twitter解析とその可視化

東海大学大学院理学研究科M2　宗像昌平・船山貴光
東海大学理学部　山本義郎

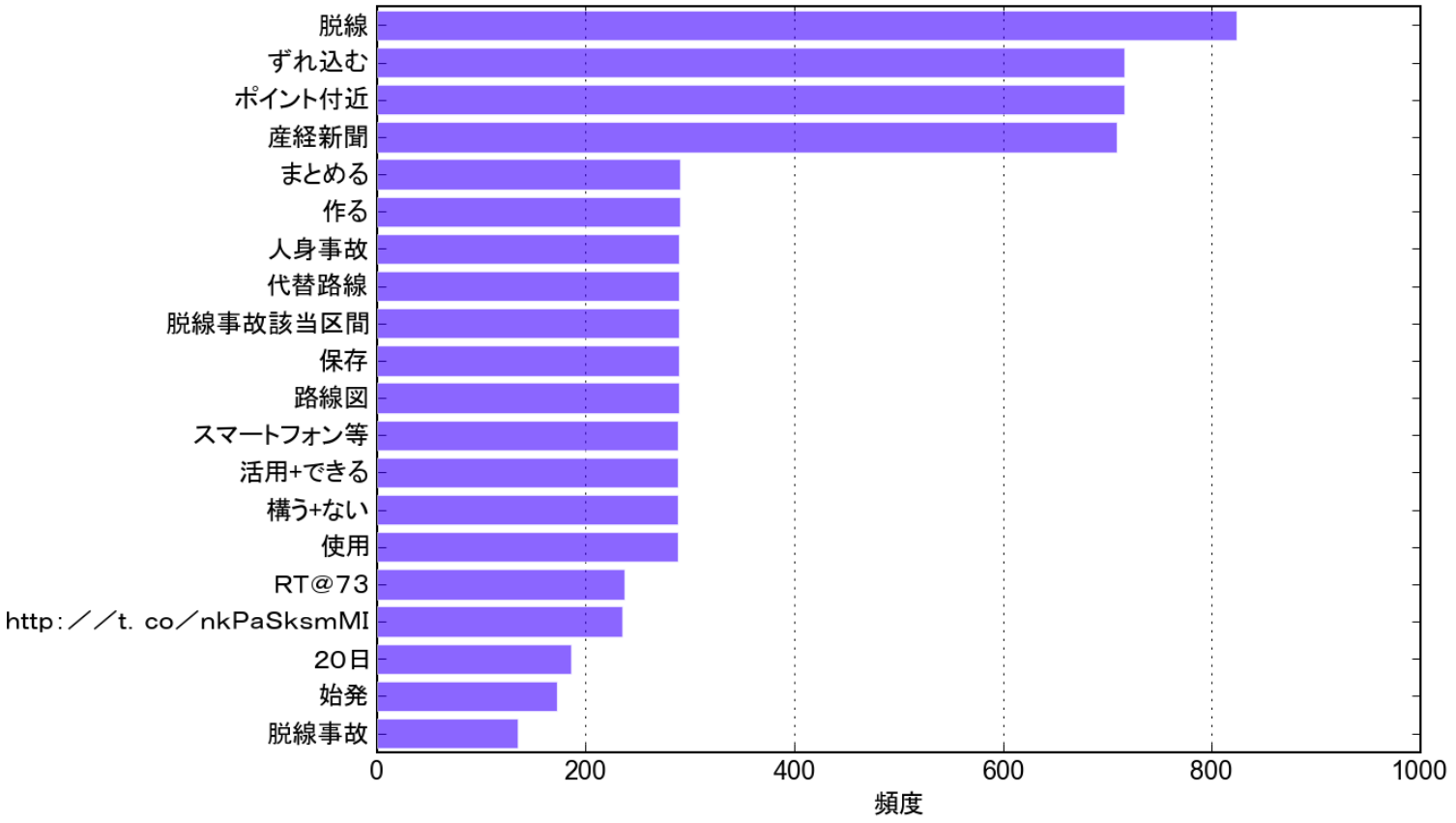
研究背景

- 近年、自然災害や公共交通機関のトラブルなどが多発している。このようなトラブル時の状況や2次災害に巻き込まれない為の情報などの防災情報を収集の際にSNSの情報を活用することを目的としている。本研究では、Twitterを用い、ツイッターのつぶやきから有益な情報をテキストマイニングにより抽出することにした。そこで、Text Mining Studioを用いて、どの様な分析が有効かを検討した。

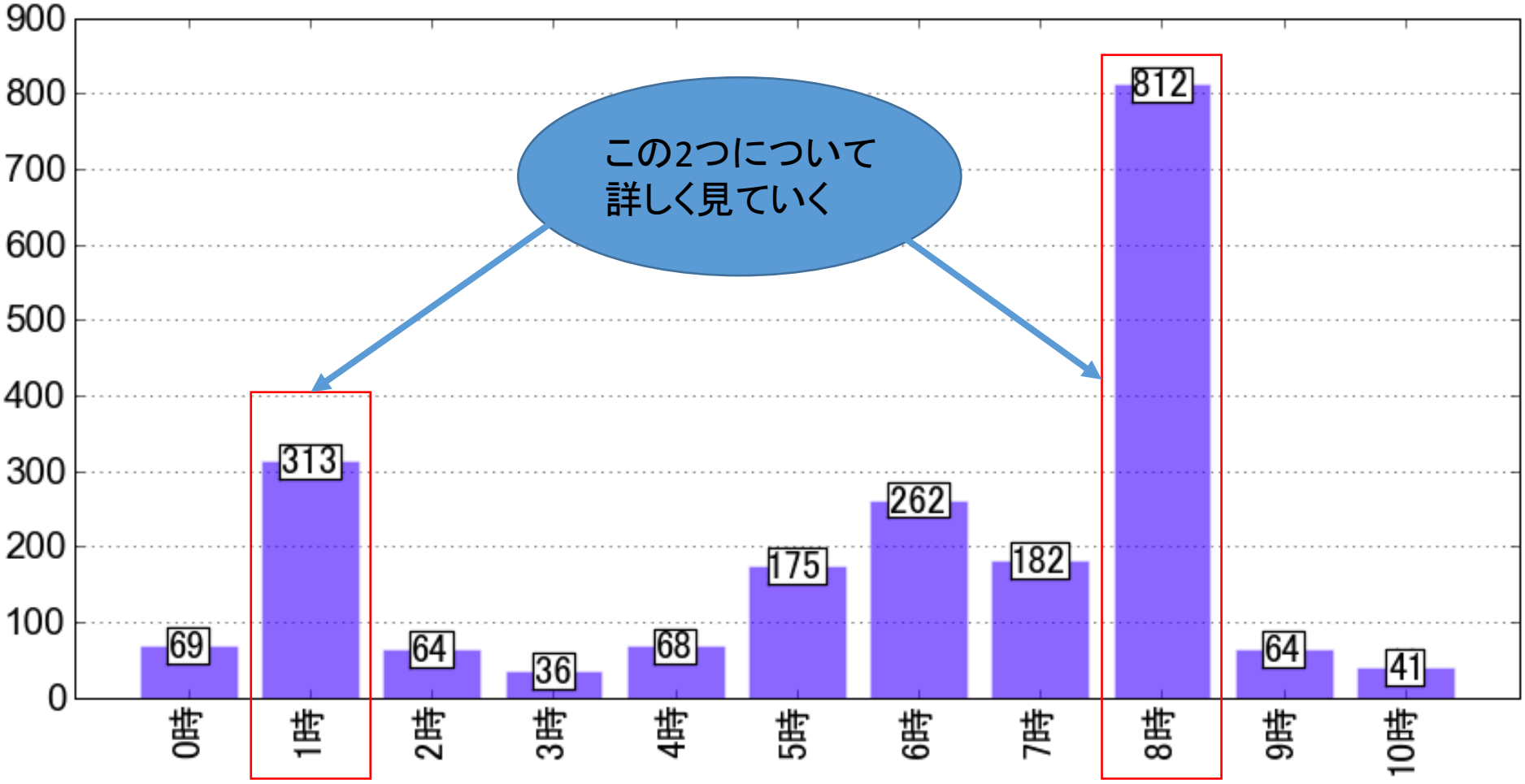
使用するデータについて

- 本研究では、2014年6月19日木曜日の18時9分頃に起きた小田急小田原線の相模大野駅構内での脱線事故後のツイートを分析した。本データは、『小田急』や『脱線』という単語が含まれるツイート(テキストデータ)である。また、データを取り始めた時刻は、事故が起きた日の深夜0時から朝10時までである。
- RのtwitteRパッケージを用いて、RとTwitterを連携し、streamRパッケージのfilterStream関数を使用してツイートを取得した。
- 今回のデータは、すべての文章に単語『小田急』が含まれているので、言葉の繋がりを見ることばネットワーク以外は、単語を除外して解析する。

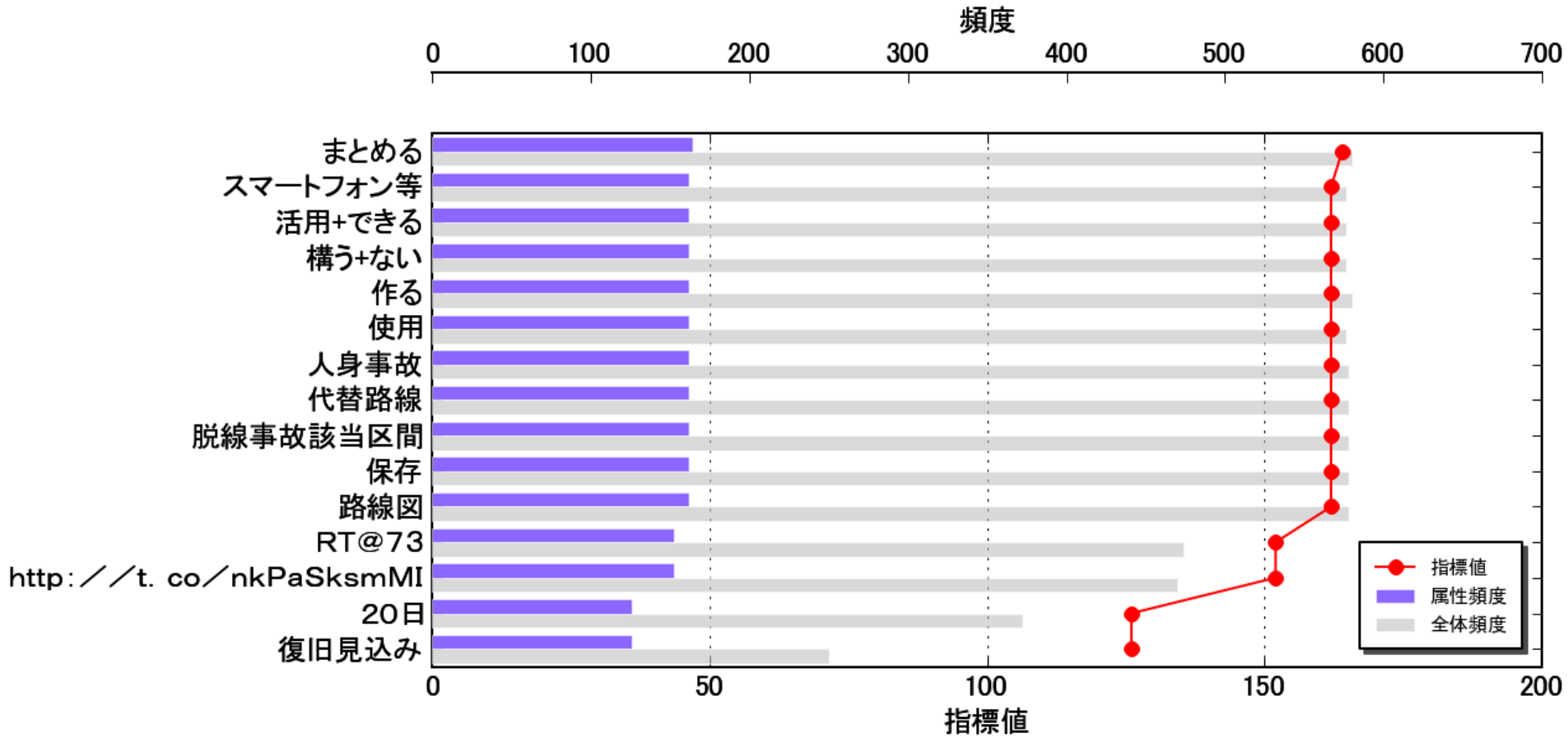
単語頻度解析(上位20単語)



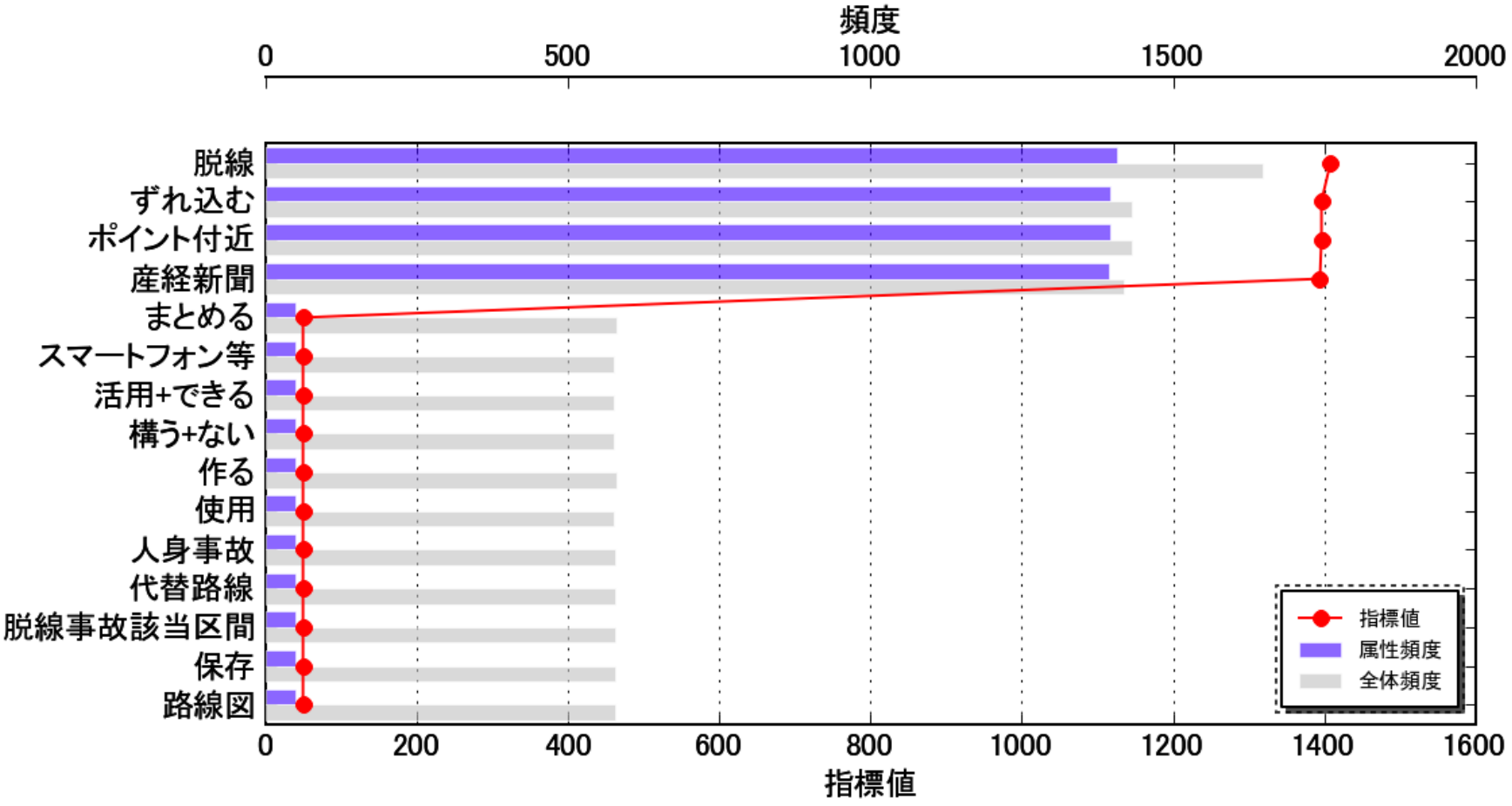
時間別のツイート数


900
800
700
600
500
400
300
200
100
0
この2つについて
詳しく見ていく
69
313
64
36
68
175
262
182
812
64
41
0時
1時
2時
3時
4時
5時
6時
7時
8時
9時
10時

1時台のツイートの特徴語抽出



８時台のツイートの特徴語抽出



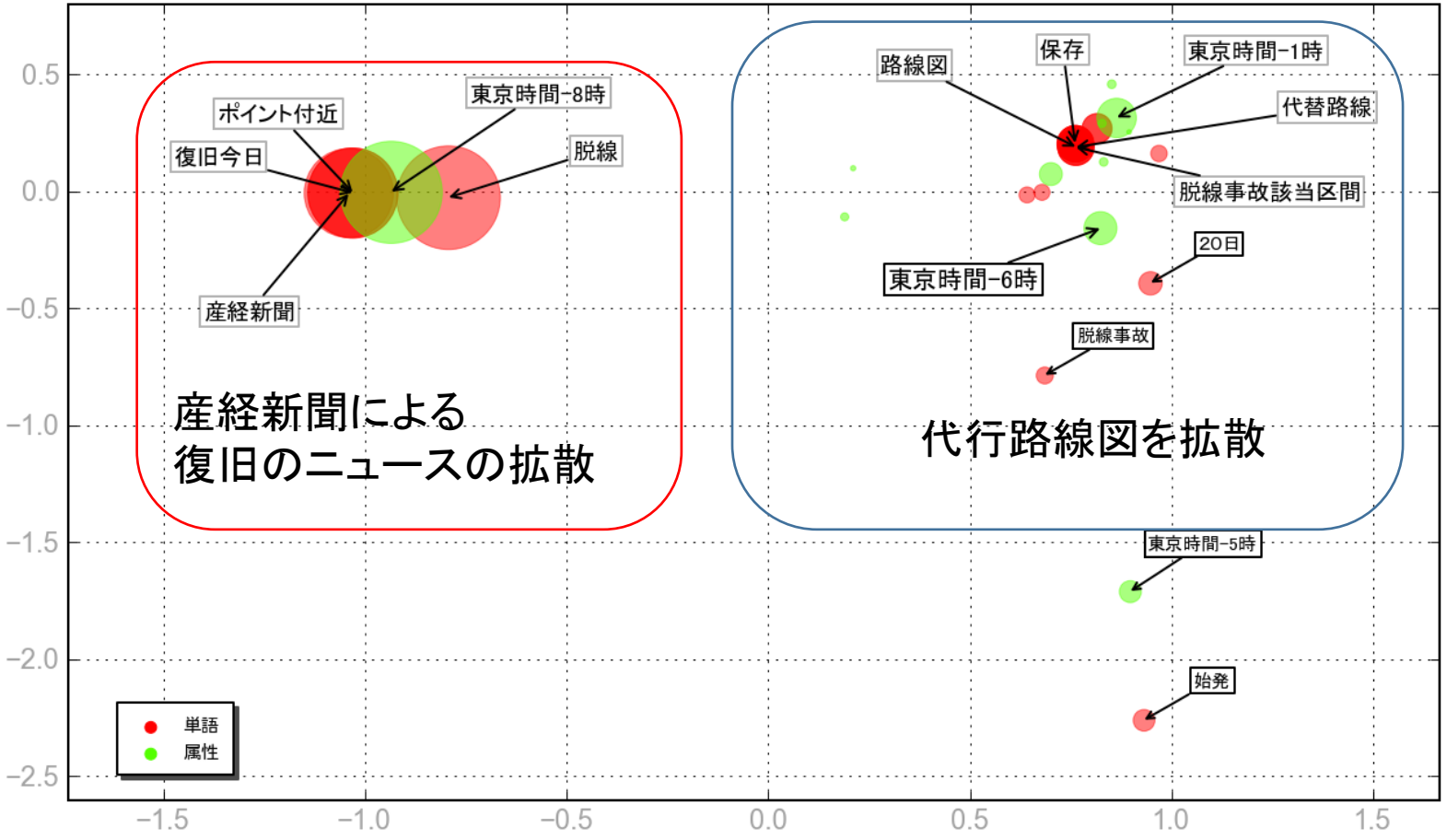
特徴語抽出の考察

- 特徴語抽出の指標値は頻度で行った。
- １時台のツイートの特徴分析からは、上位１１単語の指標値がほぼ同じ値になっており、残りの単語は大きく下がっていることがわかる。
- ８時台のツイートの特徴分析からは、上位４単語の指標値が高く、残りの単語が上位４単語に比べてかなり低いことがわかった。
- このままでは単語の頻度は分かるが、単語同士の関係が分からないのでそれぞれの原文を確認してみた。

特徴語抽出の考察

- １時のツイートの多くは『小田急線の脱線事故該当区間と代替路線を**まとめた**路線図を作りました。人身事故の時なども活用できるので、スマートフォン等に保存して使用していただいて構いません。』というRTである。
- ８時のツイートの多くは『小田急ポイント付近で**脱線**復旧きょうにずれ込み産経新聞』というツイートであることがわかった。
- 次にこのデータに対して属性を時間にした対応バブル分析を行い、結果を比較する

属性を時間にした対応バブル分析の結果



対応バブル分析の考察

対応バブル分析からは以下のことがわかった

- １時台の話題は『代行路線』『路線図』『保存』が近くに存在しているので、代行路線のツイートを拡散している時間帯だと考えられる。
- ８時台の話題は『産経新聞』『復旧今日』『脱線』が近くに存在しているので、産経新聞のニュースを拡散していることがわかった。

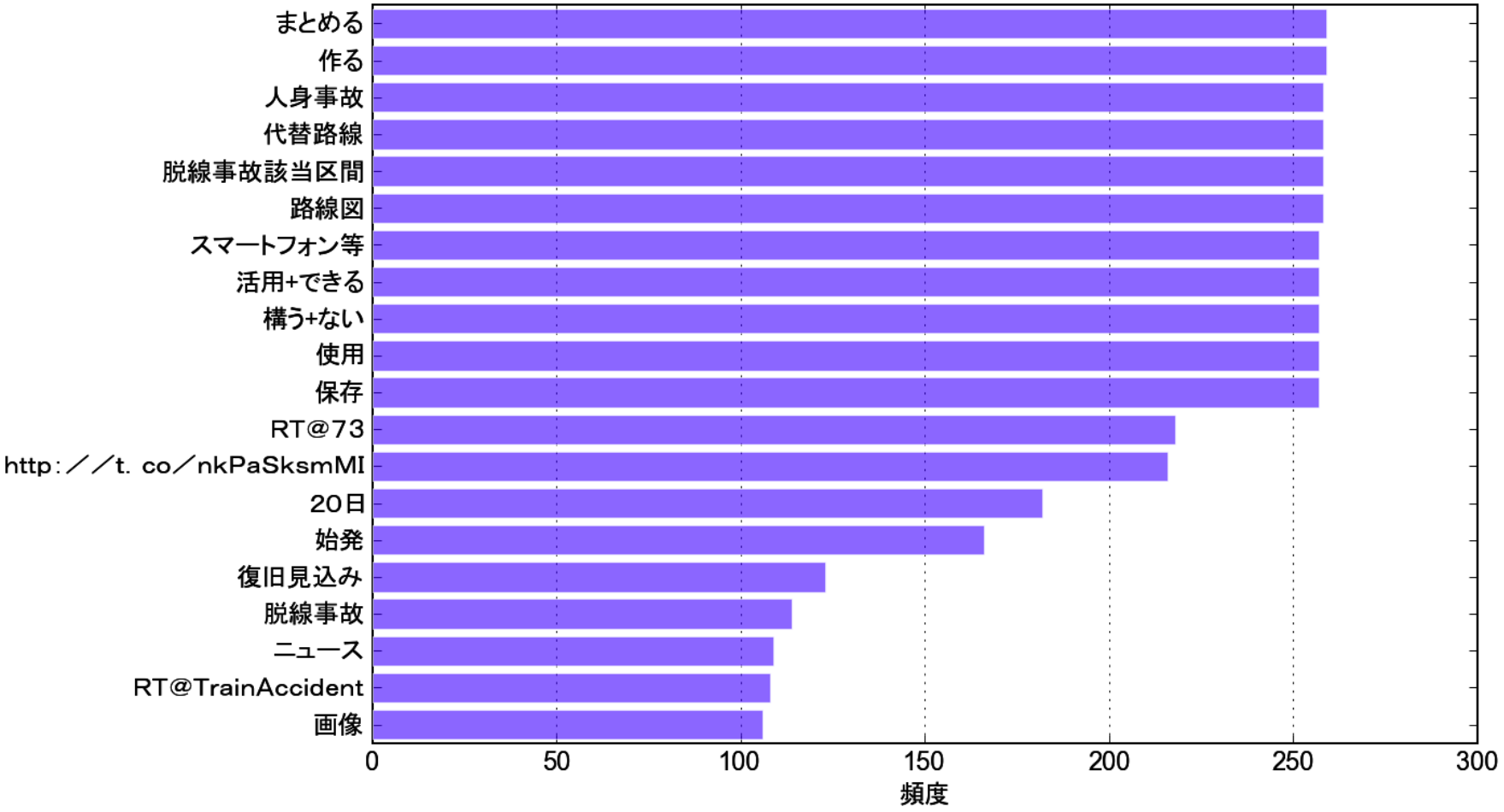
以上のことより、特徴分析の結果から原ツイートを確認した作業が対応バブルからわかった。

また、５時と『始発』が近いのは、ユーザーが始発からの復旧を気にしているからであり、原文からは『小田急始発も一部復旧せず』というツイートが拡散されていたことがわかった。

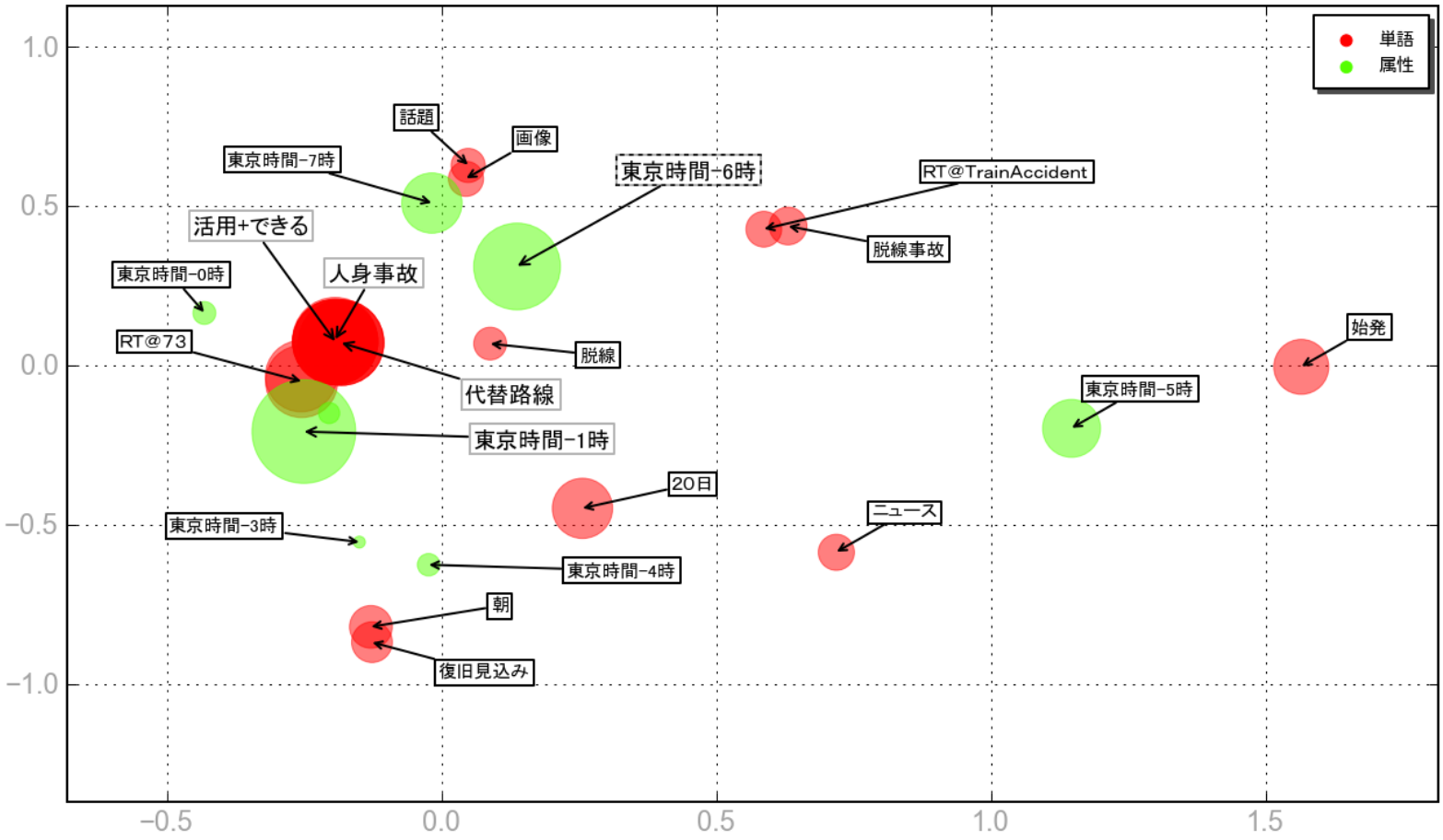
- さらに私たちは、産経新聞の復旧に関するニュースが発表される前後の反応を比べるために、深夜0時から7時までと8時から10時までの2グループに分け、詳しい分析を行った。

- 分析方法は以下の3つの方法である。
  - 単語頻度分析
  - 対応バブル分析
  - ことばネットワーク分析

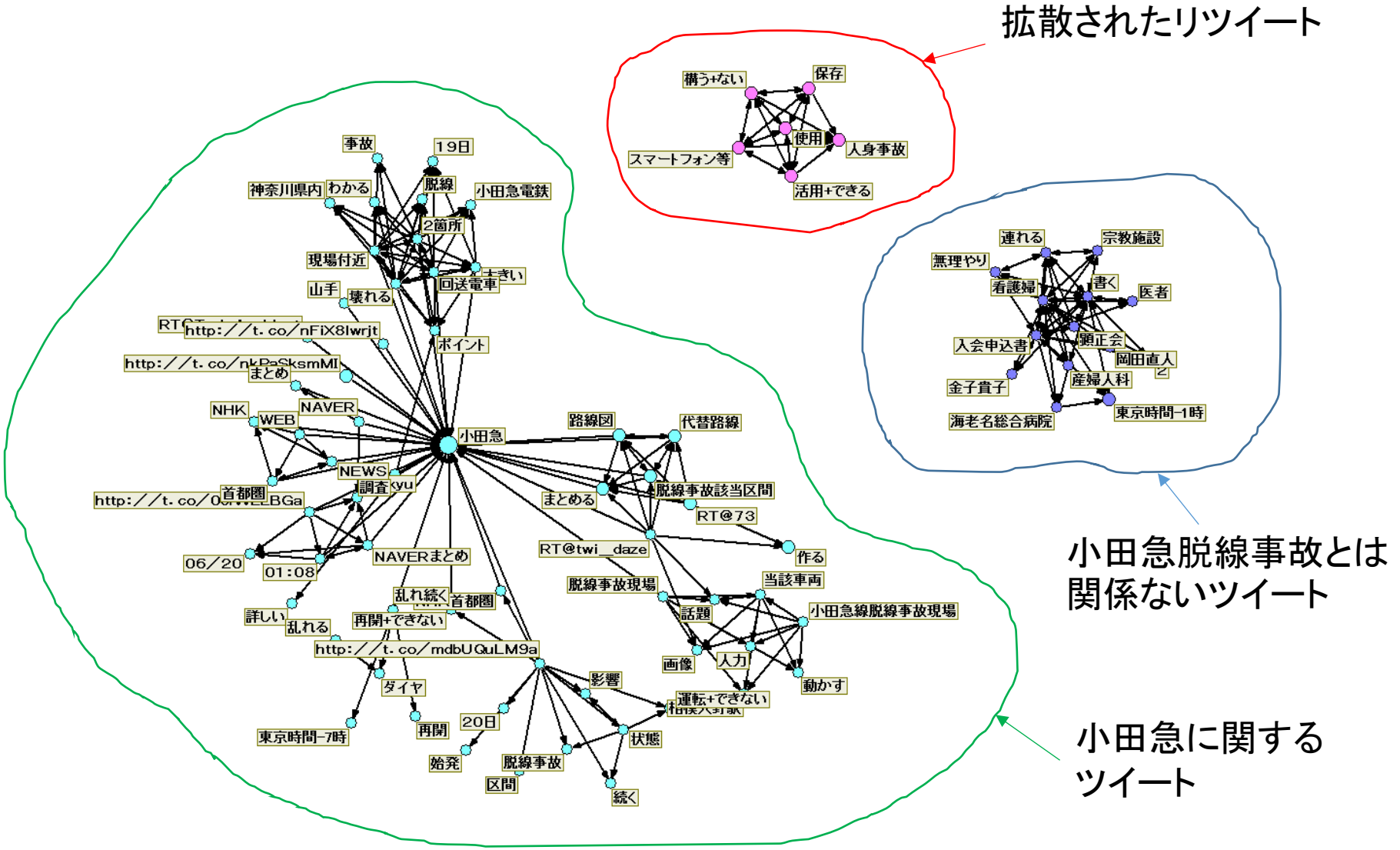
深夜0時から7時までの単語頻度解析



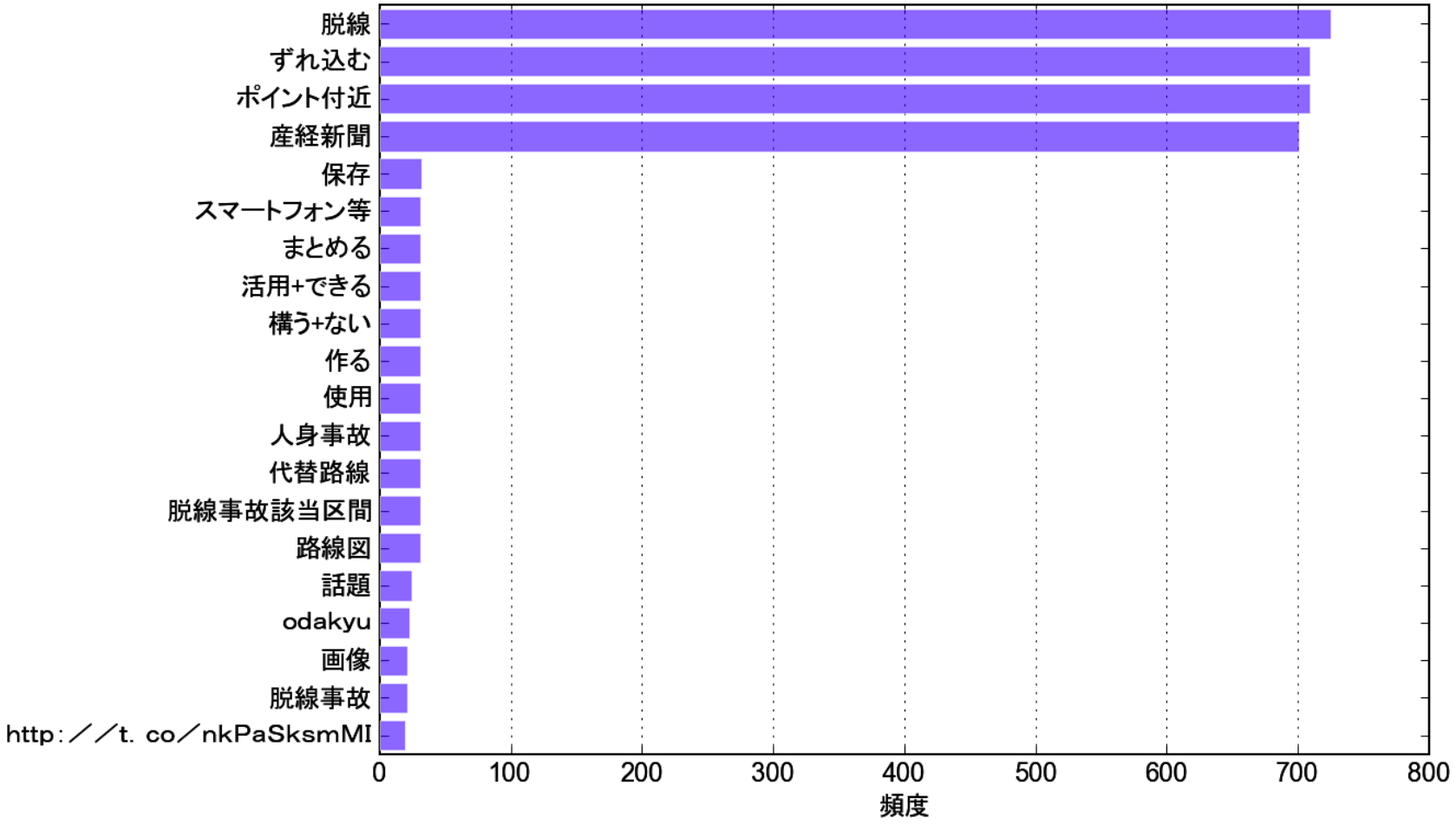
深夜0時から7時までの対応バブル解析



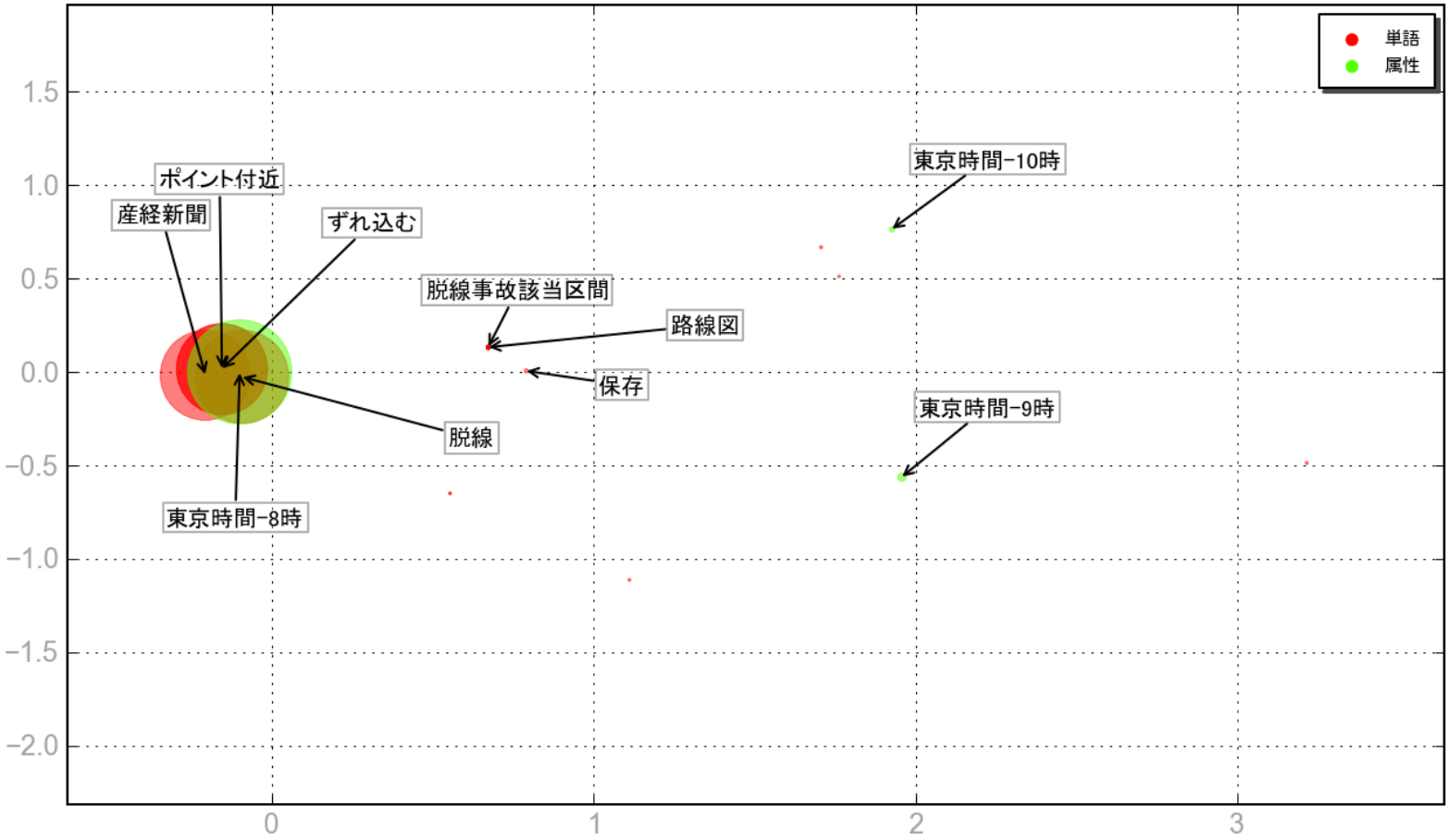
深夜0時から7時までのことばネットワーク



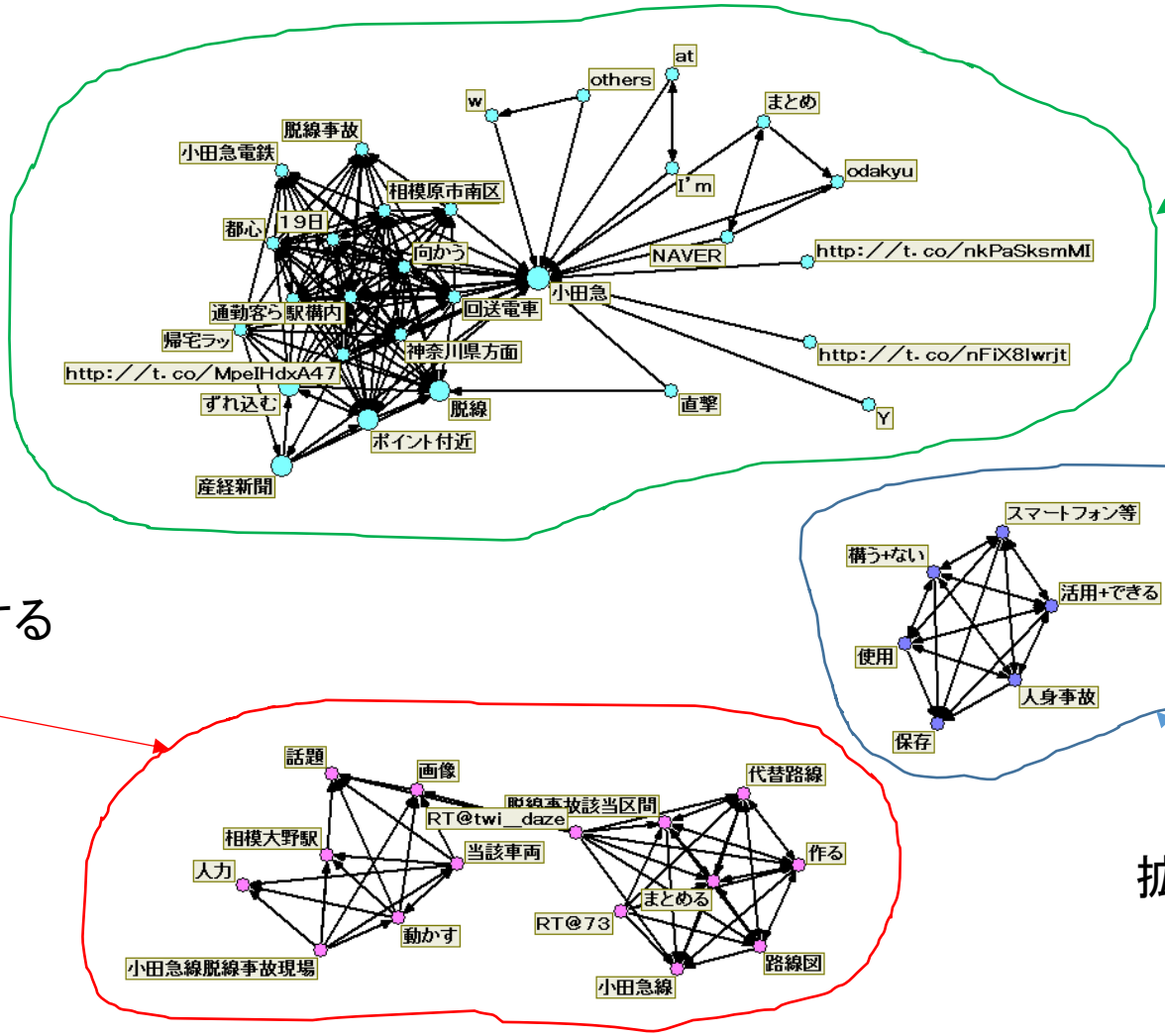
8時から10時までの単語頻度解析



8時から10時までの対応バブル解析



8時から10時までのことばネットワーク



考察

深夜0時から7時についての考察

- 対応バブル分析により時間ごとの細かい情報が読み取れるようになった。
- ことばネットワークからは、全体の分析からは確認できなかった脱線とは関係ないネットワークが見られた。

8時から10時についての考察

- 対応バブル分析からは、8時の円が大きく、残りが小さいので8時以降は小田急に対するツイートが少ないことがわかった。
- ことばネットワークからは、代行路線図が話題の画像になっていつことがわかった。

またことばネットワークからは、どちらも拡散されたリツイートが
単独のネットワークで分かれていることがわかった。

まとめ

- 分析をすることでツイート時間ごとの話題が、特徴分析と対応バブル分析の２種類でできることがわかった。

- グループごとの解析は、全体の解析では見られなかった『顕正会』や『話題』に関する情報がでてきたので、この分析方法も有益な方法だと考えられる。

- 今後は、他のツイッターデータの解析についても同様の方法を試していこうと考えている。

参考文献

- [1]石田　基広(2008)Rによるテキストマイニング入門.株式会社　森北出版
- [2]石田　基広,小林　雄一郎(2013)Rで学ぶ日本語テキストマイニング.株式会社　ひつじ書房
- [3]金　明哲(2009)テキストデータの統計科学入門.株式会社　岩波書店