修　　士　　論　　文

マイクロブログ上の話題抽出と
その真偽に関するユーザの態度分類

Discovering Topics from Microblogs and
Classifying Users' Attitude Towards
the Truth of the Topics

指導教員　　喜連川　優　教授

東京大学　情報理工学系研究科

電子情報学専攻

氏　名　　48-106448 藤川　智英

提　出　日　　平成24年2月8日

概 要

近年、Twitter などのマイクロブログでは様々な情報が共有されるようになっているが、その中には誤情報も少なくない。それらの中には人間の安全や名誉に関わるものも多いため、いち早く検出してサイトなどを通じて告知することが必要であると考えられる。

本研究ではユーザの反応からそうした情報を検出することを考えるが、そのためにはまず何が話題になっているかを知ることが必要となる。話題解析は 2 つの意味で重要である。まず、何が話題になっているかはそれ自体が多くのユーザの関心ごとになっている。また、何らかの理由で話題に対する人々の反応を知りたい場合も、何が話題になっているか、そしてそれについて言及している投稿はどれかを知る必要がある。

本論文ではこれらの問題を解決するため、何が話題になっているかを確率分布と PageRank を応用したシステムで検出し、さらに、可能ならば要約を作成、そして、関連する投稿の検出までを行う手法を提案する。

そして、ユーザの反応を信疑と根拠の有無で分類することを考える。これは、ある情報の正否を判断するにあたって、疑っている人がどの程度いるか、どのような根拠が提示されているかが重要と考えられるからである。

これらを通じて、話題の検出から誤情報の可能性があるかどうかの判定までを一貫して行うシステムを提案する。

謝辞

本論文は筆者が東京大学大学院情報理工学系研究科電子情報学専攻修士課程に在籍中の研究成果をまとめたものです。本研究の実施の機会を与えて戴いた指導教官である喜連川先生に感謝の意を表します。また、研究室の皆様は輪講などを通して研究の欠陥の指摘やアドバイスを頂き感謝します。特に、鍛治先生と吉永先生は遅れがちな私の論文に対して辛抱強くアドバイス下さり誠に感謝いたします。

2012 年 2 月 8 日

目次

図 目次

表 目次

第1章　はじめに

1.1　研究背景

Twitter などのマイクロブログの登場は情報の共有を手軽にし, 現在ネット上では様々な情報が非常に早く広がるようになっている．しかし，それらの中には根拠のない風説（流言[1]）も含まれており，特に災害などの非常時には爆発的に流言が広がることが多い. 例えば東日本大震災では以下のような流言が少なくとも 80 件広がった [1].

- ホウ酸を食べると放射線を防げる
- 外国人犯罪多発
- 東電の社員が逃走した

こうした流言は人間の安全や名誉に関わるものも多いため，迅速に発見し，サイトなどを通じてその存在を明らかにすることが重要と考えられる.

しかし，Web 情報をユーザがどう信じるかに関する研究では，正しい情報というのは必ずしも客観的に定義できないため，最終的な判断を行うのはユーザであり，正しさをシステムによって完全に客観的に決定することはできないと指摘している．なぜなら、情報の信憑性は「情報の受け手によって認知される特性」であるため、その判断基準は人によって異なるからである [2]。

そこで、情報信頼性に関する研究では、「信憑性」を直接判断するのではなく、「信憑性」を判断するための材料を提示することを目指す研究も存在する。情報分析システム WISDOM [3]では、「電気自動車は環境に良い」「裁判員制度」など

[1] 流言という用語は従来，真偽の不確かな情報一般に対して用いられうるが、本稿では文脈上誤解のない場合、誤った情報を参照する意味で用いる。

知りたい意見を入力すると、それに対する賛成意見、反対意見がどの程度存在するか、どのような人がどのような意見を述べているかなどを一覧表示することで判断の手助けを行うことを目指している。

そこで、流言の可能性のある情報を発見し、ユーザに判断材料を与えるためには、まず話題の抽出を行い、さらにその話題に対して人々がとっている反応を分析することが必要であると考えられる。

このうち最初のステップである話題の抽出は２つの点で重要である。まず、何が話題になっているかは、それ自体が多くのユーザーの関心ごとになっている。話題を解析するサービスはすでに buzztter[2], ついっぷるトレンド[3]などが存在し、Twitter 自体も公式で話題情報を公開している。

次に、流言にかぎらず、様々な話題に対して人々の反応を知りたいなら、まず何が話題になっているか、その話題と関連する投稿はどれかを知る必要がある。

URL ごとにどのような反応があるかを調べるのなら、すでに tweetbuzz[4], Ceron.jp[5]などのサービスが存在する。しかし、次のような例を考えるとこれでは不十分であることがわかる。

- A: 実質、アメリカの空を飛んでいるのは大半が倒産経験のあるエアライン …→ アメリカン航空が破産法申請 – MSN 産経ニュース http://t.co/XXX #yjfc_airlines
- B:えーっ/アメリカンエアが破産 – 毎日新聞 http://t.co/YYY
- C:げっ! アメリカンエアが倒産? (;￣O￣)

この例では、A と B は同じ話題に言及しているが、異なる新聞社の記事のため、URL,タイトルともに異なっている。また、C は同じ話題に言及しているが、URL は貼りつけていない。

このように、同じニュースが別々の新聞社によって、異なる URL、異なるタイトルで配信されていることも多いため、URL 単位では同じ話題を別々とみなしてしまう。また、話題の中にはスポーツ中継など、URL を発信源に持たないものも多いため、URL なしで話題に言及している場合を取得することができない。そのため何らかの手段で関連投稿を判別する手法が必要である。

[2] http://buzztter.com/ja
[3] http://tr.twipple.jp/
[4] http://tweetbuzz.jp/
[5] http://ceron.jp/

1.2 本研究の目的と貢献

話題の検出では先駆け的存在である [4]をはじめとし、確率モデルを仮定する手法を提案した [5]やブログに対して応用した [6]など様々な研究が存在する。また、ある話題が存在した時、その要約を作ることを目的とした研究 [7] [8]や、流言に関連する投稿を抽出、疑っている投稿を分類するという本研究に近いものも存在する [9]。ただし、 [4] [5][6] [7] [8]の研究では話題となっている単語の抽出は行なっても、それと関連する投稿の抽出は行なっていない。また、 [7] [8] [9]では何が話題になっているかは既知とした上で、話題に応じて正規表現などを手作業で作り関連投稿の抽出を行なっている。

本研究では話題の検出、関連投稿の抽出、そして可能ならば代表的な文（要約）の抽出までを自動で行う手法を提案する。また、関連投稿に現れるユーザーの反応を解析し、流言かどうかを判断するための材料を提供する手法を提案する。

まず、話題抽出では、各投稿の単語の出現数の変化から、その単語の話題度(バースト)のスコアを計算する。更に、何かが話題になっている場合は複数の単語の出現比率が同時に増大し、共起が起こることに着目し、PageRank を応用した手法を用いて各単語のスコアを再計算、クラスタリングを行い、さらに、各投稿がその話題に関するものかどうかを、含まれる単語とそのスコアの合計から判定する手法を提案する。

次に、流言の可能性があるかどうかの判定手法として、話題に関連する投稿を「信疑の有無」と「根拠の有無」という 2 つの基準で分類することを考える。

まず、「信疑の有無」の分類を行い、もし疑っている投稿が多いならば、その情報に流言の可能性があると推定できる。次に、各投稿を「根拠の有無」で分類すれば、その情報が間違っている、あるいは正しいと主張する何らかの根拠を含んでいるかを判定できる。そうした投稿を集めて提示することが出来れば、各ユーザにその情報が流言かどうかを判定する手がかりを与えることができる。

全体のシステムは、以下のようなものを想定している。まず、Streaming API で Twitter の投稿（ツイート）を収集し、後述する話題抽出アルゴリズムを適用して、話題を解析する。次に、解析した話題からクエリを作成し、Filtering API を用いて関連するツイートを収集する。それに対して「疑っているかどうか」と「根拠の有無」で分類を行う。

疑っている人がどの程度いるかを表示するとともに、根拠を伴って意見を述べているツイートを提示し、流言かどうかを判断する手がかりを与える。

1.3　本論文の構成

第 2 章　出現率の急上昇した単語の発見、PageRank を用いたスコアの再計算などを元にした話題の抽出、関連投稿の収集手法を述べる。

第 3 章　判断材料を提供するため、ある話題に対する反応について、「信疑の有無」「根拠の有無」で分類を行う手法について説明する。

第 4 章　全体のまとめと今後の課題について述べる。

第2章　話題の抽出

2.1　予備実験

話題抽出において、単語の出現数の急上昇(バースト)を検出する手法は大きく分けて 2 つ提案されている。一つ目は確率的オートマトンを用いた方法 [4]、もう一つは確率分布を仮定し、その出現回数が生じる事象の確率の低さから、バーストのスコアを計算する手法である [5] [6]。

しかし、いずれの場合も、単語の普段の出現数との比較からバーストかどうかを判定するという点では共通している。そのため、まず各単語について普段どの程度出現しているかの情報を保持する辞書を作成しなければならない。そこで、まず Twitter における単語数がどの程度かの調査を行った。

2.1.1　実験設定

Twitter の各ツイート（投稿）の収集には Streaming API の Garden hose アクセスを用いた。これは、Twitter のツイートの 10 分の 1 をリアルタイムに取得することができる API である。

ツイートのコピーである RT は元のツイートを参照し、同じツイートは 2 度目以降は無視した。また、一人のユーザからの影響を過度に受けないように、10 分以内の同じユーザの投稿は無視した。

次に、各ツイートが日本語であるかどうかを、日本語の文字（ひらがな、カタカナ、漢字）のいずれかが 2 つ連続し、かつその内の一方がひらがなかカタカナであるような文字列が含まれているかどうかで判定した。ここで、日本語の文字が 2 つ以上連続した文字列を探索するのは、外国語のツイートでも、顔文字などでひらがななどが一文字使われることがあるためである。また、2 つの文字の内

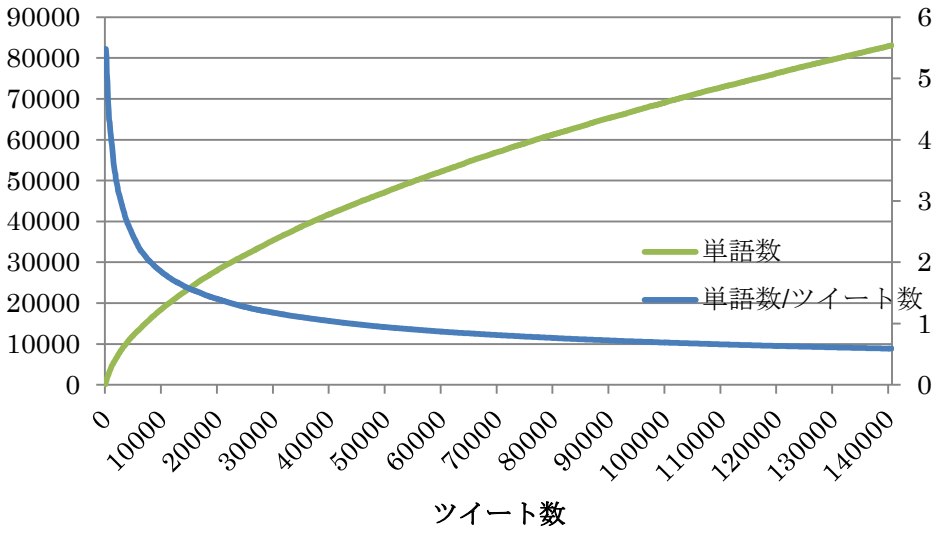

図 2.1 ツイート数と単語数の関係

一方がひらがなかカタカナとするのは、漢字の連続である中国語のツイートを除外するためである。

次に、各ツイートに対して以下のような前処理を行った。

元の文字列	変換後の文字列
http://www.yahoo.co.jp/ などの文字列	UURRLL
@username などのユーザーネーム	USERNAME
発言主のユーザーID、名前	USERNAME
RT(QT) USERNANME などの非公式RT	RTorQT
￥＾！などの全角記号	半角記号
改行	“ “(空白)
wwwww など w の 3 つ以上の連続	www

形態素解析には Kuromoji[6]を用いた。ただし、非自立語、記号(!-/:-@ など)、漢字以外の一文字単語、2 桁以下の数字は除いた。

また、単語の出現回数は、1 つのツイートに複数回出現した場合でも、1 回とカウントした。

[6] http://www.atilika.org/

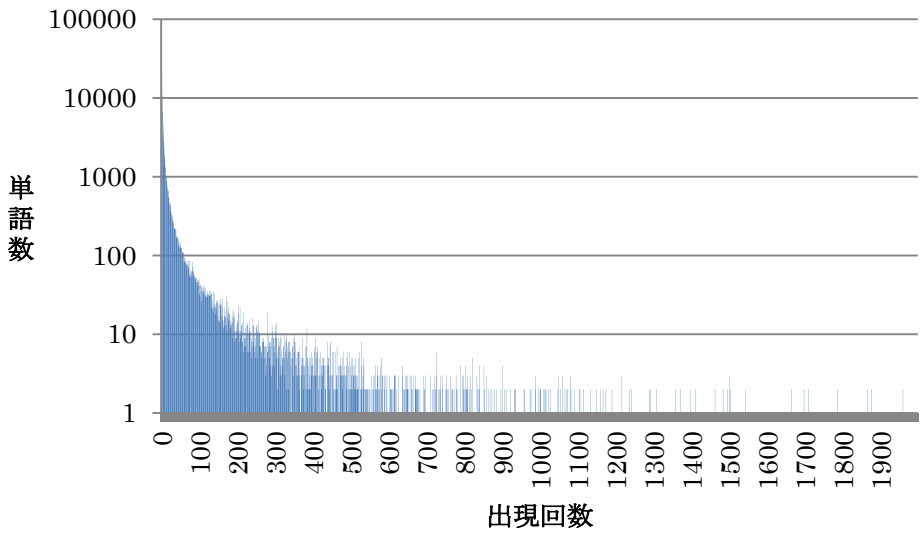

図 2.2 出現回数と単語数の関係

2.1.2 単語の出現数と頻度分布

まず､約 100 ツイートごとに､ツイート数と単語の種類数を調べた結果が図 2.1 ツイート数と単語数の関係である。このグラフから、ツイート数の増加と共に単語の種類も増加するが、その増加率は徐々に減っていくこと、ツイート数に対する単語の種類数も急激に減少することがわかる。

次に、単語の出現回数と、それに該当する単語の種類数の関係を表したものが図 2.2 出現回数と単語数の関係図 2.2、出現回数 50 回未満の単語の全体に占める割合のグラフが図 2.3 である。これらのグラフから、Zipf の法則が成り立っており､出現回数が極少数のところに単語のほとんどが集中していることがわかる。そのため、出現回数が極少数の単語を削除するだけで、辞書の容量を大幅に削減できることがわかる。

これらの結果に基づき、本研究では、100 万ツイートごとに辞書に含まれる単語のうち出現数が 1 の単語を削除し、1000 万ツイートごとに辞書に記録した出現回数を 2 分の 1 にし、さらに、その時点で出現回数が 5 回未満の単語を削除することで、辞書の容量を抑えている。

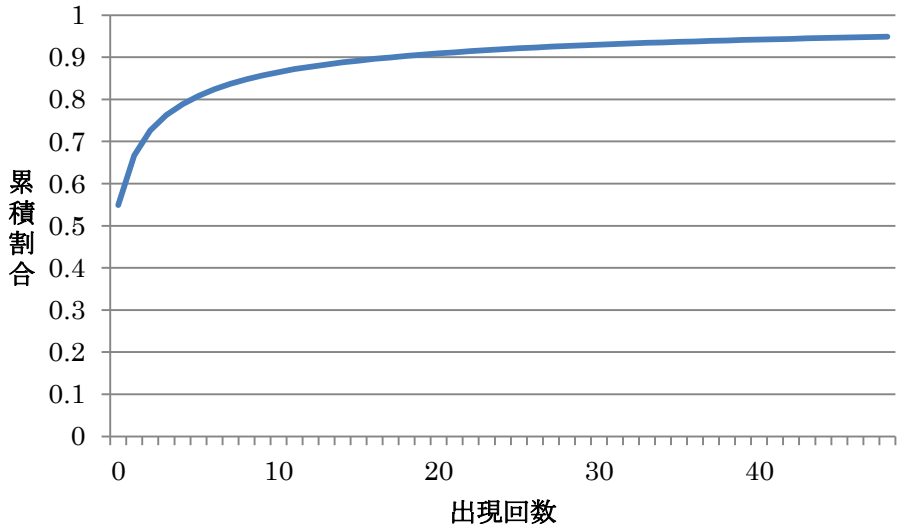

図 2.3 単語の出現回数と累積割合の関係

2.2 バーストスコアの計算

2.2.1 確率分布を用いた計算

前述したように、単語の出現率の急上昇(バースト)を検出する方法は 2 通りあるが、確率的オートマトンを用いる方法は計算コストが大きく、また、パラメータの設定を必要とする [6]ため、確率モデルを仮定する方法を用いる。

確率モデルを仮定する方法では、まず各単語についてそれまでの出現数のデータから、出現数がある確率分布に従うと仮定し、それのもとで、直近の出現数が確率の低い事象ならば、バーストとみなす手法である。この手法では、パラメータの設定を必要とせず、また、どれほど確率が低い事象かという形で、バーストの度合いを定量的に図ることができる。

2.2.2 確率分布の適合度検定

確率分布として、[5]では二項分布、[6]では正規分布を仮定しているが、まず、各単語がどのような確率分布に従うか確かめなければならない。そこで、2 つの確率分布に対して適合度検定を行った。

まず、出現率が 1000 分の 1 以上の単語をランダムに 100 個選び、10000 ツイートごとに出現数を記録したサンプルを 388 個集めた。そして、それぞれの単語

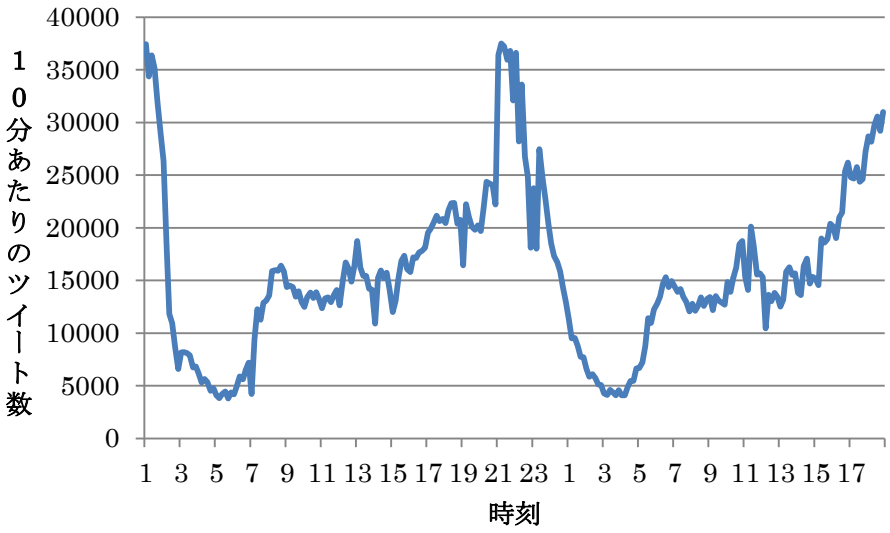

図 2.4 2 日間に渡る 10 分あたりのツイート数の変化

に対して、有意水準を 0.01 と設定して適合度検定を行った。

その結果、二項分布では 100 個中 37 個が棄却されず、正規分布では 76 個が棄却されなかった。必ずしも満足できる結果ではないが、どちらも実際の分布と大きく外れていないと考えられる。

この結果だけから考えると正規分布のほうが適切に思えるが、正規分布を用いるのは 2 つ問題がある。まず、今回の実験では適合度検定がしやすいように 10000 ツイートごとに区切ってサンプリングを行ったが、実際にはツイート数は図 2.4 のように時間に応じて大きく変化する。そのため、同じ出現数でも時間によってバーストかどうかは異なってくる。 [6]では正規分布を用いているが、これはブログを対象とし、1 日単位で計算している。1 日ごとのブログの記事数の変化はツイート数と違いそれほど大きくないと考えられるため、正規分布でも問題ないと言える。

また、正規分布では分散の計算を必要とするが、そのためには過去のデータを保持しておかなくてはならず、実際 [6]では 90 日分のデータを使用している。

対して、二項分布では以下の式で確率分布を計算する。

$$\mathrm{prob}(\mathrm{n},\mathrm{k}) = p^k(1-p)^{n-k}\binom{n}{k} \qquad (1)$$

ここで、$\mathrm{p} = \frac{\text{過去にその単語が出現したツイート数}}{\text{過去の全ツイート数}}$（単語の平均出現率）で、n はバース

トを計算したい時間帯のツイート数、k はその間に単語が出現したツイート数である。

このように、二項分布ではその時間帯のツイート数を式に含めることができるため、ツイート数の変動に追随して確率分布を計算することが可能である。また、そのために用いるパラメータは平均出現率 p だけであり、その単語のそれまでの出現回数さえ保持していれば計算できる。このような利点から、本研究では単語の出現数の確率分布として二項分布を仮定する。

2.2.3 具体的な計算手順

まず、それまでの単語の出現回数からそれが現れる確率 p を計算する。ただし、新しく登場した単語の場合、それまでに蓄積されたツイート数によって p が大きく変動してしまう。例えばそれまでに 100 万ツイートを収集している場合は 100 万分の 1 になるが、1000 万ツイートを収集している場合は 1000 万分の 1 になる。また、この場合、出現確率が極めて小さく計算されるため、1 回登場しただけでも「確率が低い事象」とみなされてしまう。そこで、下限 MINIMAL_PROB を設定し、以下のように計算する。

$$\mathrm{p} = \mathrm{MAX}\left(\frac{\text{過去にその単語が出現したツイート数}}{\text{過去の全ツイート数}},\quad \mathrm{MINIMAL\_PROB}\right)$$

本研究では MINIMAL_PROB=1/100000 と設定した。

そして、各単語について、調べたい区間に対して、まず出現率が全体の平均出現率を比べる。出現率が平均出現率以下ならばバーストの可能性はないと考える。超えているならば、n=その間のツイート数、k=その間に単語が出現したツイート数として式(1)を用いて確率を計算し、その対数を取ったものをバーストの度合い、スコアと定義する。

$$\mathrm{score} = -\log \mathrm{prob}(\mathrm{n},\mathrm{k}) = -\log p^k(1-p)^{n-k}\binom{n}{k} \qquad (2)$$

各単語について、直近の 10 分間、20 分間、30 分間……と 10 分単位で区間を増やしながら、スコアを計算していき、減少に転じる直前の部分のスコアを最終的なスコアとして設定する。また、その区間をバースト区間と定義する。例えば直近の 10,20,30 分間のスコアが 7,10, 8 なら、その単語のスコアは 10 で、直近の 20 分間をバースト区間とみなす。

ここで、その事象が発生する確率が 5%未満である、すなわちスコアが-log0.05 以上であるような単語をバーストの可能性があるとする。ただし、単語の出現数

k が 5 未満の場合は、サンプル数が少なすぎるため、バーストとはみなさない。

2.3 PageRank を応用したバースト スコアの調整

前項では確率が 5%以下の事象を「バーストの可能性がある」とみなしたが、例えば 1 万の単語があれば、そのうち 5%、500 個はたとえバーストしてなくても、この 5%の領域に入ってしまう。そのため、実際に話題になっているものと、偶然バーストしているように見えるものを区別する必要がある。

本研究では、何かが話題になる場合、複数の単語が同時にバーストし、共起が起こることが多いことに着目する。

- 例：サッカーの日韓戦
  - 「サッカー」「日本」「韓国」「本田」（選手名）
- 例：「スマトラ沖で地震」というニュース
  - 「スマトラ」「地震」「津波」「マグニチュード」「インドネシア」

そこで、共起している単語があるほどその Score を高くみなすことで、本当にバーストしているかどうかを区別することができると考えられる。

本研究では、これを実現するために、PageRank を応用する。PageRank では、以下の仮定が成り立つとして Web ページの重要性を判断する。

- 多くのページからリンクされているページほど重要なページである。
- 重要なページにリンクされているページほど重要なページである。

単語の話題の度合いの場合も、以下の仮定が成り立つと想定できる。

- 多くの単語と共起している単語ほど話題になっている単語である。
- 話題になっている単語と共起している単語ほど、話題になっている単語である。

ここから、リンクを共起度に置き換えれば、共起度が高いものほどバーストスコアが大きくなるように再計算することができる。これにより、他のバーストしている単語と共起している､本当にバーストしている単語のランクをあげられる。また、話題の中心となっている重要な単語は他の多くの単語と共起していると考えられるが、このような単語のランクをさらに上げることができる。

実際、PageRank を本来の用途とは異なり、類似度が高いものほどランクを高くするために使用した例として、VisualRank [10]がある。

2.3.1 具体的な計算手順

単語の関連性を計算するにあたっては時間的な類似性を用いる研究も多い [4] [5]。しかし、Twitter ではめまぐるしく話題が移り変わり、時間的な類似性が必ずしも関連性を意味しない。例えば同じ時間に始まる別々のテレビ番組が同じ時間的な類似性を持ちうる。また、歌番組では、そのタイトルと出演歌手の単語は類似性が高いと考えられるが、歌番組を表す単語は 1 時間の間バーストするが、出演する歌手は登場する短い間のみバーストするというように時間的類似性は異なる。

時間的類似性と単語の共起の両方を重み付けて考慮する研究 [11]も存在するが、どちらにどの程度重み付けを行うかには恣意性が加わるため、本研究では確実に関連性を表すと考えられる共起度のみを用いた。

まず、２つの単語w_i,w_jの共起度を以下の式で計算する。

$$c_{i,j} = \frac{|w_i \cap w_j|}{|w_i \cup w_j|} \cdot (idf_i + idf_j)/2$$

ここでは、共起度として jaccard 係数にw_i,w_jの idf の平均をかけたものを用いている。ここで、jaccard 係数の計算に用いるのはw_i,w_jのバースト区間のうち、より小さい方である。また、idf の平均をかけるのは、珍しい単語ほど共起しにくく、そのぶん重要であると考えられるためである。なお、$c_{ii} = 0$である。

ただし、計算コストの削減とノイズの除外のため、共起度がしきい値未満のものは 0 と設定する。本研究ではしきい値を 0.10 とする。

そして、この共起度を要素とする行列を C とおく。

$$\mathrm{C} = (\mathrm{c_{i,j}})$$

次に、推移行列 A は C を各列において正規化したものである。i 番目の列の要素の合計をs_iとおくと、

$$A = (a_{i,j}) = \begin{cases} \frac{c_{i,j}}{s_i} & (s_i \geq 1 \text{ のとき}) \\ c_{i,j} & (s_i < 1 \text{ のとき}) \end{cases}$$

PageRank を用いたバーストスコアの再計算は以下の式の繰り返しで行う。

$$\mathrm{R} = \mathrm{dAR} + (1 - \mathrm{d})R_0$$

ここで、R は i 番目の単語のスコアを表す。R_0は i 番目の要素が単語w_iの最初のバースト スコアを正規化したものであるような縦ベクトルである。d は一般的な PageRank の例にのっとり、0.85 と設定する。

R の初期値はすべての要素が等しい単位ベクトルである。この計算を繰り返し、R の変化率が 10%未満となるまで繰り返す。

2.4　クラスタリング

クラスタリングは一般的な階層的クラスタリングを用いる。ただし、クラスタ間の類似度 r の計算は以下のように、各単語のスコアに基づいた重み付けを用いて行う。

クラスタ$\mathrm{C_i}$，$\mathrm{C_j}$が存在した時、そこに含まれる単語を$\mathrm{C_i} = \{\mathrm{t_{i_1}}, \mathrm{t_{i_2}}, \dots, \mathrm{t_{i_n}}\}, \mathrm{C2} = \{\mathrm{t_{j_1}}, \mathrm{t_{j_2}}, \dots, \mathrm{t_{j_m}}\}$とすると、

$$\mathrm{r}(C_i, C_j) = \sum_{\mathrm{k}} \sum_{\mathrm{l}} \mathrm{w_{i_k}} \mathrm{w_{j_l}} \mathrm{c_{i_k} } j_m$$

ここで、$\mathrm{w_{i_k}}, w_{j_l}$は単語$\mathrm{t_{i_k}}, t_{j_l}$のスコアをそのクラスタ内のスコアの合計で割り、正規化したものである。しきい値は 0.10 とした。

また、クラスタのスコアは、そのスコアに含まれる単語のうち最大のスコアとする。

2.5　関連する投稿の抽出

次に、クラスタと関連のあるツイートを抽出する。各ツイートに対して、クラスタ内の単語の正規化された重みの合計をそのツイートのスコアとし、それが一定の閾値以上ならば関連するツイートであるとみなす。

例えば「少女時代、KARA が紅白に内定」というニュースで、各単語の重み付けが次のようになっていたとする。

- 少女：0.15
- 時代：0.25
- KARA:0.30
- 紅白：0.30

このとき、あるツイートに「少女」「時代」が含まれていればスコアが 0.15+0.25=0.40、「KARA」「紅白」が出現していればスコアは 0.30+0.30=0.60 というように計算する。

これにより、URL などと関係なく、「中心的な単語が多く含まれているかどうか」のみで話題と関連があるかどうかを判定するため、テレビ番組などが発信源の話題に対しても適用することができる。

2.6 各クラスタの詳細解析

2.6.1 トレンドの種類

トレンドにはいくつかのパターンが存在する。本研究では話題を次の５つに分類する。

- 一般的な話題
  - テレビ番組など、発信源が WEB 以外にあり、それに対して多くの人が言及している場合
- ニュース
  - 新聞記事やブログ記事などが話題となっているもの
  - 例：地震の前兆？ / 太平洋、ニュージーランドでクジラ 70 頭座礁 http://...
- 反響の多いツイート
  - 一つのツイートが大きな話題を読んでいる場合。非公式 RT で言及されることが多い
  - 例 : RT @XXX: オリンパスの歴代経営陣による悪質粉飾決算は大問題。ホリエモンは実刑で塀の中。何故オリンパスのお年寄り歴代経営陣にマスコミも検察も甘いのだろうか?
- 診断・分析系
  - 「診断メーカー」[7]を中心とした、各ユーザーに応じて何らかの「診断結果」「分析結果」を返すサービス
  - 例：USERNAME を一言で表すと……
- テンプレ系
  - 一つのツイートが話題になった時、その改変が数多く投稿されるパターン

[7] http://shindanmaker.com/

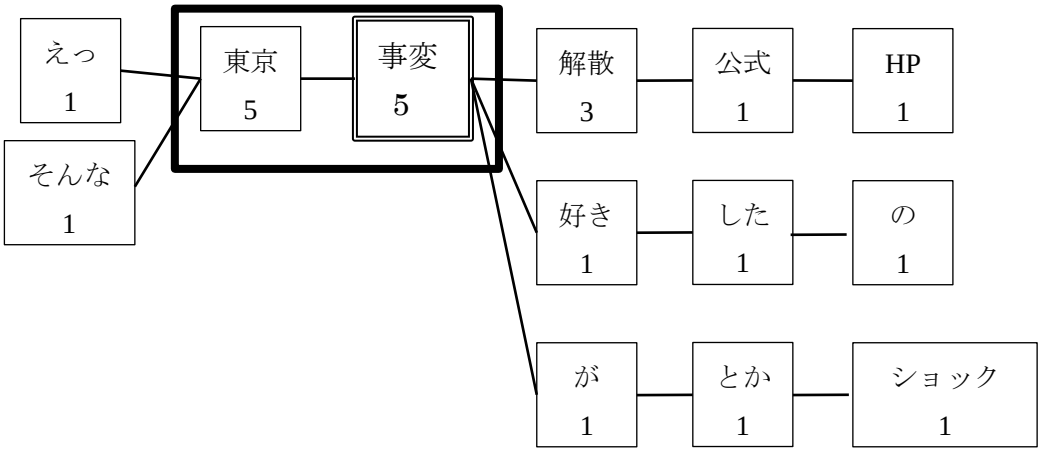

図 2.5 フレーズ、頻出文字列の抽出

2.6.2 フレーズ・代表的な文字列の取得

ニュース、反響の多いツイート、診断・分析系では同じ URL や文字列が何度も登場する。そこで、それらを取り出すため、 [7]の手法を用いる。この手法では、キーとなる単語を中心として、隣接する単語のグラフを作り、最も重みが大きくなるグラフを取り出し、それを要約とみなす。

例えば以下のような例を考える。

- 東京事変解散、公式 HP で発表 http://t.co/xxxxxxxx
- えっ、東京事変解散したの
- そんな、東京事変が……
- 東京事変解散とかショックすぎる
- 東京事変好きだったのに……

ここで、「事変」という単語を中心にグラフを作成すると図 2.5 のようになる。

元の論文では各グラフから最も重みの大きい経路を選択しているが、本研究では次のような改良を加えている。まず、出発点の単語の出現回数の 3/4 以上出現している単語の連なりを「フレーズ」（ひとかたまりの言葉）とみなす。この例では「東京」-「事変」がコレに当たるため、「東京事変」というフレーズが存在するとみなす。

次に、出現回数が全体の 1/20 かつ MIN_SUPPORT 以上の経路を探索する。本研究では、MIN_SUPPORT は 4 と設定している。

この時、次のうち少なくともひとつの条件を満たすなら「文」とみなす。

1. 「は」「が」などの係助詞、格助詞が含まれる

2. 「存在」「解散」などのサ変接続名詞が含まれている

ただし、複数の「文」が見つかった場合は、1 は満たす文のほうが 2 より優先される。それでもまだ複数の文の候補がある場合、重みの合計が大きい物を取り出す。重みの合計は次のように計算する。まず、各単語ノードの重みを count/log(length+1)で計算する。文が不必要にここで、count はその単語の出現回数、length は中心ノードからの距離である。各経路についてこの重みを合計し、それが最も大きい物を取り出す。

以上の操作を左右のノードに対して行うが、ここで、単純に左右の最も大きな経路をつなげてしまうと問題が生じる。例えば同じニュースに対して次のような２つのタイトルの記事が存在したとする。

- 応募条件「コネのある人」 岩波書店が縁故採用を公式宣言
- 老舗出版社、採用の応募条件に「コネのある人」

ここで、「コネ」を中心に前述の操作を行い、左右をつなげると、『老舗出版社、採用の応募条件に「コネのある人」 岩波書店が縁故採用を公式宣言』という実際には存在しない、また、同じ物を表す単語が左右に登場する不可解な文になってしまう。

そこで、最初に右方向の経路のみを抽出し、次に、それに左方向の経路の候補を加え、実際に存在している文かどうかを判定するという手法を取る。ここで、右方向の経路を最初に探索するのは、「東京事変が解散」「金正日が死去」のように中心となる単語（「事変」「金正日」）が○○という形の文章が多いためである。この場合では、『「コネのある人」 岩波書店が縁故採用を公式宣言』という文をまず取り出し、次に左側の経路を加えたものを作り、実際に含まれる文化どうか確かめていく。

以上の方法で文を抽出し、次にその文が含まれるツイート群に含まれている URL を調べ、最も多いものを代表的な URL として取り出す。ここで、その結果によって次のように分類する。

- 文が「RTorQT」で始まる場合、非公式 RT であるため、「反響の多いツイート」と判定する
- 文に「USERNAME」が含まれる場合、各ユーザに対して別々の結果を出すものと考えられるので、「診断・分析系」とみなす。また、特に診断メーカーの場合が非常に多いため、URL が診断メーカーのものだった場合も、「診断・分析系」とみなす。

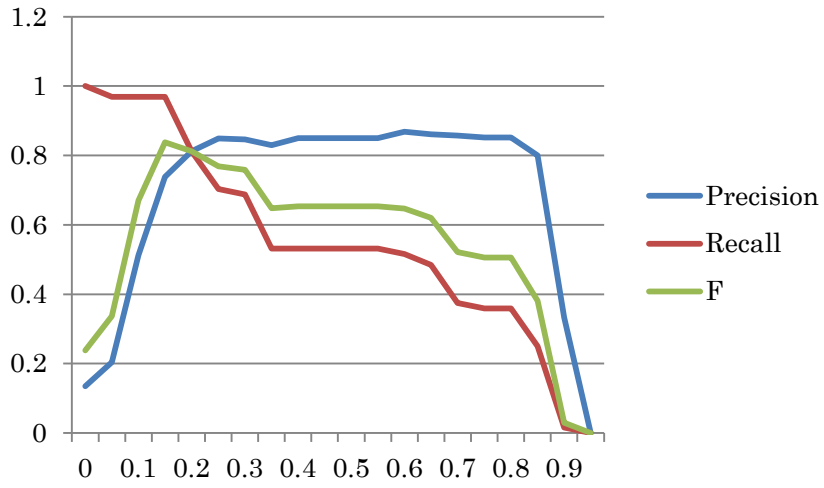

図 2.7 しきい値と各種指標の関係

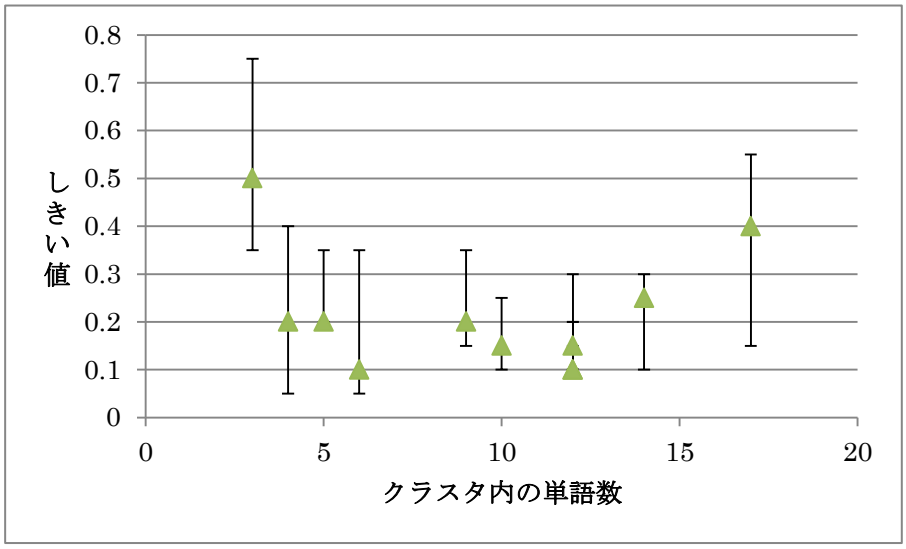

図 2.6 F 値を最大にするしきい値の分布

- それ以外の場合は何らかの記事が話題になっているとみなし、「ニュース」と判定する

2.7 関連投稿の抽出実験

関連投稿を抽出するにあたって、その閾値を設定するための実験を行った。11個のニュースに対して、関連するツイートかどうかの判定精度を調べた。なお、

すべてのツイートを調べると莫大な数になり、かつ、そのうち関連するツイートは多くても全体の 0.5%ほどであるため、クラスタ内の単語が最低でもひとつ含まれているもののみを対象とした。

各ニュースに、閾値を 0.05 刻みで変化させ、Precision, Recall, F-値を計算した。例えば「Twitter が国の状況により、検閲を行う意向を表明」というニュースでは図 2.6 のようになった。

各ニュースに対して、F 値が最大となるしきい値、更にその F 値から 0.1 引いたものを「許容値」と仮定して、それを満たす最小のしきい値と最大のしきい値を調べた。その結果をグラフにしたものが図 2.7 である。

当初クラスタ内の単語数が多いほど、スコアが分散するためしきい値は低くなると予想したが、実際には特に関連はなかった。図を見ると、どの場合でも 0.1 から 0.3 の付近に位置していることがわかる。

各閾値の中央値は表 2.1 のとおりである。なお、最大 F 値の平均は 0.827 である。

これらの結果から、閾値を 0.2 と設定した。

表 2.1 各閾値の中央値

F 値が最大になる閾値	「許容値」を満たす最小閾値	「許容値」を満たす最大閾値
0.2	0.1	0.35

2.8 実際の結果

以上の手順を行い得られる例を表 2.2 に示す。なお、これは 2012 年 2 月 14 日~16 時に得られたものである。

表 2.2 結果の一例

代表的な文	福島第一原発:2 号機の温度、79.1 度に上昇 - 毎日 jp(毎日新聞)
話題語	号機、上昇、温度、冷温、停止、注水、容器、圧力、規定、保安、上限、読売、

	原子、炉
関連ツイート	2 号機、誤差が±20℃の温度計で計測してたらしいのに 79.1℃‥‥コンマ 1℃ってどういうアレ?(゜゜;)＼(--;)
	……え? 原発 2 号機やばくね…?
	本日 12 日現在、2 号機の温度、79.1 度に上昇。80 度を超えると冷温停止状態とは完全に言えなくなる。

また、話題を bot[8] @trend_words_jp[9](図 2.8)とサイト[10]（図 2.9）を通じて公開を行なっている。

図 2.8 話題情報を投稿する bot @trend_words_jp

[8] 自動的に発言を行う Twitter アカウント
[9] https://twitter.com/#!/trend_words_jp
[10] http://www.tkl.iis.u-tokyo.ac.jp/~fujikawa/trend/

図 2.9 トレンド情報を公開するサイト

なお、これらでは試験的に、「急上昇トレンド」（直近の 10 分間でのバーストスコアが閾値以上のもの）と話題占拠率（関連ツイートがバースト区間の全ツイート数に占める割合）を提示している。

2.9 評価

「何が話題になっているか」は客観的な定義が難しく、話題抽出を扱った研究でも、評価実験を行なっていないものも多い [5] [6] [11]。本研究では、不完全ながら以下のような評価を行った。

- 得られたクラスタが実際に話題になっているものか
- クラスタリングが適切か
- PageRank を用いた調整の適切さ

ここで、クラスタリングが適切かどうかを図るにあたっては、出力されたクラスタのうち同一のクラスタにもかかわらず別々のクラスタとなったものの数をかぞえることで行った。

また、PageRank を用いた調整の適切さは、150 を閾値とし、最初ランクが閾

値より下だったが調整によって閾値より上位に入った単語の数、そして、そのうち話題と関係ない不適切な単語の数を調べた。

2012 年の 2 月 8 日 22 時と 2 月 16 日 21 時における上位 10 とニュース系のクラスタを対象とした。

表 2.3 結果の一覧

実際に話題になっているか （正解数/全クラスタ数）	21/21		
クラスタリングの適切さ （不適切なクラスタ数/全クラスタ数）	9/21		
調整の適切さ	全単語数	調整により上位に入った単語数	
		適切	不適切
	104	31	5

表 2.3 に示すように、少なくとも話題になっていないものが上位に来ていることはなかった。次にクラスタリングの適切さだが、本来同じクラスタであるものが別々となっていたのはテレビ番組の場合が多かった。これは前述したように例えば歌番組では次々に別々の歌手が出てくるといった具合に、同じテレビ番組でも話題の細分化が起こるためと考えられる。これに関しては「どこまでをひとつのクラスタとみなすか」も含めて再検討する必要があると考えられる。

PageRank を用いた調整の結果では､例えば｢福島県川内村のミミズから検出 – 毎日 jp」というような話題で､「毎日」「jp」のような共起しているものの本来の話題と関連のない単語が上位に来てしまう場合がみられた。このように誤ってスコアを上げないため、多くの単語と共起が起こることが予測できるものはスコアを下げるなどの措置が必要と考えられる。

第3章　ユーザの反応の分類

トレンドの解析を行い、話題を発見した際、それが流言の可能性があるかどうかを判定するにはどうしたらよいか？　ユーザーが疑っているかどうかを分類する研究 [9]はすでに存在するが、これに対して、本研究ではユーザの反応を「信疑」に加えて「根拠の有無」という 2 つの基準で分類することを考える。

情報に対するユーザの反応を分類することには、次のような利点がある. まず、どの程度の割合の人がその情報を疑っているかを知ることができる。もし通常より有意に多くの人が疑っている情報なら，それは誤情報である可能性が示唆されるため、情報の真偽を判断するための材料の一つとなりうる．また，流言に対する反応の中には、それが誤りまたは真実であると判断する根拠を示している発言も多いため、そうした発言を集めることが出来れば、ユーザに情報の真偽を判断するための判断材料を提供することができる。

3.1　問題設定

本研究では、流言に対する Twitter 上での反応を、情報に対するユーザの態度（信疑）、その根拠の有無という二つの観点で分類するテキスト分類問題を考える。以下で、各タスクの具体的な内容について順に説明する。

まず、情報に対するユーザの態度（信疑）の分類基準について説明する。一口に疑いと言っても完全に疑ってるものから半信半疑、「本当ならひどい話です！」のように一応疑っているがかなり信用よりのものまで幅広い。しかし、ここでは、わずかでも疑っているものは疑いに分類する。なぜなら、細かく分類するとモデルが複雑になり、ラベル付の際にも判断に個人差が出てくるためである。また、根拠のはっきりした情報なら「本当なら」のように留保をつけることもないと思

図 3.1 Twitter の情報伝播の例

われるため、真偽不明の情報特有のツイートであると考えられるからである。

例えば有名人が死んだという流言が流れたとき、「○○（有名人）が死んだって本当？」「デマだろ」のような発言を"疑い"、「○○（有名人）が交通事故でなくなったそうです。ショック……」のように、まったく疑っておらず、普通のニュースと同様に受け取っている場合は"信用"と分類する。

次に、各発言におけるユーザの態度（信疑）に対する根拠の有無の判断基準について述べる。根拠としては、具体的な URL やツイートの引用など、その根拠の内容が明白なものに限らず、発言の文面からユーザが確信を持って判断する材料を持つと考えられる場合に根拠ありとする。例えば「○○が死んだって話、ジョークサイトの嘘記事ですよ。」という発言では、単にそれが誤りであると指摘するだけでなく、なぜ誤りと言えるかの根拠も示しているため、"根拠あり"と分類する。

非公式 RT の場合、"RT"以前の部分に意見が書かれているため、最後の RT 以前の部分のみを分類に使用する。図 3.1 の 2 番目、3 番目ともに、「RT @XX: Twitter が情報統制」の部分は使用しない。ここで、最後の RT 以前の部分を使用するのは、3 番目の発言ように疑っている意見を信じている場合でも、最初部分（「そうなんだ」）だけを取り出して「信用している」と誤って判断しないためである。

3.2　分類手法

我々は、テキスト分類問題で広く使われる分類器である SVM を、教師付きデータから学習し、各分類タスクに適用する [12]。流言に対するツイートを観察した結果、おおよそ次のような特徴が見られた。

- ユーザが情報を信じている場合は「ひどい」「感動した」などの感情的表現が含まれることが多い
- 流言であると指摘しているツイートは「流言」「ガセ」など直接的な言葉を使っているツイートが多い
- 根拠を示しているツイートは、何らかの事実を提示するため、「○○は××のはず」「××じゃない」など特定の言い回しのパターンが好んで用いられる

これらの特徴を考慮して、出現する単語や文体を素性とすれば分類が可能と考えられる。実際に用いた素性の詳細は次のとおりである。

- 単語の 1、2、3-gram
- 品詞（第一階層と第二階層）の n-gram
  - 例：“東京都が株主”
  - 第一階層： 名詞-助詞-名詞
  - 第二階層： 固有名詞-格助詞-名詞
- 文章長（140 で割り、1 以下となるように正規化）

ここで、品詞(POS)の n-gram を使用しているのは、品詞の連なりは文体を表すためである [13]。根拠を提示している文では、次のように話題によって提示される根拠も異なり、当然のことながら含まれる単語も異なる。

- 株主は東京都です。都知事じゃありません。
- 現在ではザイールという国は存在しません。

そのため、共通する要素である文体を捉えることが重要と考えられる。

例えば感情的な文なら、「ひどい！」「すごい！」→「形容詞-記号」、「許せない」→「動詞-助動詞」のようになるし、事実を提示する文では、「○○は××」→「名詞-助詞-名詞」という形になる。

ここで、より長い品詞の連なりを使用すればそれだけ文体をとらえやすくなるが、その場合素性の数が爆発的に増えてしまう。そこで、長さを 3 までに固定した固定長品詞 n-gram を使用する場合と、長さは固定しないが、 [13]で提案され

ているアルゴリズムを使用して、出現するドキュメントの割合が閾値(minsup)以上の可変長品詞 n-gram だけを使用する場合で実験を行った。

3.3 素性の重み付け

単語、品詞の素性の重みとして、tf 及び tf-idf を検討する。idf は、その単語が出現するツイートの数が小さいほど大きくなる関数形になっている。そのため、特定のツイートにしか登場しない単語の重要度を上げる一般語フィルタとしての役割を持つ。例えば新聞記事のカテゴリ分類では「は」「が」のようなどんな文章にも出現するような単語を使用してもあまり意味はなく、「株価」「粉飾決算」など特定のカテゴリ(ここでは経済)にしか出現しない単語のほうが分類にあたって有用であるため、これが用いられる。

3.4 信疑／根拠タグ付きデータの作成

学習・評価用データの作成には Streaming API を用いて集めたツイートと共に、Togetter（http://togetter.com/)というツイートまとめサイトを用いる。このサイトでは誰でも議論や様々な話題に関するツイートをまとめることができ、毎日 200 件以上のまとめが作られている。また、各まとめには何に関するものなのか分かりやすいようにタグ付けが行われている。

今回の実験では、様々なページの中で、「デマ」というタグが付いたページを使用する。多様なサンプルを得るため、流言に対する 10 人以上の反応をまとめたページを用い、手作業でラベルづけを行った。

なお、この時、「マジかよ」のような極端に短く、信じている場合にも半信半疑にも使用される、人間でも判断が難しい反応に関してはラベル付けを行わなかった。

対象とした特定の流言に関するまとめページは 35 で、ラベル付けを行ったのは 1986 ツイート、内 761 が「疑い」で、410 が「根拠あり」だった。

3.5 実験結果

3.5.1 実験設定

実験ではラベルづけしたデータのうち、「東京電力株主 5 位は石原氏」というトピックに関する流言をテストデータ、それ以外を訓練データとして用いた。
また、SVM 分類器の実装には liblinear[11]を用いた。SVM を用いて分類器の学習を行う場合、ソフトマージンパラメタ C の値が分類精度に大きく影響することが知られている。そこで、C を2^{-s}(s=−4, −2, 0, … 14, 16)の範囲で動かし、陽性(疑い、根拠あり)に対する F-measure が最も高くなる C を選んだ。また、可変長品詞 n-gram における閾値 minsup に関しても、0:05 から 0:30 まで 0:05 刻みで動かし、陽性に対する F-measure が最も高くなる minsup を選んだ。
比較のため､全ての発言を疑いまたは根拠ありと分類した場合を Baseline とする。

表 3.1 信疑ごとのツイート数

	疑い	信頼	合計
訓練データ	743	1210	1953
テストデータ	211	120	331

表 3.2 根拠の有無ごとのツイート数

	根拠あり	根拠なし	合計
訓練データ	340	1613	1953
テストデータ	145	186	331

3.6 分類精度

表 3.3 表 3.4 に分類実験の結果を示す。文体を表す品詞の n-gram を使用した場合、信疑分類の精度に大きな変化は見られない(表 3.3)が、根拠の有無の分類精度は大きく向上している(表 3.4)。これは前述したように、根拠を提示している文は共通の単語はあまりないが文体は共通しているためと考えられる。

[11] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

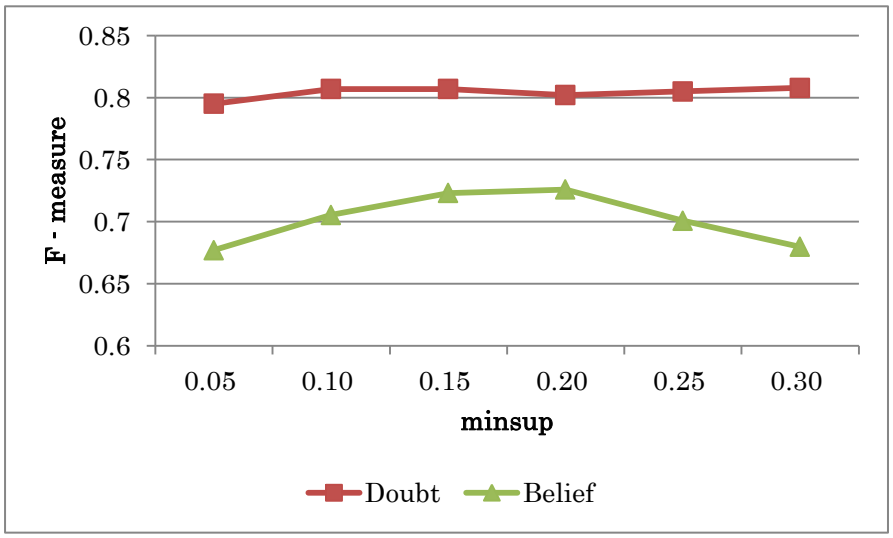

図 3.2 minsup と F 値の関係

しかし、tfidf に関しては、信疑の分類でもあまり変化がなく(表 3.3)、根拠の有無の分類では精度がかえって下がってしまった(表 3.4)。これは、前述したように同じカテゴリであっても使われている単語に多様性があり、出現割合の少ない単語だからといって重要なわけではないためと考えられる。

表 3.3 信疑の分類結果

	Baseline	tf			tf-idf		
		単語 n-gram のみ	+固定長品詞 n-gram	+可変長品詞 n-gram	単語 n-gram のみ	+固定長品詞 n-gram	+可変長品詞 n-gram
Accuracy	0.63	0.749	0.785	0.758	0.779	0.779	0.789
Precision	0.63	0.83	0.872	0.839	0.848	0.852	0.841
Recall	1	0.763	0.777	0.767	0.796	0.791	0.825
F-measure	0.773	0.795	0.822	0.801	0.822	0.82	0.833

次に、品詞の可変長 n-gram を用いた場合に minsup のみを変化させたときの結果を示す。図 3.2 は minsup と F-measure の関係を示す。また、minsup と素性数の関係を図 3.3 に示す。これらのグラフから、minsup の値を適切に選ぶことによって、少ない素性数で精度の向上が実現されていることが確認できた。図

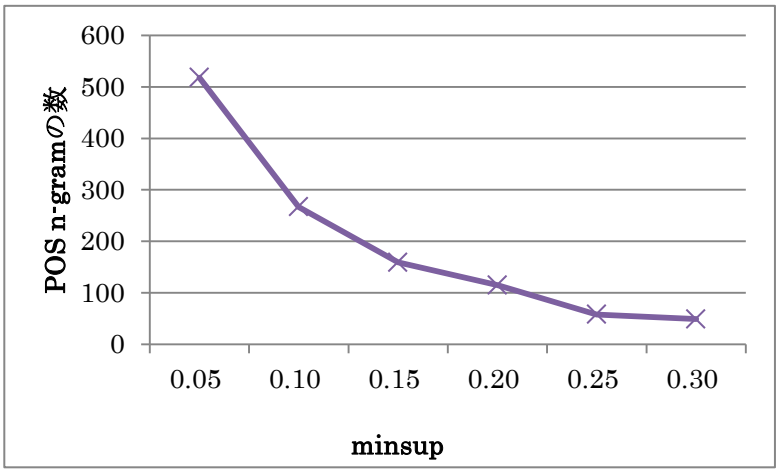

図 3.3 minsup と素性数の関係

3.4 図 3.5 は C と各指標の関係を示すが、C が一定以上になると指標にほとんど変化は見られなくなり、また、最大値との差も小さいため、C を変化させて指標との関係を調べることができない場合でも、C を一定以上の大きさにすれば最適値に近い結果が出ると予想できる。

表 3.4 根拠の有無の分類結果

	Baseline	tf			tf-idf		
		単語 n-gram のみ	+固定長品詞 n-gram	+可変長品詞 n-gram	単語 n-gram のみ	+固定長品詞 n-gram	+可変長品詞 n-gram
Accuracy	0.438	0.704	0.749	0.785	0.692	0.737	0.722
Precision	0.438	0.758	0.804	0.825	0.795	0.845	0.811
Recall	1	0.476	0.566	0.648	0.4	0.49	0.476
F-measure	0.609	0.585	0.664	0.726	0.532	0.62	0.6

3.7 議論

表 3.5 と表 3.6 に、素性の上位 10 件を示す。これらの表から次のことがわかる。まず、信疑分類に関する素性では、「デマ」「本当」「？」など直感的にも関係

があると思われる素性が並んでいる。

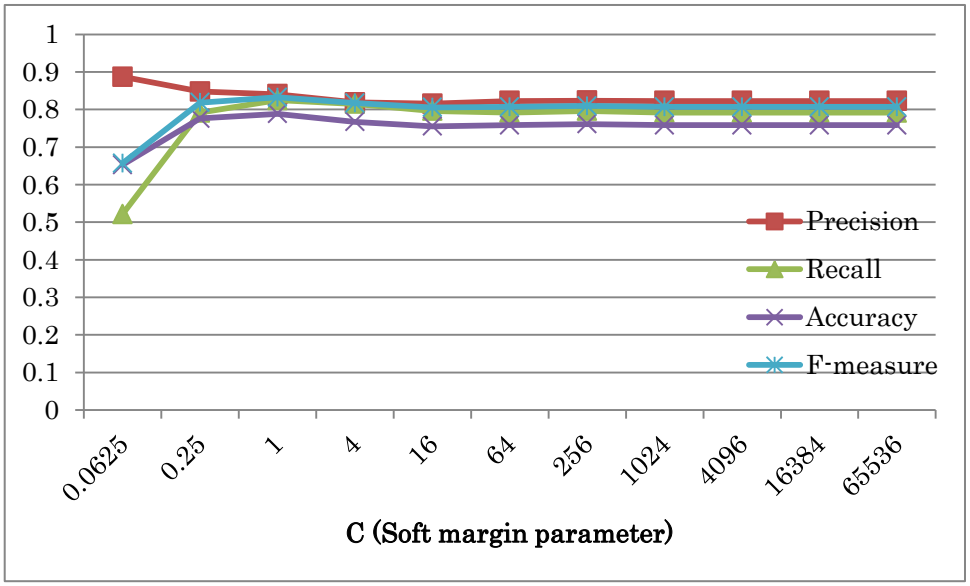

図 3.4 C と信疑の分類の精度の関係

しかし、根拠の有無の方に関しては、重複した素性や「コンゴ」のように特定のトピックとしか関連しない単語に大きな重みが与えられてしまっている。これは、根拠を示す文の場合、トピックによって出現する単語がまったく違うこと。さらに、あるトピックについて根拠を示す側はある単語を多用するが、そうでない側はほとんど使わない場合、その単語を根拠を示す際に使われる単語と分類器が誤認識してしまうためだと思われる。また、大きな重みを与えられた品詞 n-gram についても、その妥当性は直感的に判断が難しい。これに関しては今後さらに調査を進めていく予定である。表 3.7 表 3.8 に判定されたラベルと実際のラベルの数の関係を示した Confusion Matrix を示す。表 3.8 を見ると、根拠の有無の分類については、偽陰性が誤りの大きな要因となっていることが分かる。

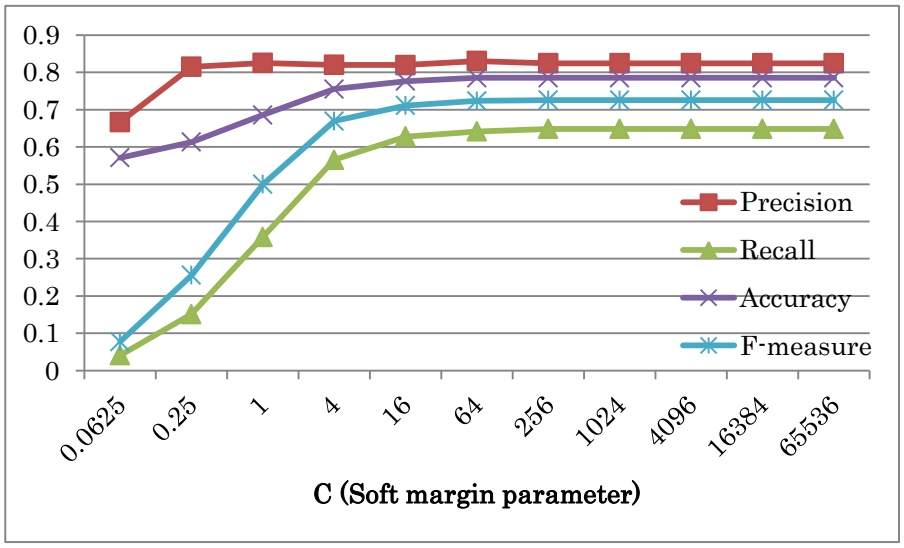

図 3.5 C と信疑の分類の精度の関係

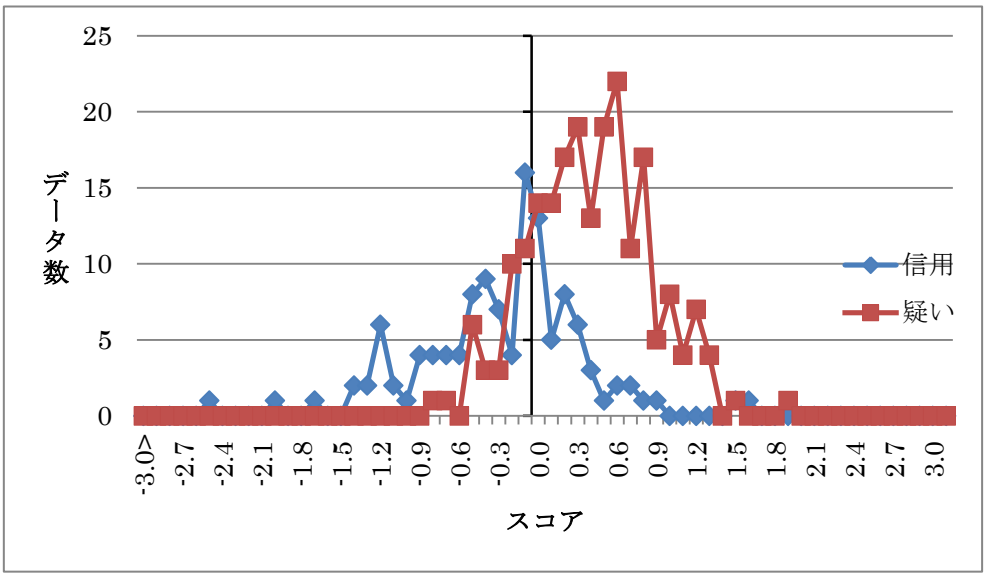

図 3.6 マージンとテストデータの分布(信疑タグ付きデータ)

そのため、今後は、根拠が提示されている文を正しく認識するための新しい素性の導入などに努めていきたい。

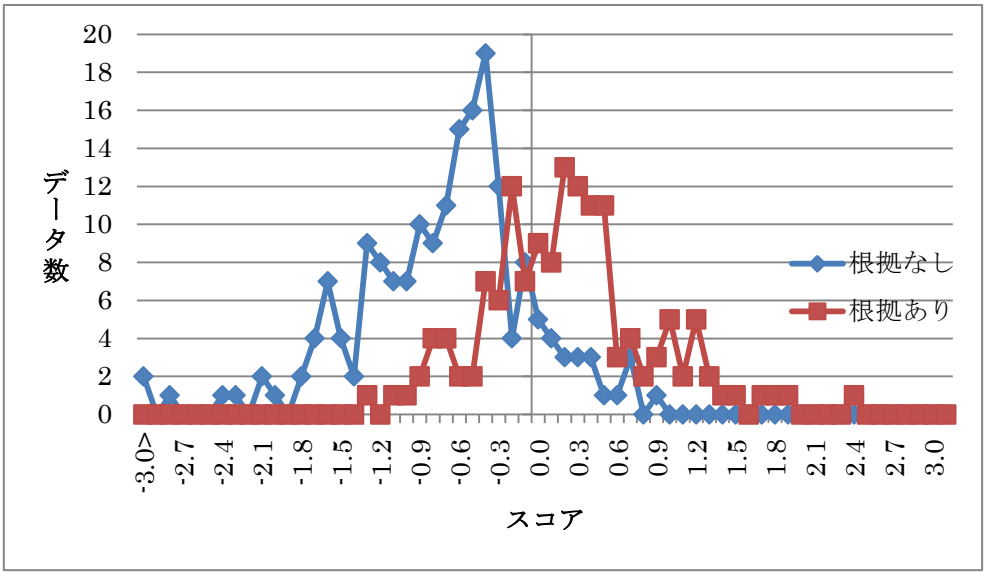

図 3.7 マージンとテストデータの分布(根拠の有無タグ付きデータ)

図 3.6 図 3.7 に、SVM を用いて分類を行ったときのマージンに対するスコア超平面からの距離とテストデータの分布を示す。信疑タグ付きデータ、根拠タグ付きデータともに、陽性のほうがピークが右よりになっており、スコアの大きさが実際のラベルと連動していることがわかる。

表 3.5 信疑の分類についての重みの大きい素性の一覧

	疑い		信用	
1	デマ	7.08	!	−2.68
2	ソース	3.96	中国	−2.61
3	本当	3	統制	−2.59
4	という	2.74	…	−2.57
5	確認	2.64	ね	−2.35
6	ガセ	2.5	新宿	−2.18
7	嘘	2.35	情報統制	−2.15
8	記事	2.06	事	−2.13
9	.	2.01	歴史	−2.1
10	ほんとに	1.99	日本	−2.04

表 3.6 根拠の有無の分類についての重みの大きい素性の一覧

	根拠あり		根拠なし	
1	/	4.27	:	−3.23
2	.	3.49	を	−3.17
3	係助詞_サ変接続名詞	3.47	一般名詞_一般名詞_一般名詞	−3
4	形容詞_助動詞_名詞	3.46	0	−2.96
5	コンゴ	3.28	!	−2.96
6	代名詞_名詞_一般名詞	3.15	トルコ	−2.75
7	ザイール	3.11	トルコが	−2.75
8	「	2.92	名詞_名詞_記号	−2.71
9	固有名詞_固有名詞	2.91	USERNAME:トルコ	−2.68
10	固有名詞_格助詞_自立動詞	2.89	う	−2.65

表3.7 信疑の分類の Confusion Matrix

	信用	疑い
「信用」と分類	96	47
「疑い」と分類	24	164

表3.8 根拠の有無の分類の Confusion Matrix

	根拠なし	根拠あり
「根拠なし」と分類	166	63
「根拠あり」と分類	20	82

第4章　おわりに

本論文では話題の解析のため、話題の抽出、関連投稿の抽出、フレーズ、代表的な文の抽出までを行う手法を提案した。この手法ではまず二項分布を仮定してその出現率が生じる確率が一定以下のものを抽出したあとPageRankを応用した手法でスコアを再計算、実際にバーストしている可能性のある単語を取得することができた。さらに、各投稿に含まれる単語とその単語のスコアから関連投稿かどうかの判定を行い、フレーズや代表的な文の抽出を行うことが可能となった。

さらにユーザの反応を流言かどうかの判断材料を与えるために関連投稿を「信疑の有無」「根拠の有無」で分類する手法を提案した。

今後の課題として、適切なクラスタリングの問題がある。似たような話題が同時に出現している時、例えば「松本復興大臣が知事を恫喝」「梶川ゆきこ氏が知事を批判」というニュースが流れている場合、「知事」という単語などが共通しており、また、後者が前者に対する反応であり、同じ話題であるとも見なせるため、同じクラスタだとみなされてしまう。こうした細かいクラスタリングをどのように行うかが課題である。また、「昼飯」「昼食」など似たような意味の単語や、テレビ番組名とそのハッシュタグのように、同じ物を意味する単語があった場合、同一のクラスタに入れるのが適切と考えられるが、「昼飯」という言葉を使う人は「昼食」という言葉を使わず、その逆も成り立つため、共起度だけを見ると別々のクラスタであるとみなされてしまう。そのため、どの単語とどの単語が似たような意味を持っているかを予め調べ、それを元にクラスタリングを行うなどの対策が必要である。

また、スコア、あるいは順位が一定以上のものを話題として抽出する場合など、パラメータの設定を恣意的でなく何らかの根拠に基づいたものにすることが望ましい。これらの問題を解決しより正確な話題分析を行うことが必要である。

参考文献

[1] (2011) 震災後のデマ 80 件を分類整理して見えてきたパニック時の社会心理. [Online]. http://news.livedoor.com/article/detail/5477882/

[2] 山本祐輔, "ウェブ情報の信憑性分析に関する研究," Kyoto University, PhD thesis 2011.

[3] 情報信頼性判断支援システム. [Online]. http://ici.wisdom-nict.jp/

[4] Jon Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 373–397, October 2003.

[5] Jeffrey Xu Yu, Philip S. Yu, Hongjun Lu Gabriel Pui Cheong Fung, "Parameter Free Bursty Events Detection in Text Streams," in *VLDB '05 Proceedings of the 31st international conference on Very large data bases* , Trondheim, Norway, 2005.

[6] Nick Koudas Nilesh Bansal, "Searching the Blogosphere," in *Proceedings of the 10th International Workshop on Web and Databases*, Beijing, China, 2007.

[7] Mark-Anthony Hutton, and Jugal Kalita Beaux Sharifi, "Summarizing Microblogs Automatically," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, California, 2010, pp. 685-688.

[8] Hikaru Yokono, and Manabu Okumura Hiroya Takamura, "Summarizing a document stream," in *ECIR'11 Proceedings of the 33rd European conference on Advances in information retrieval*, Springer-Verlag Berlin, Heidelberg, 2011 , pp. 177–188.

[9] Emily Rosengren, Dragomir R. Radev, Qiaozhu Mei Vahed Qazvinian, "Rumor has it: Identifying Misinformation in Microblogs," in *Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, July 27–31, 2011, pp. 1589–1599.

[10] Member, IEEE, and Shumeet Baluja, Member, IEEE Yushi Jing, "VisualRank: Applying PageRank to Large-Scale Image Search," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 11, no. 30, pp. 1877 - 1890, Nov. 2008.

[11] Jinfeng WEI Hong LI, "Netnews Bursty Hot Topic Detection Based ," in *International Conference on E-Business and E-Government*, 2010, pp. 1437-1440.

[12] P. Raghavan, and H. Schutze C.D. Manning, *Introduction to Information Retrieval*: Cambridge University Press, 2008.

[13] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, MIT, Massachusetts, USA, 9-11, Oct. 2010, pp. 207-217.

発表文献

[1] 藤川智英, 鍜治伸裕, 吉永直樹, 喜連川優. マイクロブログ上の流言に対するユーザの態度の分類. 電子情報通信学会データ工学研究会(PRMU DE). 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 111(77), 55-60, 2011 年 6 月..

[2] 藤川智英, 鍜治伸裕, 吉永直樹, 喜連川優. マイクロブログ上の中心的話題とそれに対するユーザの反応の抽出. 情報処理学会 第 74 回全国大会, (2012.3). （発表予定）

[3] 藤川智英, 鍜治伸裕, 吉永直樹, 喜連川優. マイクロブログ上の話題抽出とユーザの態度の分類に基づく流言検出支援システム . 第 4 回データ工学と情報マネジメントに関するフォーラム（第 10 回日本データベース学会年次大会）(DEIM 2012), (2012.3). （発表予定）