探索的画像分析ツールの実装と社会調査への応用

加藤 遼*1
Email: ryou@sfc.keio.ac.jp

*1：慶應義塾大学政策メディア研究科前期博士課程

◎Key Words 社会調査，画像分析，解釈多様性

1. はじめに

ソーシャルメディアを使って情報を発信することが一般的になったことで、そこに集まる大量のデータを様々な分野で応用しようという試みが盛んに行われている。それらのデータは、ネットワークの時代であるからこそ社会調査の文脈で応用可能である。実際にTwitter のデータを世論調査へ応用する先行例もある[1]。しかし、これらの事例において用いられているのは、テキストや位置情報などの数値データのみであり、画像データを扱った事例はない。多くのソーシャルメディアが画像を情報として持つだけでなく、画像がメインコンテンツであるソーシャルメディアも主流となっている。画像は人の認識そのものであり、その画像の集積から社会を分析することが可能ではないかと考えた。そこで Twitter の画像データとテキストデータを対象に、分析者が”コンピューターと対話的”であり、”探索的”である分析を行うことで新たな知見の発見を促進するツールを実装し、社会調査への応用を試みる。

2. コンセプト

開発コンセプトとして、以下２点を明確にする。

2.1 試行錯誤の平易化

データから新たな気づきを得るためには、膨大な試行錯誤が必要である。大量のデータを人の力だけで分類していくのは時間がかかるが、大量のデータを整理し、絞りこみ、集計し、可視化することはコンピューターが得意とする作業である。そこで本ツールでは、ツール上で全行程をノンストップで行えるようにすることで、分析者が手を止めることなく簡単に試行錯誤を繰り返すことができるようにした。

2.1 分析者による画像の文脈理解と解釈

データを分析し知見を得る際に必要なのは、対象の背景知識と文脈の理解である。画像解析においては、近年技術の進歩は著しいものの、画像の持つ意味や文脈上の画像の役割などを解析できるまでに至っていない。しかしながら、人は前提となる知識があれば画像を見ることでそれらを理解することができる。そこで画像に付随するテキスト情報をツール上で解析し、その結果と自信の持つ前提知識をもとに分析者が画像を分類する方法をとる。このようにデータと対話的に分析を行うことにより、画像のもつ意味の解釈や文脈上の微妙な差異の発見につながると考える。

これら二つのコンセプトから、分析者は自身の持つ分析対象の知識や文脈への理解をもとに、コンピューターと対話的に分析を行い何度も試行錯誤を繰り返すことで新しい気づきや知識を得ることができる。しかし、本ツールは何かを入力すれば、解析された結果が返ってくるマイニングツールとは異なる。そのため、使用するにあたり分析者には分析対象への深い理解と文脈を見分ける知識が要求される。だからこそ分析者の多様な解釈を担保し、新しい知見の発見が行えると考えている。

3. 実装

3.1 分析対象データ

分析に使用するデータは Twitter 社が公式に公開している API を利用する。後述する検索ボックスに入力されたキーワードを含む最新のツイートを、一度に最大100 件取得する。取得したツイートデータから、1,ツイートをしたユーザーのプロフィール画像、2,プロフィール本文、3,ツイート本文を分析データとして使用する。

3.2 ツールの実装

Twitter のデータを対象に画像とテキストの分析を行う場合、1.データの取得、2.データクリーニング、3.解析、4.結果による絞りこみ、5.可視化を行う必要がある。さらに何度も試行錯誤を行うには、この過程を何度も実行することが必要となり、知識発見までのハードルが高い。そこで分析者が知識発見のみに集中できるよう、コンセプト 1 で述べたように、全ての過程をツール上で実行できるよう実装した。

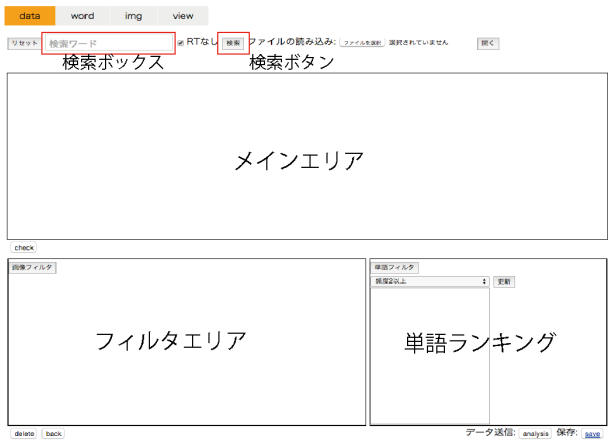

Fig .1　ツールの挙動プロセス 1

Fig .2　ツールの挙動プロセス 2

Fig .3　ツールの挙動プロセス 3

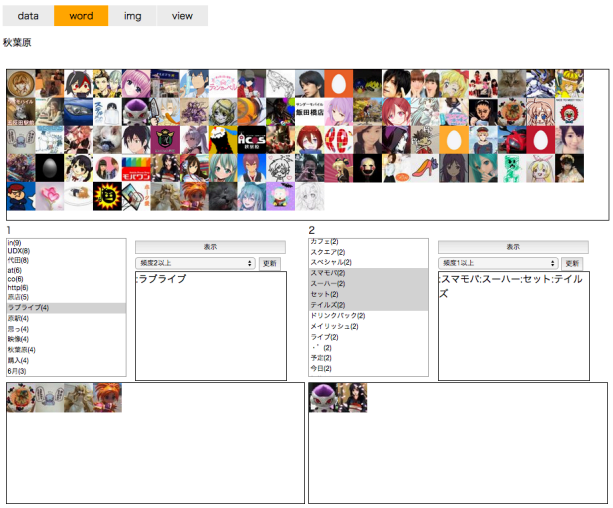

Fig .4　ツールの挙動プロセス 4

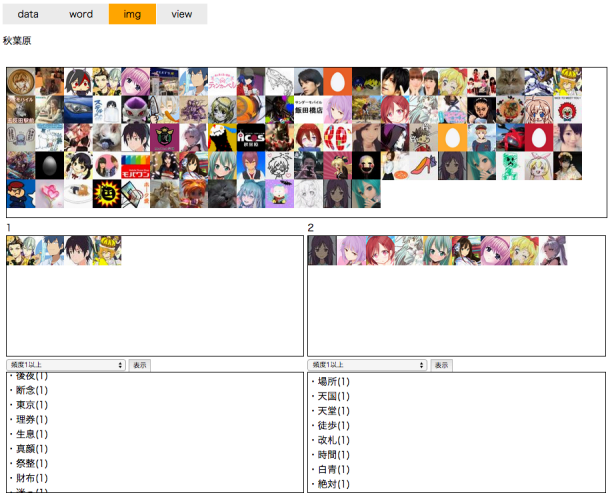

Fig .5　ツールの挙動プロセス 5

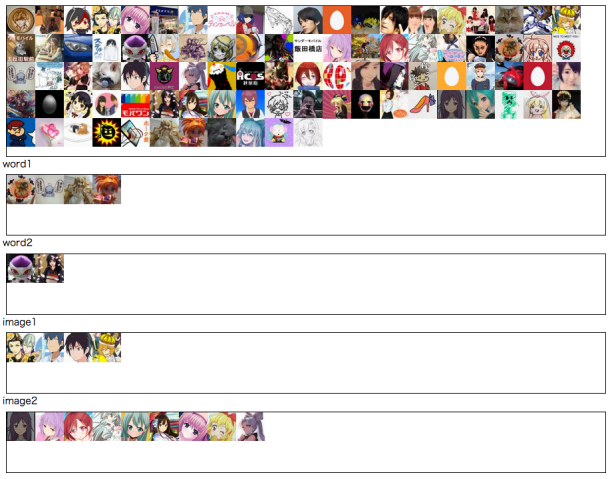

Fig .6　ツールの挙動プロセス 6

3.3　ツールの使用手順

(1) 検索ボックスに分析キーワードを入力する(Fig.1)。検索ボタンをクリックすることで検索キーワードを含む最新のツイートを 100 件取得する。100 件に満たない場合はその全てを取得する。更にデータ量を増やしたい場合は、もう一度検索ボタンを押すことで、前回取得した 100 件の次の 100 件を取得することができる。

(2) メインエリアに取得したツイートのユーザープロフィール画像が表示される(Fig.1)。この時、同じアカウントが複数存在する場合は一つだけ表示するため、取

得数と表示数が異なる。ここで取得数よりも表示数が少ない場合、そのキーワードは同じユーザーが多くツイートしているということになる。

(3) メインエリア左下の check ボタンを押すことで、取得したツイートのテキストを形態素解析した結果が表示される(Fig.2)。頻度を変えることで形態素解析の結果をしぼりこむことができる。

(4) 単語一覧から解析キーワードを選択し、単語フィルタをクリックすることで、選択した単語を含むツイートをしているユーザーの画像がフィルタエリアに表示される(Fig.3)。また、メインエリアからフィルタエリアに画像をドラッグアンドドロップすることで画像を個別に選択することも可能である。逆にフィルタエリアからメインエリアに画像を移すことで選択を解除することも可能である。

画像を選択したら画像フィルタをクリックすることで、選択した画像のユーザーのツイート本文を形態素解析した結果に単語ランキングが更新される。これらは逐次行うことが可能であり、単語で絞り込み表示された画像から不要な画像を抜いたり、新たな画像を追加したりして、再度単語を表示するといった作業を連続で行うことができる。

フィルタエリア左下の delete ボタンをクリックすることでフィルタエリアに表示されているユーザーのデータを分析データから削除することができる。これによって分析、解釈、クリーニングの３つを試行錯誤の過程で行うことが可能である。

(5) フィルタエリア右下の analysis ボタンをクリックすることで画像の比較を行う。クリック後、word タブを選択することで、分析データ内で異なる単語で絞り込まれた、二つの画像群を表示して比較する(Fig.4)。また、image タブを選択することで、二つの画像群のツイートに含まれる単語を表示して比較する(Fig.5)。view タブでは、これらの全てを表示することで全体と個別に抽出したものを俯瞰的にみることができる(Fig.6)。

(6) 上記の (4) ～ (5) を繰り返すことで探索的に解釈を繰り返し知見や仮説を発見する。データが不十分な場合 (1) ～ (3) を行うことでデータを追加し、分析データを作成していくことができる。

4. 実証実験

ツールを使った実証実験として、以下二つの調査を取り上げる。

4.1 秋葉原と池袋を利用するユーザー層の調査

Twitter は自分の日常をつぶやくなど、マイクロブログとしての役割がある。そこで地名に関するツイートから、その地域にどういう人が集まるのかを調査した。

秋葉原、池袋に関するツイートでは、アイドルの実写アイコンとアニメやイラストのアイコンが多く見られた(Fig.8)。そこで、それぞれのデータからアニメ・イラストアイコンのみを取得したところ、明確な差異が発見できた。秋葉原では女性イラストのアイコンが多くを占めたのに対し、池袋では男性イラストのアイコンが多くを占めた。「秋葉原」「池袋」ともにアニメグッズや同人誌を売る店が多く存在するが、秋葉原には萌え系や美少女系アニメなどを扱う店舗やメイド喫茶など男性向け店舗が多く、一方池袋には、「乙女ロード」と言われる通りがあるように、美少年系アニメや執事喫茶など女性向けの店舗が多く存在する。つまり、それぞれの地域に集まる人の性別の違いがここから見てとれる。一見すると、同じ種類のアイコンに見える画像群であるが、秋葉原、池袋に関する知識や、アニメにおける文脈を読み取る知識によって明確な差異を発見できることが証明された事例といえる。

「秋葉原」

「池袋」

Fig. 8 「秋葉原」「池袋」に関するツイートのユーザープロフィール画像

4.2 マイルドヤンキーの実態調査

マイルドヤンキーとは、2014 年に話題となり始めた「上京志向がなく、地元で強固な人間関係と生活基盤を構築し、地元に出たがらない若者たち」を指す、『ヤンキー経済　消費の主役・新保守層の正体』[2]の中で

原田曜平氏によって提唱された概念である。この調査では、本ツールを使いマイルドヤンキーの実態の視覚化を試みた。

分析キーワードとして「マイルドヤンキー」を選択した場合、得られる結果はマイルドヤンキーと考えられるユーザー本人ではく、マイルドヤンキーに関して発言している第三者のツイートである。そこで、『ヤンキー経済』の中でマイルドヤンキーとされる若者にとって「夢の国」と喩えられている「イオンモール」を分析キーワードとして採用した。国内 142 店舗のイオンモールについて、「イオンモール　店舗名」を検索キーワードとして検索を行い、言及しているユーザーのツイートを取得した。取得したデータから宣伝を目的とした商業アカウントであると考えられるものを排除し、分析を行ったところ実写アイコンが多いことが判明した。Twitter のアイコンに関する先行研究 [3]において用いられている分類に基づくと、実写アイコンは、”本人一人”(本人であると考えられる人が一人で写っているもの)、”本人複数”(本人と考えられる人を含めた複数人が写っているもの)、”他人”(有名人やアイドルといった本人ではない第３者が写っているもの)の３タイプに分類できる。

ここで、マイルドヤンキーの特徴であるとされている「絆」「仲間」「家族」といった意識が高いと考えられる「本人複数」に該当するアイコンをもつユーザーをマイルドヤンキーの定義に近い層と想定して抽出した。(Fig.7)

Fig.7 抽出したユーザープロフィール画像例

抽出後、検索により取得できたデータから商業アカウントを排除したアカウント数に占めるマイルドヤンキーと想定されるユーザーの割合を、各店舗におけるマイルドヤンキー度として算出した。

算出したマイルドヤンキー度を 3 段階に分類し、各店舗を地図にプロットした。(Fig.8)

※プロット全体 https://www.google.com/maps/d/edit?mid=zsuIFeLWWHK4.kap0uGhUA_m8

これにより、概念としてのマイルドヤンキーの実像を定量的に視覚化し、地域間の特徴に関して調査することを可能にした。こちらは、背景知識として提唱されている概念をもとに探索的に仮説を発見していくことで新しい知見を発見した事例といえる。

Fig.8 プロット例

5. おわりに

両実験では、取得されたデータから、分析者の持つ背景知識や文脈を介することで新しい気づきと解釈を行うことができた。マイルドヤンキーの例では、提唱されている概念をもとに調査を進めることで、インタビューなどの定性調査しか行えなかった対象に対して大規模な定量調査を行うことができることを示唆する。

一方で、本ツールには細かいものを含めると多くの課題が発見された。その中でも本質的なものを二つ挙げる。

第１に、結果の正当性を客観的に保証することができない点である。本ツールはコンピューターだけでは判断できない点を分析者の判断と解釈で補うというコンセプトのもと開発されているが、客観性を示す要素がないため発見した知見の正当性を判断することができない。また、分析者自身の想定する仮説に沿うようにデータを整形することも可能である。そこで、判断の際に参考となるような数値を占めす必要があるだろう。

第２に、データの一過性の問題である。これは Twitter の特徴であるリアルタイムの結果の影響を大きく受けることが原因である。現状最新のツイートのみを取得する仕様のため、データを取得する時間が変わるだけで、結果が大きく変わることもあり得る。これはデータ取得時に最新のものだけでなく、ある程度の期間のデータを取得しランダムサンプリングをしたデータを返す、または外部で長期にわたり取得したデータを読み込むなどの機能を追加することで解決することができるだろう。これによって、ツールの信頼性と結果の説得力が増すと確信している。

参考文献

(1) 山本仁志, 小川祐樹, 宮田加代子, 池田謙一：“Twitter における意見表明の規定要因”, 研究報告知能システム, 2013-ICS-170, 1-7 (2013)

(2) 原田曜平：“ヤンキー経済　消費の主役・新保守層の正体”, 幻冬舎新書(2014)

(3) 富永登夢, 土方嘉徳, 西田正吾：アイコン画像に注目した Twitter 研究の提案, 人口知能学会全国大会論文集, 28, 1-4 (2014)