

Twitterにおける学術関連情報発信者の特徴分析と分類

筑波大学

図書館情報メディア研究科

2014年3月

荒川唯

## 目次

1. はじめに.....	1
1.1. Web 2.0 からソーシャルメディアへ.....	1
1.2. Twitter とは.....	1
2. 先行研究.....	4
3. データと手法.....	8
3.1. データ.....	8
3.2. 専門家アカウント調査.....	8
3.2.1. 専門家アカウント属性 .....	8
3.2.2. 専門家アカウントのツイート内容の分析 .....	8
3.2.3. ツイートのテキスト分析 .....	9
3.3. 専門家/非専門家分類実験概要.....	9
3.3.1. ランダムフォレスト機械学習.....	10
3.3.2. 評価.....	11
4. 結果と考察.....	11
4.1. アカウントの属性調査.....	11
4.2. アカウントのツイート調査.....	14
4.3. ツイートのテキスト分析.....	19
4.4 専門家/ 非専門家アカウント自動分類実験.....	25
5. おわりに .....	27

# 1. はじめに

## 1.1. Web2.0 からソーシャルメディアへ

日本におけるインターネットの人口普及率は2012年末、79.5% (総務省, 2013) となり、多くの人々がインターネットを利用している。情報技術の進展とともに、インターネットを用いたサービスは多様化してきた。2005年に登場したWeb2.0はアメリカのティム・オライリーが『What is Web2.0』で提唱した言葉である。Web2.0以前 (Web1.0) におけるウェブはマスメディア的であり、情報・コミュニケーションの流れは情報発信者から受信者への一方通行が中心のメディアであった。Web2.0では受信者は情報を受信するだけでなく、情報発信者にもはたらきかけ参加できる、個人メディア中心へと移り変わった。当時、利用者を急激に増やしていたソーシャルネットワーキングサービス (SNS) やブログはそれを象徴するものであった。インタラクティブなコミュニケーションのソーシャルメディアも個人が情報発信者になりうるという点で流れを汲んでいるといえよう。

ソーシャルメディアはSNS (Facebook, mixi, google+, LinkedIn など) やブログ、マイクロブログ (Twitter など)、ソーシャルブックマーク (SMB)、RSSといった、利用者によるインタラクティブな活動ができるメディアとして注目を集めた。本研究ではソーシャルメディアの中でもTwitterを対象としている。

今となつては当たり前になっているソーシャルメディアだが、数年前まではマスメディアを通さず政治家や有名人の考えを知ったり、話しかけたりはできなかった。Web2.0の波が起き、ブログが広がり、現在においてはTwitterをはじめとしたソーシャルメディア上で直接一般の人と著名人がコミュニケーションを交わせる。ソーシャルメディアは人と人との距離を狭められるメディアであると考えられる。本研究ではそんなソーシャルメディアの中でも、140字以内で記事を投稿するシンプルな機能を持つTwitterを対象とする。

## 1.2. Twitter とは

Twitterはマイクロブログまたはミニブログと呼ばれるウェブサービスの一種である。2006年6月にアメリカ、カルフォルニア州サンフランシスコの会社、Obvious社 (現Twitter社) によって始められたサービスである。日本語版の公開は2008年、携帯電話向けのサイト開設は2009年6月であり、日本においては2009年後半からユーザー数が急激に増加した。2012年末には、月間アクティブユーザー数が2億人を超えた<sup>1</sup>と発表され代表的なマイクロブログであると言える。

1件の記事 (ツイート) に対し140字までしか投稿できないという制限がある。また、自分が読みたい人のツイートはフォローと言う形で登録すれば時系列を追って自分のTwitter

---

<sup>1</sup> <https://twitter.com/TwitterJP/status/281052580849778688>

ページで読むことができる。Twitter の利点として、フォローに相互認証を必要としないため、興味、関心が高い相手をフォローできる特性がある。そのため興味に基づくソーシャルネットワークが形成されることが多いと言える。シンプルでリアルタイムな情報のやりとりがなされるため、速報性や拡散力が高い特徴も備えている。一方で一時的な絆が生まれやすいプラットフォームとも言える。利用者はプライベートで利用する一般の人々はもちろんのこと、企業や行政など団体も利用している。

2010年にはイスラム政権の連鎖的な民主化運動（ジャスミン革命）で、ソーシャルメディアを用いたコミュニケーションが世界的に注目された。2011年3月11日に発生した東日本大震災では、新たな情報収集ツールの1つとして使われた。個人が情報を交換するだけでなく、これを契機に開設した団体アカウントもある。例えば、東京電力や首相官邸は震災情報や原子力発電所の事故に関する情報を発信するために利用を開始している。震災以前にアカウントを開設していたアカウントでも、震災前後で変化が見られた。被災地域自治体や周辺地域自治体のアカウントにおいて3月11日以降にフォロワー数が急増していることが分かっている（総務省、2011）。こうした公式アカウントでは災害情報やライフラインに関する情報発信がなされた。個人としては人探しや救援、救援ボランティアについての情報、震災の恐怖といった心的ストレスの緩和などマスメディアでは補いきれない情報がやりとりされた。また地域の食料や生活用品の店舗在庫状況の情報といった局所的な情報のやりとりもなされた。しかし、有用性が注目された一方、「コスモ石油の爆発により有害物質が雲などに付着し、雨などといっしょに降る（ママ）」といったデマや流言が広がった負の側面も指摘されている（総務省、2011；立入、2011；荻上、2011）。

このように Twitter は有益な情報を収集するためのツールにも、誤情報の発信・受信ツールにもなりうる。膨大な量のツイートから必要な情報を取り出すことは、緊急時はもちろん、普段の生活の中でも重要である。Twitter を用いた情報収集は一定数のユーザーで行われていることが確認されており、SNS などほかのソーシャルメディアと比較してもその割合は高いことが報告されている（インターネット協会；2012、トライバルメディアハウス、クロス・マーケティング；2012）。しかし、ユーザー数からも分かる通り、Twitter ではあらゆるユーザーのツイートがリアルタイムで更新されている。意識的に活用すれば、有益な情報を入手にすることができる。どの情報が重要で、どの情報が重要ではないのか、すばやく判断するのは難しい。そのため前もってフォローする候補であるユーザーの特徴を知り、フォローするか否かを考える必要がある。Twitter は属人性が強いメディアであるため、コミュニティに属すると考えられるアカウント、そのアカウントの投稿するツイートの特徴を知り探し出す必要がある。

特定の情報の1つとして学術関連情報があげられる。青山ら（2011）の行った学術とソーシャルメディアの利用動向のアンケート調査がある。アンケートは日本化学学会に訪れた人々が対象であり、質問は Twitter やソーシャルブックマーク（SBM）などのソーシャルメディア（論文内では新コミュニケーションメディアと定義されている）の利用に関するこ

と、学術関連情報で用いるときの受発信の意識についてである。

この調査から、意識的に Twitter を学術情報発信に用いているアカウントは少ないことが分かる。しかし、学会発表関連ツイートを分析した研究 (Weller et al., 2011; Letierce et la, 2010) や論文の URL 含むツイートを分析するなどが存在していることから、意識、無意識にかかわらず学術関連の情報を投稿している可能性はある。それは Twitter が人の興味に左右されることが多いことから十分に考えられる。また、この調査では分野は化学 (自然科学分野) のみに限ることから、分野を広げ実際のツイート、ユーザーの特徴を分析することは重要であると考えられる。

そこで、本研究は学術関連情報を多く含むユーザーの特徴を捉え、それに基づいたユーザーの自動分類実験を行った。自動分類実験では用いる特徴量の種類やその組み合わせによって結果が左右される。アカウントの持つ特徴を考慮した特徴量を加えることによってより高い性能の自動分類が期待できる。学術情報を多く発信するアカウントのフォローをすることで、継続的な情報収集手段として Twitter を用いることができると考える。

本論文は第 1 章で Twitter のデータの利用に関する概観を行い、第 2 章では先行研究を踏まえて本研究の立ち位置を確認する。第 3 章では本研究で用いるデータと手法に関する説明を記述し、第 4 章は行った各分析、実験について報告し、結果を踏まえて研究の考察を行う。第 5 章では研究全体のまとめと今後の展望について述べる。

## 2. 先行研究

Twitter に関する研究は様々な分野, 様々な観点から行われている. ここでは, 本研究に  
関係の深い, ユーザーに関する研究, Twitter における科学コミュニケーションの研究を取  
り挙げたい.

Twitter の情報を有効活用する論文としては, ツイートで未来予測を試みた, 豚インフル  
エンザの流行予測や株の予測を行った Ritterman et al. (2009), インフルエンザ流行予測を複  
数の機械学習手法を比較しながら行った荒牧ら (2011), 大統領選挙に関する予測を行った  
Tumasjan et al. (2010), サッカーの実況ツイートを用いて試合の自動要約を試みた高村ら  
(2010)などがある.

Twitter のデータ活用として, Twitter のデータのみではなく, そのほかのデータと組み合  
わせて利用している場合もあり, 早川ら (2011) では Yahoo!ニュースと Twitter のフォロー  
関係をデータとして用いている. 研究の目的はユーザーの興味・関心の高いニュース記事  
を Twitter のフォロー関係を用いて推薦することとしている. 指標としては Yahoo!ニュース  
とフォローしている友人のツイートを用いて友人のツイートと記事の関連度を算出して  
いる. 友人にはユーザーがあらかじめ重要度を設定し, 重みとして用いていた. 評価は 3 人  
のユーザーによるユーザスタディが行われており, 平均クリック数をカウントし比較する  
とユーザー自身のツイートを用いた場合より研究で作成したシステムによる推薦の方が高  
い値が得られたことが示されていた. 筆者は重要度が高い友人がユーザーの興味と近いこ  
と, 友人たちの中で局所的に盛り上がっている話題を推薦できると考察し, ユーザーのツ  
イートを用いた場合より, 重要度の高い友人のツイートを用いた方がユーザーにとって有  
用な記事を推薦できることが分かった.

山口ら (2012) は Web ページがどの分野の専門性をもつユーザーに共有されたかと言う  
指標をもって, Web ページの推薦手法の提案を目的とする. 指標によって Web ページのト  
ピックや重要度を推定し用いた. 予備調査では URL 含有率は 14% (収集期間 2011 年 12 月  
13 日-14 日). 評価実験データは 2011 年 12 月 15 日 14 時から 18 日 14 時までの 3 日間であ  
る. 限られたページが頻繁に共有されている. 提案手法ではどのような専門性を持つユー  
ザーをフォローしているかによってユーザーの興味を判定がなされている. Web ページが  
どんな専門性を持つユーザーに共有されたかで重要度を算出していた. ユーザーの興味と  
Web ページのトピックの類似度によって推薦するか否かの決定をくだしていた. ユーザー  
の興味判定にはユーザーリストを用い, 専門性の判定には特定の語が含まれるリストに属  
する数に基づいている. ユーザーの興味はフォローユーザーと特定の語が含まれるリスト  
に基づかれていた. 推薦スコアについては, ユーザーの興味スコアと専門性スコアのそれ  
ぞれ上位を候補ユーザー集合から取り出し, 和集合を作成する. 評価は被験者 (12 人) によ  
る評価実験にて, 3 つの手法によって各々 Web ページの推薦を行い, 評価を行った. 手法は  
提案手法, フォローに基づく手法, マイクロブログ上でどれほどその Web ページが共有さ

れるかを用いた手法である。有名ユーザーをフォローしている数が多いグループでは提案手法が高い評価を受け、有名ユーザーのフォロー数が少ないグループはフォローの方が提案手法を上回る結果となった。提案手法はユーザーリストの登録から専門性を算出するためこのような結果になったと考えられる。

Twitter のユーザーに着目した研究には Golder ら (2009) の類似性による接合に基づいて、ネットワーク構造に着目したユーザー推薦を提案や後藤ら (2010) の意外な情報提供者の推薦、白木原ら (2010) の情報先取り発信者の研究などがある。

後藤らは (2010) はソーシャルメディアにおける意外な情報提供者の推薦手法の提案を目的とし、Twitter を対象に推薦システムを実装し、検証を行っている。具体的には返答、引用、お気に入りと言うユーザーの行動の表出を元に推薦システムを構築した。人のつながりについては先行研究 (Golder et al, 2009) と同様、二次ネットワークまでを用いた。検証には Twitter 公式など他のシステムを含む 3 つのシステムの評価を被験者実験 (15 人) に行った。評価はアカウントの既知、未知、好奇心の刺激の有無を問うた。結果、どのシステムにおいても、未知のアカウントより既知のアカウントの方が高い割合を占めた。しかし、未知で刺激的なアカウントは提案システムが最も高い値を示した。また、今回の研究においては引用が最も有用である可能性があることが示唆された。

白木原ら (2010) は Twitter ユーザーの中でも流行語に敏感な人 (trendspotter) を検出することを目的とし、システムを作成、検証を行った。検出にはバースト検出のアルゴリズムを用い、流行語の選定には Twitter 上で一定時間内に多く投稿された語を検出する buzztter<sup>2</sup> から行っている。提案システムを用いた実験は Twitter におけるバースト検出の妥当性の検証、trendspotter の検出、検出した trendspotter の有効性の確認を行っている。バースト検出の有効性については、実験に用いた 3 語において有効にバーストの検出がなされ、Twitter の検出においても有用であることが示された。trendspotter の有効性については提案システムで検証が得られた。trendspotter の検出では、流行語を含む投稿を複数回行っているユーザーを検出することで想定通りの trendspotter を検出できた。検出されたユーザーの傾向として、一時的に流行語を投稿するだけでなく、非 trendspotter よりも他の流行語も含む投稿を行う傾向があることが明らかになった。

田中ら (2010) はツイートの有用な情報の分類の提案を目的として、ツイート投稿者の指向 (一般への情報発信/コミュニティ内コミュニケーション) によってユーザー分類および、ツイートも分類を判定した。判定は 2 種類行われ、1 つはフォロー関係を考慮した手法、他方はお気に入りのつけられた (被お気に入り) 割合を用いた手法である。友人知人のフォロー関係を取り除いた、ネットワークを利用している。HITS アルゴリズムを適用している。

Twitter のデータを用いた科学コミュニケーションの研究としては Letierce et al. (2010) や Weller et al. (2011) がある。Weller et al. (2011) 研究の目的は Twitter における引用や参照について、Twitter を用いた科学コミュニケーションを指標の提供を目指している。この研究は

---

<sup>2</sup> <http://buzztter.com/>

研究者 (科学者) の Twitter 利用と伝統的な科学コミュニケーションがマイクロブログのコミュニケーションにどう表れているか、もしくは新しい行為が表れているかを調査する初歩的な研究である。2つのタイプのデータセットを利用しており、1つ目は人をベースとしたデータで Bradley の研究者リストと人で手作業で認識された研究者のアカウントを利用した。大学教員のみでなく、大学院生、企業研究者を含む。本研究ではアカウントを個人ではなく、研究者コミュニティとして捉えている。もう1つのデータは学会時のツイートデータで、4つの学会発表 (WWW2010, MAL2009 (Modern Language) Association Conference) の開催期間中にそれぞれ学会のハッシュタグがつけられたツイートを収集したものである。学会発表の選定基準は参加者の多さと専門領域を基になされていた。行われた分析は、ウェブサイトの URL リンクのタイプ、学会関連のデータセット外のツイートからのリツイート回数、高頻度リツイートユーザー、被リツイートユーザーについてである。Twitter における引用行動は従来の科学的流通や引用行動と全く持って異なり、特有の分析や指標の確立が必要であるとしている。特に科学的な情報を含むツイート内で行われる引用に着目しており、論文への直リンク、目的のソースとツイートの仲介ページ (RT) を分析していた。

Letierce et al. (2010) の研究では科学的な内容を含む情報が Twitter でどのように広がっていくかを明らかにすることを目的としている。3つの学会に関するハッシュタグを含むツイートを抽出することで、研究者がつぶやいたと考えられるツイート内容などを調べた。まず、ウェブ上のメディアについて対象に調査を行い、Twitter がもっとも選ばれているメディアと示された。そして、Twitter 利用に関して複数の観点から調査を行った。Twitter が学会の科学コミュニケーションで利用されていることは、複数の研究でも確認されているが、ここでは研究者が情報を広げるために、どのようにして Twitter を利用しているかにフォーカスをあてている。分析は会議中におけるキーイベントと出席者について、タグとハイパーリンクがイベントの間はどう用いられているかについてである。データは3つの会議の公式ハッシュタグが付与されたツイートをクロールし収集しており、ネットワーク分析には HITS アルゴリズムを利用し、発言力やハブとなるユーザーについて分析した。なお、対象は大学院生や CEO など大学における研究者のみに限定していない。

Twitter は正解中のオーディエンスに情報を届ける可能性を持っているが、現状では特定のコミュニティ内を出ていない、ツイート量は会議で行われている内容に応じて量が変わることなどが分かった。ハブと authority の両方を持っている人もいれば、authority のみ高い人もいる。URL は論文などの文書の他にスライドなど周辺のものも含んでいる。#は何が公式なのか分からなくつけていない人もいる。この論文の中では Twitter は世界中のオーディエンスに科学的なメッセージを広げるのに妥当なサービスである可能性を述べていた。

Uysal (2011) はリツイート (RT) に基づいたユーザー推薦の手法の提案を行っている。実験には決定木 (J49) のアルゴリズムを用いて学習を行った。ランクづけには Coordinete Ascent (CA) algorithm を用いた。特徴量は大きく分けて4種類あり、Author-based (投稿者本人)、Tweet-based (ツイート情報)、Content-based、User-Based (投稿者と関係する人の情報)

であった。本研究では情報過多の問題から情報を掬い出す2つの手法を提案した。4つのグループを設け **coordinate ascent learning** を用いて訓練を行った。

このように、関連研究がある中で本研究は学術関連情報を発信する可能性のあるユーザーの特徴およびツイートの特徴をイベント如何に関わらず分析し、Twitter ユーザーの研究としては特定の情報発信者の特徴を分析し、ユーザーのフォローの支援に向けた基礎的な研究に位置すると考えている。

## 3. データと手法

### 3.1. データ

本研究で用いたデータには2種類あり，専門家アカウントと非専門家アカウントおよびツイートである．専門家アカウントの収集は，まず Twitter のユーザー検索を用い，「大学教員」をキーワードとし検索した．検索結果，上位 119 アカウント（当初 120 アカウントを収集したが，ツイート数が 0 であったため除外）を対象とし，2012 年 7 月 3 日 11,485 ツイート（約 97.29 ツイート/1 アカウント）取得した．非専門家アカウントは，2012 年 8 月 19 日にパブリックタイムラインより無作為に 119 アカウントを抽出し，合計 11,900 ツイート（100 ツイート/1 アカウント）取得した．なお，非専門家アカウントは分類実験の比較アカウントとして用いた．専門家アカウントの属性の調査，対象アカウントの相対出現頻度といったツイートのテキスト分析，専門家非専門家の分類実験を行った．

### 3.2. 専門家アカウント調査

アカウントの特徴をさぐるためにアカウントの属性やツイートの内容の分析を行った．

#### 3.2.1 専門アカウント属性

属性については，プロフィールなど Twitter アカウントから得られる情報を始点として行った．実名アカウントの場合，本務校としている大学のサイトや Read&Researchmap<sup>3</sup>（要資料参照）といった研究者データベース，科研費データベース（KAKEN<sup>4</sup>）などインターネット資源も利用し集計に加えた．集計項目は実名/匿名，専門分野，所属，職務（職名・職階）である．専門分野は科学研究費で用いられている系・分野・細目表<sup>5</sup>を基に分類を行い集計した．本研究で利用した分類区分は分野である．

#### 3.2.2 専門家アカウントツイート内容の分析

専門家アカウントのツイートに関しては専門情報が含まれている可能性が高い分類として，研究，教育のカテゴリを設けた．明示的に分かる情報についてはこの2つのカテゴリにカウントし，曖昧なものも含めてそのほかのツイートはその他として集計した．曖昧なものと言うのは例えば，政治を専門としているアカウントが政治の時事について感想や意見を述べているといったものである．また，資料の利用可能性を考え，論文に関するツイートの集計も同時に行った．

---

<sup>3</sup> <http://researchmap.jp/>

<sup>4</sup> <http://kaken.nii.ac.jp/>

<sup>5</sup> 日本学術振興会．“別表系・分野・分科・細目表（1）平成 23 年度科学研究費補助金系・分野・分科・細目表”．

[http://www.jsps.go.jp/j-grantsinaid/03\\_keikaku/data/h23/download/j/05.pdf](http://www.jsps.go.jp/j-grantsinaid/03_keikaku/data/h23/download/j/05.pdf)

### 3.2.3 ツイートのテキスト分析

ツイートのテキスト分析については、語の相対出現頻度(Bag-of-words, BOW), 品詞構成比率, Twitter の特有の記号 (@, RT, QT, #), URL の出現率, @におけるフォロー/フォロワー関係を分析した。形態素解析には MeCab<sup>6</sup>を用いた。これらの項目については専門家, 非専門家, データ全体の数値の比較を行った。

### 3.3. 専門家/非専門家分類実験概要

本実験は専門家クラス, 非専門家クラスの2クラスの分類実験を行った。実験にはRのランダムフォレストのパッケージを用いて行った。与える特徴量のタイプを変化させ, 全実験に用いた特徴量は大きく分けて, 従来から用いられている (1) テキストの特徴量, Twitter の分析であることを考慮した (2) Twitter の特徴量の2つである。(1) テキストの特徴量は単語の相対出現頻度, 1 ツイートあたりの文字数である。(2) Twitter の特徴量は Twitter 特有の記号 (@, RT, QT, #), URL の出現率, リプライ (@) におけるフォロー/フォロワー関係である。@の関係とは, 該当アカウントと@先のアカウントのつながりを指しており, 相互フォロー, フォローのみ, フォロワーのみ, 双方ともにフォローしていない状態に分けられる。

#### 3.3.1. ランダムフォレスト機械学習法

ランダムフォレストとは Breiman (2001) が考案した手法であり, ブートストラップサンプリングを用いた決定木による集団学習の1つである。ランダムで選択された変数を用いて決定木を複数作成し, 得られた結果より多数決で判定を行う。具体的な手順は以下の通りである。まず  $i$  行  $j$  列特徴量行列を作成する。データの中から重複を許し, 行に対してをブートストラップサンプリングを行う, 変数には  $j$  個の中から重複なしで  $\sqrt{j}$  個をランダムに選択し用いて, 決定木を成長させ分類する。作成したサンプルより  $2/3$  は学習データとして分類器を構築, 残りを評価データとする。これを Out-of-Bag (OOB) テストと呼ばれ, 評価に用いたデータは OOB データと呼ばれる。ランダムフォレストでは分類における重要度 ( $VI_{acu}$ ) が計算され, 分類に影響を与える程度が数値計算される。 $VI_{acu}$  は OOB データにおける正解数 ( $C_{oob}$ ) と変数をランダムに置換した場合の正解数 ( $C_{per}$ ) の差の平均を標準誤差で割った値である。

$$VI_{acu} = \frac{\text{mean}(C_{oob} - C_{per})}{s.e.}$$

この手法の利点は上記の重要度が算出できることのほかに, 欠損のあるデータでも一定の性能を保つことができることや, 変数が多いデータでも効率的に処理を行えること, 過学習しにくいといったことが挙げられている (鈴木, 2010; 金, 2007)。

<sup>6</sup> <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

### 3.3.2. 評価

分類実験結果の評価には精度 ( $P_i$ ) , 再現率 ( $RR_i$ ) ,  $F_\beta$  値を用いる. 精度とは分類結果にどれだけ「ゴミ」がないかを示し, 正であると判定された分類結果のうち, 実際のクラスで正であったものの割合を示す. 再現率とは分類結果にどれだけ「漏れ」がないかを示す指標であり, 分類結果が実際のクラスの正の集合をどれほど反映できたかを表している.  $F_1$  値とは再現率と精度の調和平均であり, この値が大きいと性能が良いと評価する. それぞれの式は以下のように定義されている (徳永, 1999; 高村, 2010) .

$$P_i = \frac{a_i}{a_i + b_i}$$

$$RR_i = \frac{a_i}{a_i + c_i}$$

$$F_\beta = \frac{(\beta^2 + 1) \times P_i \times RR_i}{\beta^2 \times P_i + RR_i}$$

表 1. 交差行列

		分類結果	
		True	False
実際のクラス	True	$a_i$	$c_i$
	False	$b_i$	$d_i$

## 4. 結果と考察

### 4.1. アカウント属性調査

専門家アカウント 119 アカウントに関するアカウント属性調査の集計結果を述べていく。対象とした専門家のアカウント名が実名であるか、匿名であるかについては、97%が実名あり、3%が匿名であった。これはアカウントの選定の方法による影響と考えられる。Twitter のユーザー検索ではプロフィールに明示的に検索キーワードが記されていないとも、関係する単語が含まれているアカウントが上位に表示される場合も多々ある。

表 2 はアカウントの専門分野に関する集計結果である。アカウントによっては複数の分野を研究フィールドとしている場合がある。そのため、専門分野の集計については、複数の分野を専門分野としている場合はそれぞれの分野をカウントし、重複ありの集計となっている。不明は匿名のためアカウントのプロフィール以外で情報を得られなかったものである。なお、専門分野の集計に関しては、ユーザー名は匿名ではあるが、プロフィールに分野の記載があったアカウントがあったため集計でカウントした。そのため、専門分野のみ不明の数が他の集計と数値が異なる。専門分野の分布は社会科学が 0.52 と最も多く、全体の半分以上を占める結果となった。続いて、総合領域 (0.20) , 人文科学 (0.12) , 数物系科学 (0.07) , 工学 (0.04) , 農学 (0.02) という順となった。そして今回のデータに含まれていた分野のうち、最も少なかったのは生物学分野であった。その他、系・分野・細目表に記載はあるがこの表に記載がない分野は今回のデータには含まれていない分野である。1 番多かった社会科学の細目としては政治学、経済学、経営学、法学などが含まれていた。次点の総合領域の細目は、情報学や教育工学、科学教育、脳神経科学などがここに含まれていた。

表3は所属の分布における集計の結果である。所属先は私立が1番多く0.52あった、次いで国立大学(0.35)、公立(0.03)の順となった。その他は大学以外の研究所所属の研究員を本務としている場合などがここにカウントされており、大学では講義を受け持っていたり、共同研究者として席を置いていたりしていることが確認されている。先に述べた所属の分布は、割合は違えども、人数の多さの順は実際の分布と同じ順を示している。文部科学省が行っている学校基本調査によると平成25年度調査では私立大学教員が最も多く6割弱を占めており、次いで国立大学教員が3割強、公立大学教員は1割弱となっている(文部科学省, 2013)。この調査は本務校として大学に勤めている教員が対象であり、非常勤や特任教授として勤めている教員の人数が入っていない、大学のみで研究機関は含まれていないなど、項目の立て方の違いはあるが、教員数は大学数に起因すると考えられるため、おおよそ同じ結果となったと言えるだろう。

表4は職階に関する集計である。1番多い割合を占めているのは教授であり0.51で半分を超える結果となった。次いで多いのは准教授であり0.21である。この2つの項目だけ合わせると7割を超えることとなった。そのあとは非常勤/特任(0.11)、その他(0.06)講師(0.04)、その後は助教授(0.02)、とつづき、助教は0.01となった。助教授は日本では2007年4月以降、廃止となった職階であるが、ここで集計されている助教授と言うのは海外の大学に所属しているアカウントを示している。その他の内約については、サーチャーや主任研究員が含まれている。

表2. 専門分野

分野	度数	割合
工学	5	0.04
社会科学	66	0.52
人文科学	15	0.12
数物系科学	9	0.07
生物学	1	0.01
総合領域	25	0.20
農学	2	0.02
不明	3	0.02

表 3. 所属の分布

所属	度数	割合
国立	42	0.35
公立	4	0.03
私立	62	0.52
公立(海外)	2	0.02
その他	4	0.03
不明	5	0.04

表 4. 職階の分布

	度数	割合
教授	62	0.51
准教授	25	0.21
講師	5	0.04
助教	1	0.01
助教授	2	0.02
非常勤/特任	13	0.11
その他	7	0.06
不明	4	0.03

## 4.2. アカウントツイート調査

専門情報が含まれると考えられるツイートを計量した結果を示した (表 5) . 学術関連情報を明確に含むツイートの割合は研究 0.01, 教育 0.03 という結果となった. 多くを占めるその他には, 明確には学術関連情報とは言い難いが, 専門情報に近い一般的なニュースに関するツイートといった不明確ではあるが専門としている分野に関することも含まれている.

アカウントのツイートの内容を個々に見ていくと, どの分野においてもニュースに関連するツイートが多いことが分かった. ニュースはソースとして URL が付記されている場合がよく見られた. ニュースにしても, 時事であるだけでなく, アカウントの専門分野に近い場合があり, 特に政治学や経済学といった社会科学分野ではその内容について本人の見解や意見などコメントが付記されていた.

表 5. ツイートの内容の分析結果

	度数	割合
その他	10,990	0.96
研究	169	0.01
教育	326	0.03

表 6. アカウントごとのツイート内容の特徴一覧

アカウント	内容の特徴
総合領域 1	AKB 論や論壇についてのコメント.
総合領域 2	大学の予算や学問について. 震災後の安全や反原発を呼びかけるなど.
総合領域 3	科学コミュニケーションについて. 主に, それに関する学会や学内外向けのシンポジウムの告知. そのほか, 科学に関するニュースを少々.
総合領域 4	放射能についてデータをもとに論じる.
総合領域 5	教育 (学習指導要領: 数学や音楽) と放射能について.
総合領域 6	@が多い. 自分の意見について連続ツイートを行うことがある.
総合領域 7	科学技術について.
総合領域 8	スポーツ科学の研究や時事について.
総合領域 9	特に脈略なく日常について.
総合領域 10	時事ニュース (IT, 情報関連が多い), シンポジウムの告知.
総合領域 11	日常について, 学生との@が多い.
総合領域 12	学会中継, 個人情報 (※アカウントの研究領域) について.
総合領域 13	授業へのコメント質問への応答が多い. ハッシュタグを利用している.
総合領域 14	スパコンなど情報系, ポスドク問題.

総合領域 15	研究のことや日常 (食べ物) について.
総合領域 16	日常, 教育, 授業のお知らせ.
総合領域 17	日常, @, 雑誌宣伝.
総合領域 18	科学, blog 更新通知, 本務校について.
総合領域 19	学内のシンポジウムの告知. 国際会議について. 写真多い. 英語その他外国語でツイートすることあり.
総合領域 20	論文, 書評紹介.
総合領域 21	教育や学習について, それに関する時事は URL つきでツイート多し.
総合領域/社会科学	情報関連を主として, ニュースやネット記事についてコメント.
人文科学 1	特定の内容はつぶやいていない.
人文科学 2	マンガ学. マンガ関連のツイートあり.
人文科学 3	ジェンダー論, 原発問題. @が多い.
人文科学 4	@が多い.
人文科学 5	政治的な時事について.
人文科学 6	マンガやアニメ文化について. シンポジウムなどの告知もあり.
人文科学 7	政治や言論について.
人文科学 8	スポーツや文化について.
人文科学 9	政策を公共の観点でツイートしている, またその学問についての問いに答えている. 本の紹介が多い. シンポジウムやセミナーの告知.
人文科学 10	大学や自分の研究分野に関することについて. その他雑記的なツイート.
人文科学 11	時事 URL つきでツイート, 本務校での教育や研究について. 本に関するツイートも.
人文科学 12	仏教や哲学. 時事に関してもその目線で. 大学の地震の教育についてもツイートあり.
人文科学 13	コンテンツ (マンガ・ゲームなど), タロット, 本 (小説) について.
人文科学/社会科学	震災復興について.
社会科学 1	政治情勢について自分の意見を述べる.
社会科学 2	国際活動とジェンダー. 研究<活動.
社会科学 3	@が多い, 原発関連の意見.
社会科学 4	認知科学系アカウント. 論文ツイートあり, コメントあり.
社会科学 5	教育工学. 授業実況や受講生との交流あり.
社会科学 6	経済学者. 政治に関する評論や言論, テレビ出演に関するツイートを多く含む.

社会科学 7	法学。学会のテーマについてのつぶやきあり。主は刑事事件に関する見解。
社会科学 8	政治や白書の見解を述べている。
社会科学 9	法律や学問に関する内容。
社会科学 10	政治の思想について、議員といった個人の思想や行動に関して論じているツイートが多い。「w」が多い。
社会科学 11	自著からの引用フレーズを本文の一部のリンクと一緒にツイート。
社会科学 12	時事について、取り上げているツイート多い。ときおり、マスコミに対して批判的な視点で見ているものもあり。
社会科学 13	デジタルコンテンツ、情報教育政策について。
社会科学 14	経済学、政治学についてのツイート。自著についての紹介もあり。
社会科学 15	教育について（授業の仕方など）、写真が多い。
社会科学 16	経済学の話や本、ブログの紹介。研究会などのお知らせもあり。
社会科学 17	勤務校の教育や行事に関して。ほか、マネジメントについて。
社会科学 18	英語教育関連。
社会科学 19	日本とフランスのニュースについて多くリンクを交えてツイートすることが多い。
社会科学 20	原発関連について。
社会科学 21	ニュースや政府の報告書のデータを URL つきでツイート。
社会科学 22	英語のフレーズをツイート、自身の講演についての宣伝も。
社会科学 23	ニュースを URL とともにツイート。人材についてのツイートもある。
社会科学 24	政治や経済について、ニュース元のリンクを張り自分の意見を書きつつツイート。特に原発問題について。
社会科学 25	政治や経済について。
社会科学 26	政治や経済について。どちらかと言うと、理論で固めるより批難的な目線で。
社会科学 27	電力、原発について。
社会科学 28	教育や労働について。気になるブログやニュースやニュースを一部引用と URL をつけたツイート。
社会科学 29	原発についてリスクの視点から。非公式 RT を用い@による議論が多い。
社会科学 30	日常を経済学の視点で語る。
社会科学 31	日常について綴っている。たまに自分の日記の更新通知。
社会科学 32	授業やゼミについて、経営を人材や人間関係の視点で考えるツイート。
社会科学 33	シンポジウム、放送、自著についての告知が多い。ニュースをリンクでツイートすることも。

社会科学 34	スポーツについて，フォーラムの告知あり．@多い．
社会科学 35	@が多い．シンポジウムについての告知あり．
社会科学 36	市の政策について．@による問いへの回答も多い．
社会科学 37	自己啓発と自著の紹介へのお礼．
社会科学 38	福祉，医療関連のニュースや意見をツイート．自身の教育についても．
社会科学 39	本務校関連，ニュースについてリンク付ツイート．
社会科学 40	日常，メディアについてのニュースや催しについてリンク付でツイート．
社会科学 41	ニュースをリンク付でツイート．自分の出演や自著の紹介．
社会科学 42	ライフログ（スポーツ，音楽），日常，情報・メディア系のニュース．
社会科学 43	クラシック音楽に関するツイート．音楽鑑賞のコメント．
社会科学 44	学内のニュースやエンタメ系のニュースを URL つきでツイート．
社会科学 45	自身の blog 記事，ネットニュースの紹介．本務校関連の情報も少々．
社会科学 46	パソコン関連について，日常について．
社会科学 47	ニュースについてリンク付でツイート．
社会科学 48	時事を法の観点で見解をツイート．情報法が多い．
社会科学 49	研究とそれにかかわる生活について．
社会科学 50	政治や経済について．シンポジウムの告知．特定の話題にはタイトルを冠して，連投．
社会科学 51	皇族について．研究会の告知．
社会科学 52	経済や時事について．海外ニュースサイトの記事について URL つきでツイート．
社会科学 53	法について，時事について．
社会科学 54	モバイルや PC などについてやアプリなど電子メディアについて．
社会科学 55	政治や時事について．また，ゼミについての連絡．
社会科学 56	法学研究，学術，大学について．（シンポジウム？）実況．
社会科学 57	政治，社会事情．ジェンダー（フェミニズム）の視点から．原発についての言及．シンポジウム告知．
社会科学 58	経済学について．本の紹介や自分の出演番組の宣伝有．
社会科学 59	経済，財政について，趣味について．
社会科学 60	料理や食の安全について，とある市の振興の一貫として．社会政策について．
社会科学 61	裁判の判定など，弁護士関連，時事を url つきでツイート．
社会科学 62	経済や政治（国会）について，時事については url つきでツイート多い．
社会科学/ 総合領域	ニュース記事（主に原発）のリンクつきツイートが非常に多い．

数物系科学 1	原子 (物理) 関連ツイートや写真ツイート多数.
数物系科学 2	物理関連. 物理教育についても.
数物系科学 3	情報科学, 技術科学の予算費, 放射線についてなど.
数物系科学 4	放射能について.
数物系科学 5	放射能に対する@ (議論)
数物系科学 6	放射能関連
数物系科学 7	日常や除染について
数物系科学/総合領域	土壌汚染の観点から, データに基づき放射の影響について多くツイートしている
数物系科学/社会科学	統計や理科教育の必要性について
工学 1	放送大学に関する行事, 他のユーザーからの質問に答えるのが主. (教育についてやり直し)
工学 2	電子工学について
工学 3	時事 (政治, 原発) について. シンポジウムの告知もあり.
工学 4	研究生活, 日常
工学/人文科学/総合領域	デザイン, 美について
農学 1	教育, 漁業について
農学 2	@が多い. 日常
生物科学	研究や大学の講義, 教育について

## 4.3. ツイートのテキスト分析

### 4.3.1 相対出現頻度

ここからは、ツイートのテキスト分析の結果について述べていく。1ツイートあたりの文字数は専門家が59.65字、非専門家が30.08字であった。

表7-9はツイートテキストの相対出現頻度上位20位を示したものである。上から表7はすべてのツイートにおける相対出現頻度、表8は専門家アカウントのツイートにおける相対出現頻度、そして表9は非専門家アカウントの相対出現頻度である。

全体の傾向をみると上位に入っている語はいずれも助詞や助動詞など機能語であり、この結果は通常のテキスト分析の上位にくる語の傾向と同じと特徴と考えられる。次に専門家と非専門家アカウントの出現頻度上位語を見、比較していく。専門家アカウントの相対出現頻度では、非専門家の上位相対出現頻度では見られなかった、「です」や「ます」といった丁寧語の語尾が観察された。非専門家アカウントでは語尾として「ね」や「よ」といった砕けた表現で使われる語が上位にくる結果となった。また、「！」が上位に存在した。

表 7. 相対出現頻度上位

	語	相対出現頻度
1	の	0.056535
2	は	0.034807
3	に	0.034506
4	が	0.031752
5	て	0.031582
6	た	0.028152
7	。	0.027329
8	を	0.024084
9	、	0.023835
10	で	0.023579
11	と	0.019767
12	な	0.014136
13	も	0.013721
14	ない	0.012239
15	か	0.012073
16	だ	0.011323
17	です	0.010627
18	し	0.009222
19	ます	0.008612
20	から	0.007929

表 8. 専門家アカウント相対出現頻度上位

語	相対出現頻度
1 の	0.065318
2 は	0.039548
3 に	0.037127
4 が	0.036169
5 。	0.03229
6 て	0.031197
7 を	0.030464
8 、	0.028564
9 た	0.026213
10 で	0.025697
11 と	0.022765
12 も	0.014268
13 な	0.013201
14 ない	0.011803
15 です	0.011425
16 か	0.011006
17 ます	0.009836
18 し	0.009209
19 だ	0.008459
20 』	0.007888

表 9. 非専門家アカウント相対出現頻度上位

語	相対出現頻度
1 の	0.038023
2 て	0.032392
3 た	0.032238
4 に	0.028982
5 は	0.024815
6 が	0.022441
7 で	0.019115
8 だ	0.017359
9 。	0.016871
10 !	0.016582
11 な	0.016107
12 か	0.014322
13 、	0.013867
14 と	0.013448
15 ない	0.013159
16 も	0.012570
17 を	0.010635
18 よ	0.010529
19 ね	0.010074
20 から	0.009472

#### 4. 3. 2 品詞構成比率

表 10 は品詞構成比率を示したものである。全体の傾向として、助詞、助動詞の比率が多い。専門家の方が高い品詞は助詞、連体詞、記号であり、そのほかの助動詞、名詞、動詞、福祉、形容詞、フィラー<sup>7</sup>、接頭詞、感動詞、接続詞、その他は非専門家アカウントのツイートの方が高い比率であることが分かった。助動詞は用言や体言にいろいろな意味をそえるのに対し、助詞は語と語の関係を示したり、いろいろな意味を付け加えたいりする性質がある。

動詞が少ないのは実用的な文章の特徴であると言われており、非専門家のツイート文と比べ、その傾向があることが推測される (橘, 1991)。

表 10. 品詞構成比率

	助詞	助動詞	名詞	動詞	副詞	形容詞	フィラー
全体	0.230	0.079	0.385	0.109	0.019	0.018	0.002
専門家	0.244	0.077	0.381	0.104	0.018	0.015	0.000
非専門家	0.199	0.084	0.393	0.118	0.022	0.025	0.004
	接頭詞	感動詞	接続詞	その他 <sup>8</sup>	連体詞	記号	
全体	0.007	0.007	0.005	0.000	0.005	0.134	
専門家	0.007	0.003	0.004	0.000	0.006	0.139	
非専門家	0.008	0.014	0.005	0.000	0.004	0.124	

\*小数点第4位で四捨五入したため0と記述しているが、正確には0ではない

<sup>7</sup> 「うん」、「えと」といった話し言葉における相槌などを示す

<sup>8</sup> 間投詞

### 4.3.3 Twitter の記号, URL の相対出現頻度

表 11 は Twitter 独自の記号 (@, RT, #) と URL のツイートあたりの出現率を示したものである。@, URL, RT は専門アカウントの方が非専門家アカウントより高い頻度で出現していることが分かった。

ここで言う RT とは非公式 RT といわれている使い方である。非公式 RT とは Twitter サイトに RT 機能実装される以前から他のツイートからの引用を示す記法として使われていたものである。書き方には諸説あるが、「RT @相手のユーザー名: 引用ツイート」という示されることが多い。RT が公式サイト他クライアントに実装されてからは公開して会話を行う場合やツイートの引用とコメントを 1 つのツイート内で行うときに用いられる。専門家アカウントの場合、非専門家アカウントに比べ、議論を行ったり意見を述べたりする傾向があることから非公式 RT である「RT」の記号が頻出したと考えられる。

もう 1 つ顕著な差として見られたのは URL の相対出現頻度である。ツイートにおける URL の登場する割合が示されている研究, 吉田ら (2010), 山口ら (2012), Boyd et al. (2010) ではそれぞれ 0.20, 0.14, 0.22 とされている。これらの研究はいずれも Twitter 全体をクロールして収集したデータを利用している。専門家アカウントは比較的 URL が頻出する傾向があると言える。

表 11. Twitter の記号および URL の出現率

	@	RT	#	URL
全体	0.462	0.139	0.085	0.206
専門家	0.543	0.248	0.094	0.322
非専門家	0.383	0.033	0.076	0.095

## @におけるフォロー/フォロワー比率

表 12 は@相手とのフォロー/フォロワー比率を示したものである。表の記号は対象アカウントから見たフォロー/フォロワー関係を示しており、左がフォロー、右がフォロワーの関係を示している。有 (t) , 無 (f) である。全体では tt と互いにフォロー/フォロワー関係である場合 (tt) が 6 割を超える結果となった。次に多かったのが対象アカウントは@相手をフォローしていないが、@相手は対象アカウントにとってフォロワーである場合 (ft) で 2 割弱であった。それぞれの専門家アカウント、非専門家アカウントそれぞれの数値を見ていくと、専門家アカウントは約半数が tt であることが分かる。次いで、ft が多く 0.271 がそれに該当している。一方、非専門家クラスは 0.869 が tt であり、ほとんどが相互にフォローしあっているアカウント同士で@を行っていることが伺える。専門家アカウントで多かった ft は 0.023 であり、他のどの項目よりも少ない割合となっている。アカウントによっては Twitter 上で議論を日常的に行っていたり、意見を述べていたりすることから、そのレスポンスとしてフォローからの@が送られ、それに返答していることが予測される。非専門家アカウントは相互につながっているクローズな関係の中で@がなされていると言え、自分がフォローしていない相手に@を送ることは少ない (もしくは@が送られて来ない) ことが想定される。

表 12. @におけるフォロー関係

	tt	tf	ft	Ff
全体	0.629	0.066	0.188	0.117
専門家	0.509	0.069	0.271	0.151
非専門家	0.869	0.060	0.023	0.048

#### 4.4. 専門家/非専門家アカウント分類実験結果

専門家/専門家アカウント分類実験の結果は表 13 である。特徴量として相対出現頻度 (BOW) のみを用いた実験, BOW に文字数を加えた実験, BOW に Twitter の特徴量を加えた実験, そして BOW, 文字数, Twitter の特徴量を用いた実験, いずれも精度, 再現率,  $F$  値 9 割を超える性能を得ることができた。個別の実験を見ると, BOW と Twitter を用いた実験が最も性能が良く, 精度, 再現率,  $F$  値いずれも 96.22, すべての特徴量を用いた実験と BOW のみで行った実験が精度, 再現率,  $F$  値が 95.60 で同値, もっとも低かったのは BOW+文字数の実験で精度, 再現率,  $F$  値いずれも 94.96, という結果になった。通常, 機械学習では特徴量が多い方が良い性能が得られるといわれている。このことから Twitter の特徴量の中で分類において影響を及ぼす特徴量が存在していることが推測される。

表 13. 分類実験結果

	精度	再現率	$F_1$
BOW	95.80	95.80	95.80
+文字数	94.96	94.96	94.96
+Twitter	96.22	96.22	96.22
+文字数+Twitter	95.80	95.80	95.80

## 分類における重要な特徴量

表 14 は全特徴量を用いた実験より得られた分類に有用な特徴量上位 20 を抽出したものである。ft, tt は Twitter の特徴量, フォローフォロワー関係を示したものである。そのほかは BOW が該当する。これらの特徴量は専門家か非専門家かを判定するときに有用な変数として算出されたものである。1 番分類に寄与した特徴量は @フォローフォロワー関係を示す, ft であった。これは表 12 で示されている通り, 専門家アカウントに顕著に出ていた特徴である。自動分類においても分類に寄与していることが分かった。16 位に位置づけられた, tt も同じく @におけるフォロー/フォロワー関係を示しており, これは非専門家の特徴である。「という」や「について」は非専門家アカウントに比べ, 専門家アカウントに多く出現する語であり, 意見を述べるなど説明をしているときに用いられているのではないかと推測される。

表 14. 分類に寄与の大きい特徴量 (上位 20 位)

	語	タイプ	$V_{acu}$
1	ft	Twitter	0.0072490
2	を	BOW	0.0052572
3	者	BOW	0.0049119
4	一	BOW	0.0048632
5	の	BOW	0.0047865
6	バンドネオン	BOW	0.0046945
7	という	BOW	0.0043200
8	社会	BOW	0.0035740
9	いる	BOW	0.0035374
10	教育	BOW	0.0032775
11	」	BOW	0.0032298
12	。	BOW	0.0032043
13	問題	BOW	0.0030212
14	わ	BOW	0.0029444
15	研究	BOW	0.0029369
16	tt	Twitter	0.0028664
17	日本	BOW	0.0028137
18	について	BOW	0.0027849
19	木山	BOW	0.0026875
20	ω	BOW	0.0024863

特徴量数 : BOW = 35279, Twitter = 8, 文字数の全 35288

## 5. おわりに

本研究は学術情報関連の情報発信者の特徴分析と自動分類実験を行った。アカウントの特徴では社会科学が多く、所属は現実の大学設置数を反映しており、職階についても所属と同じ様子がみられた。ツイートに関しては明らかに専門情報を含むものは少なかったものの、個々のアカウントのツイート内容を見ると専門分野に近い内容の時事や専門に近い内容の日常ツイートが観察され潜在的に学術関連情報があることが示唆された。自動分類実験では Twitter の特徴量を入れた方が専門家アカウントを自動分類するには有用であることが考えられる。個々の特徴量としては、「教育」や「研究」といった語が上位にあり、ツイート内容の分析結果と合わせて推測するに、どの分野でも共通の話題として利用されている語である可能性がある。このことから自動分類においても重要な特徴量として現れる結果となった。

また今回は手作業による特徴分析を行ったが、自動で特徴を抽出するためには分野ごとの分類も重要となり、名詞を用いたテキスト分類や LDA といった自動要約を用いての表示が必要であると考え、今後の課題である。

## 謝辞

指導教員である芳鐘冬樹准教授，助言をくださいました副指導教員である池内淳准教授，審査を快諾してくださいました辻慶太准教授にこの場を借りて深く感謝いたします。ほか芳鐘研究室のメンバーにも感謝の意を表します。ならびに，お世話になりました先生方，同期の皆様にもこの場を借りてお礼を表します，ありがとうございました。また，第 60 回日本図書館情報学会研究大会，JADH2013 にて貴重なご指摘やご意見を頂きました皆様にも深く感謝いたします。

## 文献一覧

- 青山幸太, 林和弘. 特集,第7回情報プロフェッショナルシンポジウム: C13 新コミュニケーションメディア利用動向調査について: 日本化学会第 90 春季年会展示における, Twitter 等のメディア・ツール利用に関するアンケート調査報告 (セッション C1 情報と教育研究, 一般発表概要), 情報の科学と技術, 2011, vol.61, no.3, p.95.
- 荒牧英治, 増川佐知子, 森田瑞樹. Twitter Catches the Flu: 事実性判定を用いたインフルエンザ流行予測. 情報処理学会研究報告. SLP, 音声言語情報処理, 2011, no.1, p.1-8.
- D. Boyd; S. Golder and G. Lotan. Tweet, Tweet, Retweet: conversational aspects of retweeting on Twitter, *Hawaii International Conference on System Sciences*, 2010. p.1-10.
- L., Breiman. "Random Forests". *Machine Learning*, Vol.45, No.1, 2001, p. 5-32.
- L., Breiman; Adele Cutler. "Random forests - classification description", Random forests [http://stat-www.berkeley.edu/users/breiman/RandomForests/cc\\_home.htm](http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm), (accessed 2014-01-07).
- Golder, S. A.; Yardi, S.; Marwick, A.,; Boyd, D. A structural approach to contact recommendations in online social networks. In Workshop on Search in Social Media, SSM. (2009).
- 後藤清豪, 高田秀志. ソーシャルメディア上での行動に基づく「意外な情報」の提供者になり得る人物の推薦手法. 情報処理学会研究報告. データベース・システム研究会報告. 2010, no.41, p.1-8.
- 早川豪, 岡部誠, 尾内理紀夫. Twitter を利用したソーシャルニュース記事推薦システム. 情報処理学会研究報告. データベース・システム研究会報告. 2011, vol.2011, no.16, p.1-4.
- インターネット協会. インターネット白書 2012. インプレス R&D, 2012, p.255.
- 金明哲, 村上征勝. ランダムフォレスト法による文章の書き手の同定. 数理統計. 2007, vol.55, no.2, p.255-268.
- J. Letierce ; A. Passant and J. G. Breslin, Understanding how Twitter is used to spread scientific messages. *Web Science conf.* 2010
- 文部科学省. “学校基本調査 年次統計 平成 25 年度 大学の職名別教員数 (本務者)”. <http://www.e-stat.go.jp/SG1/estat/Xlsdl.do?sinfid=000023610162>, (参照 2014-1-15).
- 荻上千キ. 検証東日本大震災の流言・デマ. 光文社, 2011, p.204.
- J. Ritterman; M. Osborne; E. Klein. "Using prediction markets and Twitter to predict a swine flu pandemic". 1st international workshop on mining social media. 2009.
- 白木原渉, 大石哲也, 長谷川隆三, 藤田博, 越村三幸. Twitter における流行語先取り発言者の検出システムの開発. 情報処理学会研究報告. データベース・システム研究会報告. 2010, vol.2010, no.2, p.1-8.
- 総務省.”第1部 第4節 情報通信が果たした役割と課題”.情報通信白書平成 23 年版. 総務省. <http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h23/pdf/n0040000.pdf>, (2014.1.15 参照).
- 総務省.”第4章 第3節 インターネット利用動向”.情報通信白書平成 25 年版. 総務省.

- <http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/pdf/n4300000.pdf>, (2014.1.7 参照) .
- 細谷舞, 鈴木崇史. "女性シンガーソングライターの歌詞の探索的分析", じんもんこん 2010: 人文科学とコンピュータシンポジウム論文集. 東京, 2010-12-11/12, p.195-202.
- 橘豊. "第 I 編 10. 実用文の文体分析". 文体論の世界. 日本文体論学会. 三省堂, 1991, p.114-124.
- 高村大也. 言語処理のための機械学習入門, コロナ社, 2010, p.211.
- 高村大也, 横野光, 奥村学. Summarizing Micoroblog Stream, 第22回セマンティックウェブとオントロジー研究会人工知能学会 セマンティックウェブとオントロジー研究会,2010.
- 立入勝義. 検証 東日本大震災 そのときソーシャルメディアは何を伝えたか?, ディスカヴァートゥエンティワン, 2011, p.267.
- 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999, 234p.
- トライバルメディアハウス, クロス・マーケティング. ソーシャルメディア白書. 翔泳社, 2012, p.251.
- A. Tumasjan; T. O. Sprenger,; P. G. Sandner,; Welpe; M. Isabell. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM. 2010, vol.10, p.178-185.
- Uysal, I.; Croft, W. B.. "User oriented tweet ranking: a filtering approach to Microblogs". *Proceedings of the 20th ACM international conference on Information and knowledge management* . Glasgow, United Kingdom , 2011-10-24/28, 2011, p. 2261-2264.
- 山口裕人, 天笠俊之, 北川博之. "マクロブログにおけるユーザ専門に考慮した Web ページ推薦". 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回データベース学会年次大会) . 神戸, 電子通信学会データ工学研究専門委員会, 日本データベース学会, 情報処理学会データベースシステム研究会, 2012-3-3/5, 2012, F2-4.
- 吉田光男, 乾孝司, 山本幹雄. リンクを含むつぶやきに着目した Twitter の分析, 第2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8回データベース学会年次大会) . 神戸, 電子通信学会データ工学研究専門委員会, 日本データベース学会, 情報処理学会データベースシステム研究会, 2010-2-28/3-2, 2010, A5-1.
- K. Weller; C. Puschmann. Twitter for Scientific Communication: How Can Citations/References be Identified and Measured?. WebSci'11, 2011 (Poster) .