未満の市町村は宇治市と八幡市を除く 10 市町村である．これらについてはツイートの内容を 1 つ 1 つ目視で確認して関連ツイートと無関連ツイートに分別し，無関連ツイートを取り除いたものを関連ツイートとして保存する．

3.2 収集されたツイートが 10 万件以上の場合

収集されたツイート数が 10 万件以上となるのは宇治市と八幡市の 2 市である．ツイートのテキストデータに含まれるワードをもとにツイート内容を推測して関連ツイートと無関連ツイートに分別した．

まず，このワードが含まれると関連ツイートと判定する「関連ワード」と，このワードが含まれると無関連ツイートと判定する「無関連ワード」を決定する．ワードの選定には，収集されたツイートから 4000 ツイートをランダムに抽出し，それらを 1 つ 1 つ目視で関連ツイートと無関連ツイートに分別したテスト用ツイートを用いる．市町村名が含まれるワードに着目する方法と自然言語処理オープンソースライブラリ GiNZA[1]を用いてテキストデータから固有名詞や時間表現などの固有表現を抽出する方法によってワードの候補を絞り，その候補から関連ワードと無関連ワードを選定する．

選定したワードを用いてツイートを分別することになるが，その流れについては図 1 に示すとおりである．最初に実行すべきは関連ワードと無関連ワードが本文中に含まれているか否かの確認である．関連ワードのみが含まれるツイートを関連ツイートに，無関連ワードのみが本文に含

†1 京都府立大学
†2 大阪成蹊大学

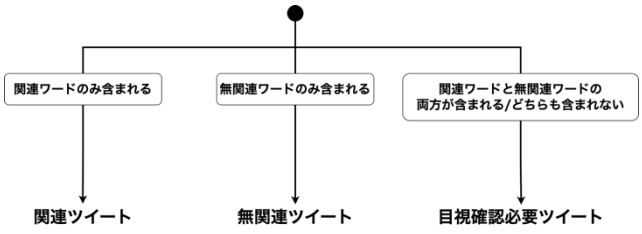

図 1　ワードによる分別の流れ

まれるツイートを無関連ツイートに分別する．関連ワードと無関連ワードの両方が含まれるツイートと関連ワードと無関連ワードのどちらも含まれないツイートについてはワードによる分別が困難なため，目視による確認が必要なツイートとする．目視による確認が必要なツイートは，１つ１つツイートデータを目視で確認して，関連ツイートと無関連ツイートに分別する．そして，全てのツイートを分別した後に，無関連ツイートを取り除いたものを関連ツイートとして保存する．

4. 市町村関心度

関連ツイートに分別された件数から山城地区 12 市町村に対する市町村関心度を算出した．Twitter 上に市町村名の登場回数が多いほど，その市町村に対する関心が高いと考えるのは妥当である．しかしながら，市町村名の単純な登場回数である関連ツイート数では市町村の規模が全く考慮されていないため, 市町村が Twitter 上での広報戦略を考える際の直接的な指標にはなり得ない．そこで，１つの関連ツイートが市町村の内外どちらから投稿されたものかは問わず，市町村の 1 人の住民によるものと見なし，関連ツイート数を人口で除すことで得られる住民 1 人当たりの平均投稿数を市町村関心度とした．市町村の人口については Wikipedia に掲載の数値を用いた[2].

表 1 は，山城地区の 12 市町村に関して，GCP で収集されたツイート数，関連ツイート数，人口，市町村関心度をまとめたものである．GCP で集められたツイート数は宇治市が最も多く 431,448 ツイートであり，逆に宇治田原町が最も少なく 4,148 ツイートであった．関連ツイート数も宇治市が最も多く 366,461 ツイートであり，逆に井手町が最も少なく 1,909 ツイートであった．関連ツイート数と人口をもとに算出した市町村関心度は笠置町の 6.467 が最も高く，逆に最も低い木津川市は 0.133 であった．笠置町の関連ツイートは笠置キャンプ場絡みのもの，南山城村も関連

表 1　数値化した結果

対象 市町村	ツイート 数	関連 ツイート 数	人口	市町村 関心度
精華町	42,425	11,613	35,911	0.323
宇治市	431,448	366,461	177,229	2.068
城陽市	23,367	19,235	73,625	0.261
八幡市	309,065	19,945	69,680	0.286
京田辺市	22,676	21,833	74,331	0.293
木津川市	17,373	10,493	79,010	0.133
久御山町	13,325	12,932	15,014	0.861
井手町	34,973	1,909	7,183	0.266
宇治田原町	4,148	4,134	8,664	0.477
笠置町	19,567	6,823	1,055	6.467
和束町	4,409	4,050	3,313	1.222
南山城村	6,363	6,030	2,311	2.609

ツイートは道の駅みなみやましろ村絡みのものが支配的であり，宇治市の関連ツイートについては宇治茶や歴史的建造物絡みのものが多く，市町村関心度の高い 3 市町村に共通するのは観光スポットや特産品が強みという点である．対して，木津川市は観光スポットや特産品の少ないベッドタウンであるため，市町村関心度が低いと考えられる．

5. まとめ

本研究では GCP サービスを用いて Twitter から山城地区 12 市町村に関連すると考えられるツイートの候補をすべて収集し，その中から本当に関連するものだけに絞り市町村関心度を算出した．

観光スポットや特産品が市町村関心度に影響を与えるのは明らかとなったが，それら以外で市町村関心度を左右する要因が本研究において解明できたわけではない．関連ツイートの内容まで分析することで市町村関心度を左右する要因が把握できると思われるが，この点については今後の課題の 1 つである．また，本研究では対象としていない人口 50 万人以上の大都市に対しても提案手法が有効に機能し，本研究と同じように市町村関心度が算出できるかなども検証したい．

参考文献

[1] “GiNZA”. https://megagonlabs.github.io/ginza/, (2023 年 05 月 26 日アクセス).
[2] “Wikipedia 山城地区(南山城地域)”. https://ja.wikipedia.org/wiki/山城地区, (2022 年 10 月 27 日アクセス).