# ソーシャルビッグデータの研究

○石川 博[*1)]

## 1. 目的・背景

まさにこの瞬間も現代の情報洪水は、社会の様々なセクタに大量の情報を生み出し続けている。この現象をビッグデータという。このビッグデータには、現実世界由来のデータとソーシャルメディア由来のデータが含まれる。これらを統合的に分析することで、どちらか単独だけの分析では得られないような価値を発見し、それをビジネスをはじめ防災や科学など多様な応用に生かすことが期待できる。

## 2. 研究内容

（1）実験方法

コンサートに参加した人々の多くは、会場の最寄りの駅から電車に乗って帰宅する。すると IC カードを通して実世界データとして乗降データが急激に増える。一方、参加者たちは、その時の興奮や電車の混雑に関する不満や情報をソーシャルデータに書き込む。すると、ソーシャルデータとして投稿数が同様に急激に増える。このように両者には疑似的相関がある。このとき、鉄道の管理者が乗降データの急激な増加の原因を知りたくなれば、乗降データとソーシャルデータの疑似相関を利用して、同一時間帯に同一地域で急激に増えた記事集合を適切な手段で分析すれば、共通原因である、参加したコンサートという顧客の関心にたどり着くことが期待できる。

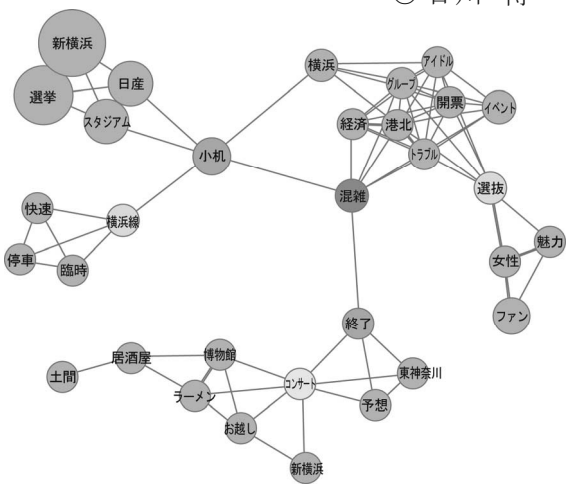


図 1. 共起ネットワーク

表 1. 観光地ランキング

| 順位 | 英語 | 仏語 | 露語 | ポルトガル語 | 韓国語 |
|---|---|---|---|---|---|
| 1 | 渋谷 | 原宿 | 原宿 | 神田 | 神田 |
| 2 | 原宿 | 渋谷 | 八重洲 | 原宿 | 日比谷 |
| 3 | 神田 | 神田 | 銀座 | 渋谷 | 渋谷 |
| 4 | 新宿 | 新宿 | 神田 | 永田町 | 新宿 |

また、日本を訪れる外国人がよく行く場所は、必ずしも日本人がよく行く場所とは限らない。そこで、東京の主要な地名を各国語の表記にして、条件としてツイッターの記事（ソーシャルデータ）を収集し、分析する。その結果として地名ごとに言及した記事数を言語別にランキングすれば、各言語圏の人々にとっての人気スポットの順位が分かる。さらに時間及び利用者（すなわちその同一性）を考慮すれば、人気のある観光コース（いわゆる黄金ルート）の発見にもつながる。

また、オープンアクセスジャーナルは、閲覧回数やダウンロード回数だけでなく、被引用回数も伝統的な雑誌に比べて早く手に入れることができる。そこで閲覧回数やダウンロード数の時系列データ間の類似性だけで、未来において、ある閾値（90）以上の被引用回数を持つことになる論文（Highly cited papers、HC）を発見できるか確認するために予備実験を行う。

（2）結果及び考察

まず駅名（新横浜など）と時間（日時）で収集したツイート集合の中で頻出する単語や互いに共起する単語を発見し、その単語をグラフのノードに、さらに互いに共起する 2 個の単語に対応するノードをエッジで結び、全体として共起関係を基にしたグラフを構築した。ただし、それぞれの頻度が適切な閾値以上の単語や共起関係だけをこのグラフの要素にするものとする。このグラフに対して、媒介中心性の高いノード（すなわち重要度の高い単語）に注目することで、混雑原因の概要が把握できることを確認できた（図 1）。

また、記述言語別に、都内の主要な地名（観光地）を含むツイートの件数を集計したところ、言語によって地名のランキングは異なることが分かった（表 1）。

さらに出版されてから 3 か月までのダウンロード回数のデータを持つ論文のサンプル（Public Library of Science から集めた 48,261 件の論文）について、その時系列データの類似度を基にスケーラビリティを有するクラスタリングを行ったところ、HC の論文を多く含むクラスタ 1 個を発見した。このクラスタにおける HC の数は、サンプルの論文集合における HC の総数（398 件）の 97.74%以上であったことから、最初の 3 か月分のダウンロード数だけの分析で、全体の 97%以上の HC が発見できる可能性を確認した。

## 3. 今後の展開

複数の実験を通して、実世界データとソーシャルデータを合わせて分析することで、新たな知見が得られることが確認できた。今後は、汎用的な分析技術を構築し、その普及と観光や防災などへ応用の拡大を行い、それを通してダイナミックな産業構造の創出に寄与する。

*1)首都大学東京