Twitter上のデマ情報の検出

3班
齋藤和孝, 田嶋脩平, 中村裕
アドバイザー教員：遠藤靖典

- **背景・目的**
- 関連研究
- データの収集と分析
- 分析の結果

背景

Twitterの普及

- **Twitterとは**
  - インターネット上のコミュニケーションツール
  - ユーザが短文（ツイート）を投稿 ⇒ ユーザ間で共有
  - 他人のツイートを再投稿（リツイート）できる
  - アカウント数：2億以上 (2010年)
  - 1日のツイート数：1.5億以上 (2010年)
  - 情報取得、伝播ツールとしても利用
- **誰でも気軽に情報発信**

⇒伝達スピードは既存のメディアよりも速い

緊急時にも活躍

背景

東日本大震災におけるTwitter利用

- 携帯電話がつながらない状況下での連絡手段
- テレビやラジオが使えない状態での情報取得・伝播
- 孤立した避難所からの救助・援助物資の要請
- 自治体によるライフラインなどの情報提供

Twitterは様々な場面で活躍
盛んな情報交換が行なわれた

しかし

多くの間違った情報や
デマの伝播が問題視

背景

Twitterにおけるデマ拡散

- **東日本大震災でも多くのデマが伝播**　⇒混乱と不安を招く
  - 放射線対策にヨウ素が有効
  - 石油タンク爆発による有害物質の放出
- **海外でもデマの伝播は発生**
  - 2010年のチリ地震時
  - イギリスのテロ予告
- **緊急時以外にもデマは伝播**

最悪の場合に備えて：成人でヨウ化カリウムを1日130mgをヨウ化カリウム錠剤で飲むのがよいが，ない場合は市販のうがい薬のルゴール液でも代用可能．数滴（うがいで使う時の半分以下）を水で薄めて飲めば十分
3月12日

【拡散希望】千葉市近辺にいらっしゃる方！コスモ石油の爆発により有害物質が雲などに付着し、雨などと共に降る可能性があるので外出の際は傘やカッパなどを持ち歩き、雨などに身体が接触しないようにして下さい！
3月11日

Twitter上のデマ拡散の顕在化…
そのリスクを調査し，評価する

目的

Twitter上から得られる情報のみを利用してデマの持つ特徴を抽出し、デマを自動で検出する

Twitter上から得られる情報

- ユーザの登録情報（登録日数・フォロワー数など）
- ネットワーク構造（リツイート伝播経路など）
- メッセージの情報（肯定語・否定語数など）

デマの定義

- 本来の意味：悪意を持って流される嘘の情報

↓ **社会的影響を考慮**

- 本研究：投稿者の意図に関わらず内容が事実と異なる情報

- 背景・目的
- **関連研究**
- データの収集と分析
- 分析の結果

信用性の自動評価に関する研究

Castillo et al. “Information Credibility on Twitter” (2011)

- **目的**
  - ソーシャルメディア内の情報のみからツイートの信用性を自動評価する
  - ツイートの信用性評価に利用できる特徴を抽出
- **手法**
  - Twitter Monitorで盛り上がっている話題を取得
  - NEWS(報道価値あり) とCHAT(友達との会話)に分類
    ⇒信用性評価はNEWSクラスのツイートのみ行う
  - 学習データの作成 (Mechanical Turkを利用)
    - 情報が信用できるかを4段階で評価してもらう
  - 信用性に関わる特徴をピックアップ (先行研究から, 全68種類)
  - 信用性の有無を識別する分類器作成

信用性の自動評価に関する研究

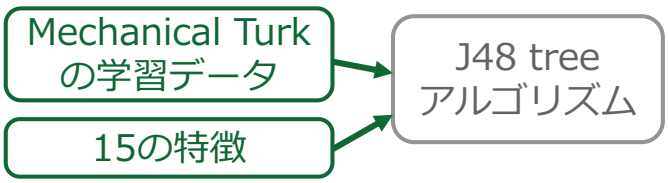
信用性に関わる特徴

- 最良優先探索により15種類の有用な特徴を選定

User-based	Topic-based	Propagation-based
•登録日数 •ツイート数 •フォロワー数 •フォロー数	•URLを含む頻度 •Positive表現の頻度 •Negative表現の頻度 •？マークの頻度　等	•RT最大深さ

信用性の自動評価

- 決定木の作成



信用できる情報	信用できない情報
•RT多い •フォロー数多 •Negativeニュアンス	•URL無 •ツイート数少 •Positiveニュアンス

- F値86%で情報の信用性の有無を判断

自作自演ミームの検出に関する研究

Ratkiewicz et al. “Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams” (2010)

*ミーム(meme)：ネットワーク上を拡散する噂, デマ

目的

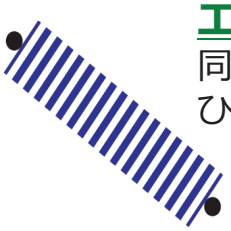
- Twitter上での“自作自演”を検出
- “自作自演”ミームの拡散ネットワークの特徴を明らかにする

手法

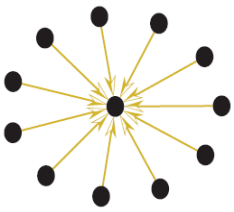
- Truthyシステムの作成
  - Twitter APIにより与えられたデータの監視, 収集, 処理
  - ミームの検出
  - Truthy site (疑わしい情報にWeb上でユーザが注釈をつける)
- 自作自演を検出する分類器の作成

自作自演ミームの検出に関する研究

- **自作自演ミームの特徴**
  - 自然発生した情報と拡散ネットワークの様子が異なる
  - ミームの起点が多い
  - 平均次数が高く星のような形になる
  - 2要素間のエッジの重みが大きい



エッジの重み大きい
同じアカウントのツイートを
ひたすらリツイート
ex) Twitter bomb
webサイト上昇



起点多い・次数高い
10のボットによる
リプライ
メッセージ中に同じ
URL含む

ミーム検出によりアカウントを強制停止された例も存在

本研究との関連

- **信用性に関わる特徴**
  - 関連研究：海外のツイートを分析
    ⇒ 抽出された特徴が日本語でも有用か検討
  - 日本語環境で有用な特徴を利用した分類を実施

- **分類手法**
  - 関連研究：教師あり学習による分類
    ⇒ 膨大な学習データ・ユーザによる評価が必要

コスト増

**学習データを必要としないクラスタリング手法
による分類を実施**

- 背景・目的
- 関連研究
- **データの収集と分析**
- 分析の結果

データの収集

- **収集するデータ**
  - **ニュース**
    社会で実際に起こった出来事や情報
  - **デマ**
    投稿者の意図に関わらず内容が事実と異なる情報

判断基準：大手新聞社による報道や公的機関などの発表

- **定義を満たす情報をWeb上で検索**
  - リツイート数の多い情報
- **収集した情報**
  - 84ツイート (デマ30、ニュース54)
  - リツイートをした16958人分のユーザ情報

肯定語・否定語の検出

- **先行研究：肯定語・否定語数が信用性に影響**
- **肯定語・否定語を自動検出するシステムを作成**
  1. **形態素解析による品詞抽出**
  2. **評価極性辞書との比較**
     評価極性辞書：品詞の持つ意味を評価 (肯定的 or 否定的)
     用言 + 名詞 = 約 13,500 語
  3. **評価値の算出**
     ツイートに含まれる否定語、肯定語の数をカウント

イソジンではなくとろろ昆布わかめ味噌玄米との情報も
RT：イソジン誤。昆布とろろ昆布やわかめ海藻。放射線に
添加なしの昔ながらの味噌や玄米も効くRT :福島1号基中央制御室
通常1000倍。放射能漏子供にヨウ素イソジン3滴水薄1日1回

肯定語・否定語の検出

- **先行研究：肯定語・否定語数が信用性に影響**
- **肯定語・否定語を自動検出するシステムを作成**
  1. **形態素解析による品詞抽出**
  2. **評価極性辞書との比較**
     評価極性辞書：品詞の持つ意味を評価 (肯定的 or 否定的)
     用言 + 名詞 = 約 13,500 語
  3. **評価値の算出**
     ツイートに含まれる否定語、肯定語の数をカウント

イソ ジン だ は ない とる る る 昆布 わか め 味噌 玄米 と の 情報 も
RT：イソ ジン 誤る 。 昆布 と る る 昆布 やわ かめ 海藻 。 放射線 に
添加 ない の 昔ながら の 味噌 や 玄米 も 効く RT：福島 1 号 基 中央 制御 室
通常 1000 倍 。 放射能 漏る 子供 に ヨウ 素 イソ ジン 3 滴水 薄 1 日 1 回

肯定語・否定語の検出

- 先行研究：肯定語・否定語数が信用性に影響
- 肯定語・否定語を自動検出するシステムを作成
  1. 形態素解析による品詞抽出
  2. 評価極性辞書との比較
     評価極性辞書：品詞の持つ意味を評価 (肯定的 or 否定的)
     用言 + 名詞 = 約 13,500 語
  3. 評価値の算出
     ツイートに含まれる否定語、肯定語の数をカウント

イソ ジン だ は ない とる る る 昆布 わか め 味噌 玄米 と の 情報 も
Pos
RT：イソ ジン 誤る 。 昆布 と る る 昆布 やわ かめ 海藻 。 放射線 に
Neg Neg
添加 ない の 昔ながら の 味噌 や 玄米 も 効く RT：福島 1 号 基 中央 制御 室
通常 1000 倍 。 放射能 漏る 子供 に ヨウ 素 イソ ジン 3 滴水 薄 1 日 1 回
Neg

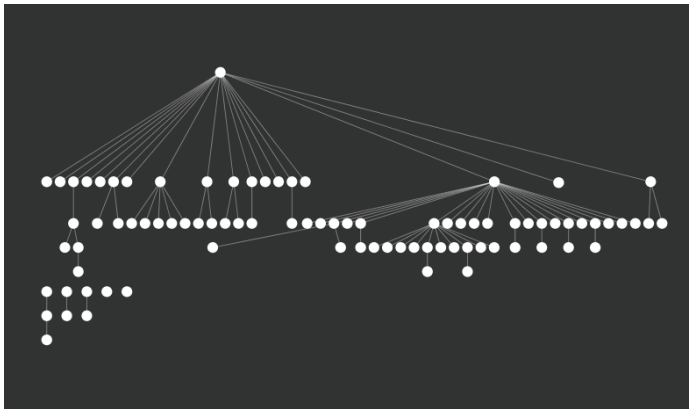
ツイート伝播経路の取得

1. リツイートを行ったユーザの取得
2. ルートノードの取得
3. ルートから順にユーザのフォロワーを探索

↓

リツイートツリー(RT)の作成

例



デマとニュースの分類の指標

- **デマ・ニュース間の有意差を確認**
  - t検定 有意水準5%
  - ユーザの登録情報、感情値、ネットワーク構造

有意差が見られた3つの指標

Np
- **否定語数 ＋ 肯定語数**
  ツイート内の感情値スコア

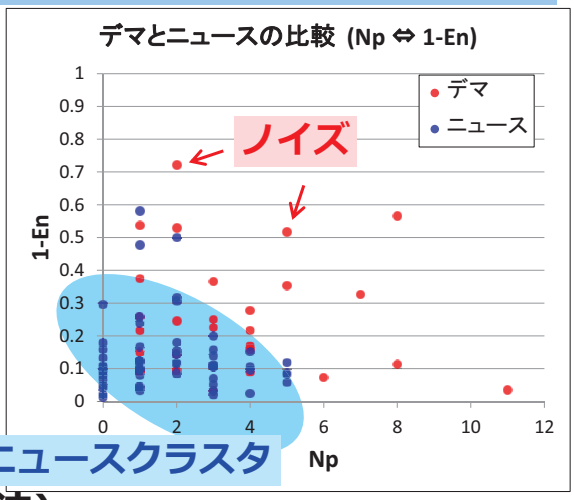
1-En
- **1 － RTエッジ数 / RTノード数**
  RTの構造的特徴

Rn
- **RT最大深さ / RTノード数**
  RTの構造的特徴

分類アルゴリズム

- **指標を用いたプロット**
  - ニュースは1点に集中
  - デマは全体的に分散 (ノイズ)

クラスタとノイズに分類できないか？



- **ノイズクラスタリング**
  (ノイズデータに対するファジィ c-平均法)
  - 1つのクラスタ (ニュース) とノイズ (デマ) に分類

目的関数

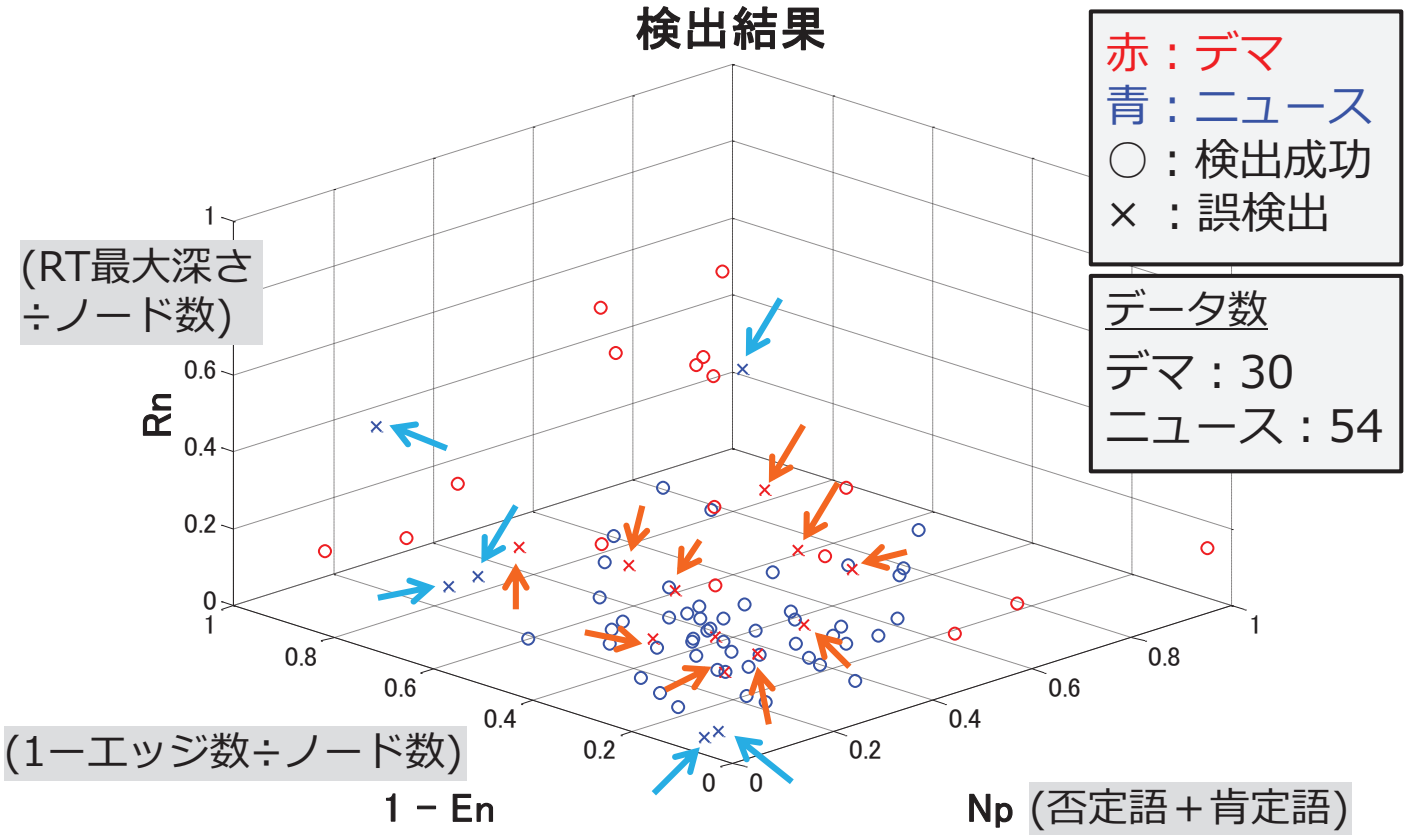
$$J(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m D_{ki}^2+\sum_{k=1}^{n}(u_{k0})^m\delta^2$$

n：データ数
c：クラスタ数

目的関数を最小化するU, Vを求める

- 背景・目的
- 関連研究
- データの収集と分析
- **分析の結果**

デマの検出 ～クラスタリング実行結果～



検出率

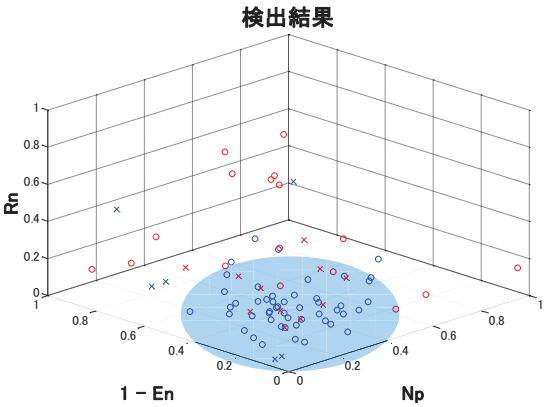
	デマ	ニュース	全体
T	19 (64%)	49 (89%)	67 (80%)
F	11 (36%)	6 (11%)	17 (20%)

- **80%の精度でデマとニュースの分類が可能**

- **デマの検出**
  - 0か1か (2値) で判定
  - さらに…
    デマの「度合い」を表す関数があってもいいのでは?

デマ度の推定

- **プロット結果の特徴**
  - 原点周辺に「ニュース」が集中



  - 原点からの距離でデマの度合いを推定

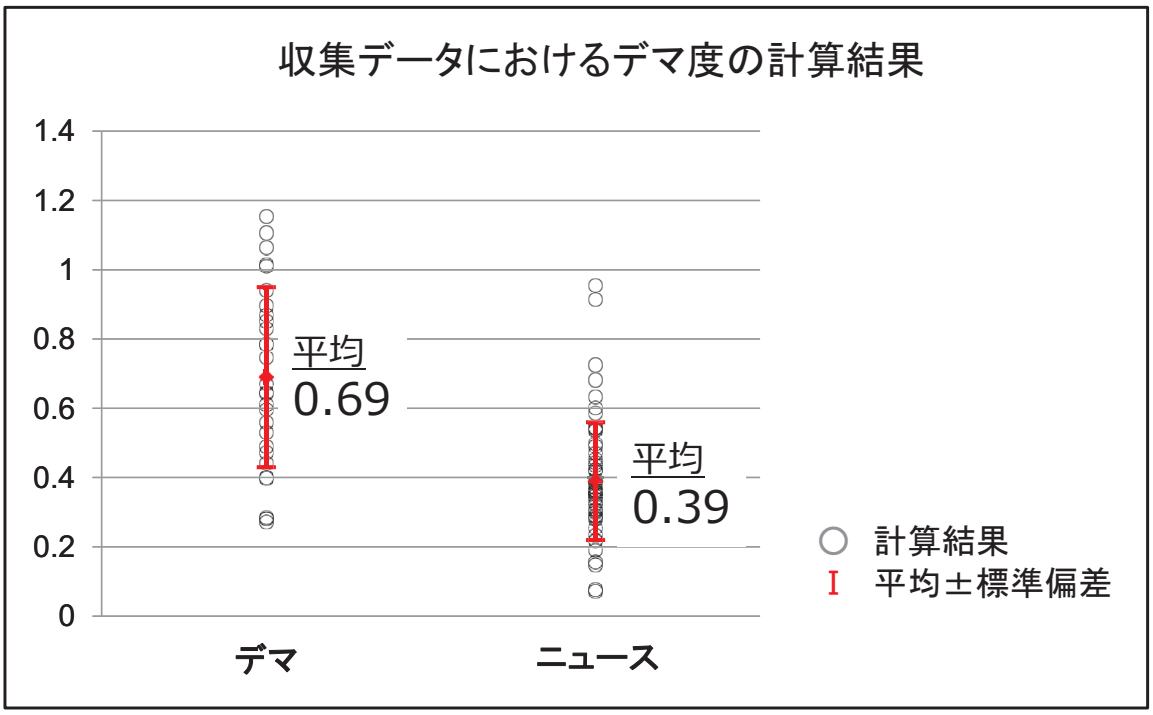
$$D = \sqrt{(\alpha Np)^2 + (\beta(1-En))^2 + (\gamma Rn)^2}$$

D : デマ度, α, β, γ : 正規化のための係数

$\alpha = 0.0909, \quad \beta = 1.38, \quad \gamma = 4.38$

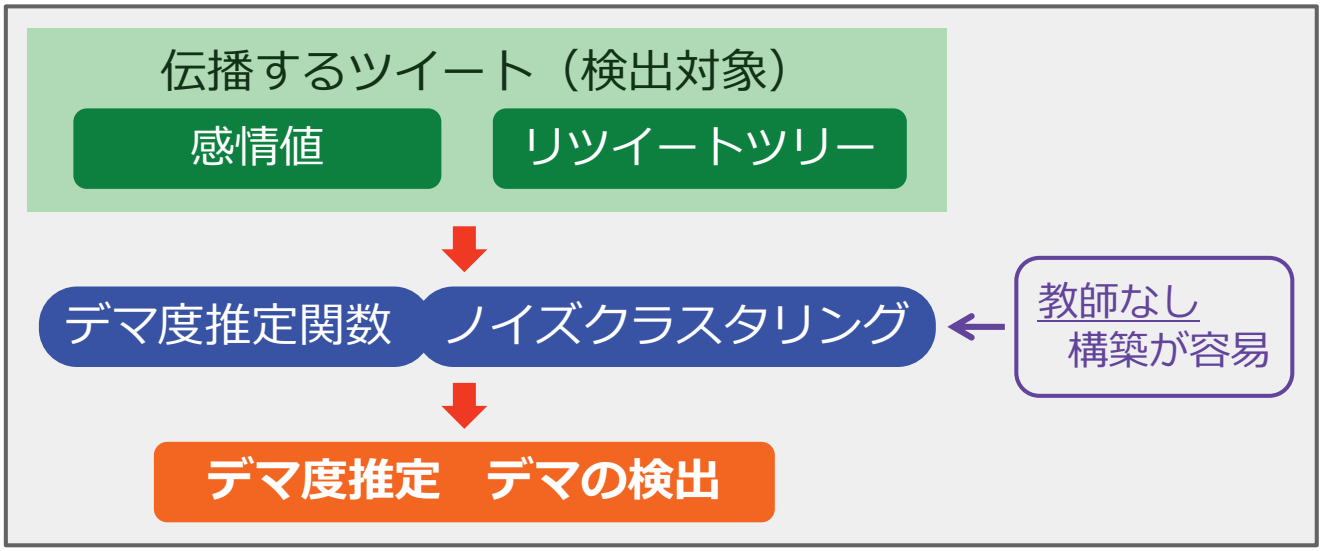
デマ度の推定

デマ度の計算結果



分析のまとめ

デマ検出の流れ



簡易的な手法で80%の精度を達成

まとめと今後の課題

目的
- Twitterにおけるデマの特徴を抽出
- 得られた特徴を利用したデマの分類

手法
- 「デマ」と「ニュース」に分けてツイートを収集
- 有意差の見られた3つの指標を利用
- ノイズクラスタリングの適用

結果
- 80%の精度でデマとニュースを識別
- 原点からの距離を利用したデマ度の推定

今後の課題
- 非公式リツイートによる伝播の考慮
  RT @... : メッセージ
- ユーザへフィードバックする仕組みの検討

参考文献 (1/3)

[1] T.Sasaki, M.Okazaki, Y.Matsuo: "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors", SOM (2010).

[2] 震災に伴うメディア接触動向に関する調査（野村総合研究所）
http://www.nri.co.jp/news/2011/110329.html

[3] 藤江利彦: 'はじめてのマスコミ論', 同友館 (2006).

[4] 香内三郎, 山本武利: '現代メディア論' , 新曜社 (1987).

[5] 佐藤卓己: '現代メディア史', 岩波テキストブックス (1998).

[6] 鈴木みどり: 'メディア・リテラシーの現在と未来', 世界思想社 (2001).

[7] 富山英彦: 'メディア・リテラシーの社会史', 青弓社 (2005).

[8] M.Schmierbach, A.Oeldorf-Hirsch: "A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions", AEJMC (2010).

参考文献 (2/3)

[9] C. Castillo, M. Mendoza, B. Poblete: “Information Credibility on Twitter”, Proceeding of the 20th international conference on World wide web, pp.675-684 (2011).

[10] J. Ratkiewicz, M. Conover, M. Meiss: “Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams”, arXiv (2010).

[11] 早川洋行: ‘流言の社会学’, 青弓社 (2002).

[12] 川上善郎, 松田美佐, 佐藤達哉: ‘うわさの謎’, 日本実業出版社 (1997).

[13] Twitter API wiki, http://usy.jp/twitter/index.php?Twitter%20API

[14] 日本語評価極性辞書
http://cl.naist.jp/inui/resuarch/EM/sentiment-lexicon.html

[15] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: “意見抽出のための評価表現の収集”, 自然言語処理, Vol.12, No.2, pp.203-222 (2005).

参考文献 (3/3)

[16] 小林のぞみ, 乾健太郎, 松本裕治: “述語の選択選考性に着目した名詞評価極性の獲得”, 言語処理学会第14回年次大会論文集, pp.203-222 (2005).

[17] 茶筅
http://chasen.naist.jp/hiki/ChaSen

[18] 荒井健太: “逐次的クラスター抽出アルゴリズムの開発と比較検討”, 筑波大学大学院システム情報工学研究科修士論文 (2009).