4Q-3

Twitterからの話題抽出と元記事の多種性からの信頼性の推測について

岡部　崇央†　　西原　典孝‡　　上嶋　明‡

岡山理科大学大学院工学研究科情報工学専攻†　　岡山理科大学工学部情報工学科‡

1.　はじめに

現在，情報社会の普及により，様々な人がインターネットを利用し，情報の収集や発信を行っている．近年では Twitter などのマイクロブログから情報を発信することが顕著で，ユーザー数は現在約 5 億人と言われている．このように多くのユーザーが情報を提供することが容易になった．反面，その信頼性が問題となってきた．例えば 2011 年の東日本大震災では多くの情報が飛び交う中，同時に嘘やデマも多く飛び交い，情報に対する信頼性が浮き彫りとなった．

このような観点から本研究では，まず Twitter 上で話題となっている出来事を抽出し，さらに Twitter の始まりにある引用先の記事（元記事）の一致性を測定することで，情報の信頼性の 1 つの目安とすることを考えた．例えば新聞記事にもなっているような話題は元記事が色々あるが，恣意的な話題や怪しい話題は情報源が単一であることが多いとみなせる．

2.　研究背景

近年嘘やデマが意図的に流れることが多い．またこのような情報も Twitter などを通して一挙に様々な場所に拡散してしまう．嘘やデマ，話題の信憑性を客観的に判断することは極めて困難である．従来研究として宮部らの研究 [1] がある．この研究では流言情報の発信を元とし流言データベースを構築するが，流言情報データベースの作成に時間が掛かることや，嘘やデマが大きく回った後に判定するなどの問題がある．そこで本研究では，判定の１つの目安として，話題となっている出来事の大元の情報源（いわゆる元記事）を収集し，その文章の引用元が同一なものなのか，色々あるのかを計測し，元記事の単一性，多種性を判定する．これを持って，その話題が恣意的に広められたもとなのか，実際に広く話題になっているものかの判定基準にできると考えた．

3.　研究内容

本研究で作成したシステムは「話題抽出」部門と「元記事の多種性の判定」部門からなる．

3.1　話題抽出

本システムでは Twitter 上での発言（以降，Tweet と呼ぶ）を 24 時間観察することで，話題を抽出する．各 Tweet に含まれる URL から元記事を辿ることにより，辿られた元記事をひとつの話題とする．話題収集には主に Twitter 社が提供している Twitter Trends API と Twitter Search API を使う．Twitter Trends では，今 Tweet されている話題を Twitter 社が独自のアルゴリズムにより，単語 10 件をまとめ，ユーザーに提供するサービスである．Trends で収集した単語を「Twitter Search API」により検索することで単語に対してのユーザー Tweet 一覧を取得することができる．それらの Tweet に含まれる元記事の URL を抜き出す．

> [Tweet]
> 16GB、19,800 円の方が魅力。♪(´▽`)→ISP 各社、Android タブレット「Nexus 7」8GB モデルを販売 - ケータイ Watch
> http://t.co/ubMj7VTg

図 1　Tweet

次に同じ話題がどれだけの数，Twitter 上で取り上げられているかを以下の手順でカウントする．

①同じ URL は同じ話題であるので，同一の URL でまとめ，その数をカウントする．

②URL で参照されている記事本体からタイトルを取り出す．

③異なる URL でも同じ内容と思われるものをタイトルから推測し，同じ話題としてグループ化する．

話題の度合いは①の数を合算した数（すなわちその話題の Tweet 数）とする．同じ内容か否かの判定は，以下のように行う．

記事のタイトル中に含まれる名詞リストを形態素解析システム MeCab[4]を用いて抽出する．この名詞リストがある程度一致すれば同じ話題と判断する．

以下に話題抽出の結果の一例を示す．データベースには話題タイトル，記事の URL，記事の内容，話題が Tweet された数，話題の名詞リスト，グループ番号となっている．

title	url	contents	counts	words	groups
PhotoshopなどのCS2製品が無料？　アクティベ	http://www	AdobeがPhoto	219	2,Adobe,CS,	1
AdobeがPhotoshopなどのCS2製品を無料で配布	http://jin11	309RT【宇宙ヤ	147	2,Adobe,CS,	1
【乞食速報】PhotoShop、Illustrator等のCS2製	http://0taku	3名前：名無し	126	2,CS,Illustra	1
Twitter / touri_69: 今回発表された新作、ポケ	http://twitt	Twitter/touri_	97	69,Twitter,X	5
メッシが史上最多となる4年連続4度目のFIFA(	http://www	FIFA（国際サッ	92	4,8211,FIFA	6
Twitpic - Share photos and videos on Twitte	http://twitp	ニンテンドー3	79	Twitpic	7
必要システム構成： Photoshop CS2	http://helpx	AdobePhotosh	58	システム,必要	8
【悲報】漫画サンデー 廃刊決定！！｜オタクニ	http://otane	少年サンデーの	47	サンデー,廃刊	9

図 2　グループ化後のデータベース

3.2　元記事の多種性の判定

抽出した Twitter 上での話題の信頼性判定の一つの目安として，同じ話題の元記事の種類数を測定する．これを元記事の多種性とここでは呼ぶ．

例えば同じ話題でも元記事が色々ある場合（同じ事件を扱った各新聞社の HP 上の記事など）は，元記事の多種

Topic extraction and reliability projection by the variations of the quoted articles on Twitter

†Takachika Okabe ‡Noritaka Nishihara ‡Akira Uejima

†Department of Information and Computer Engineering, Okayama University of Science Graduate School

‡Department of Information and Computer Engineering, Okayama University of Science

性は高く，信頼できると見なせる．一方，話題の度合いは高いが，引用 URL が同じ場合（芸能人に関するセンセーショナルな噂など）や引用 URL が異なっても，その元記事の文章が同一である場合（同じ記事のコピーをあちらこちらに貼り付け，恣意的に流そうとした話題など）は元記事の多種性は低く，信頼性は低いと見える．具体例を以下に示す．

事例 1：

9 月 10 日に自民党の谷垣総裁が総裁選への不出馬を表明した．その際 TV や新聞で大きく報じられ Twitter 上でも話題となった．Twitter 上での記事を分析した結果以下の結果が出た．

① 元記事が様々なサイトに存在すること
② それぞれ記事の内容構成が違うこと

この話題は元記事が様々な場所から出ているため，信頼性が高いと判断出来る．

事例 2：

9 月 19 日に「天皇陛下，訪韓も謝罪も両国民のためなら」という記事が Twitter 上で話題となった．様々な Twitter 上での元記事を分析した結果，元記事は 1 つの週刊誌ということが分かった．したがって以下のことが言える．

① 元記事が同じであること
② 各サイトの記事の内容が似ていること
③ 記事の始まりがほぼ同じこと

この話題は元記事が 1 つであるため，信頼性は低いと判断できる．

元記事の多種性判定の第一段階として，同じ話題グループで元記事を参照する URL が異なっている数をカウントし，これを元記事の種類数とする．さらに第二段階として異なる URL の元記事を比較し，同じ文章であるか判定する．同じ文章のものはまとめ，最終的に文章的に異なる元記事の種類数を出す．これを多種性と見なす．具体例を以下に示す．

例：２つの元記事

・大リーグ 松井秀喜選手が現役引退へ　NHK ニュース

大リーグとプロ野球の日米を通じて強打者として活躍した松井秀喜選手が現役を引退することになりました．松井選手は 27 日にニューヨーク市内で記者会見を開いて現役引退を正式に表明することにしています．

・松井秀喜が現役引退！今日ＮＹで引退会見 - MLB ニュース ： nikkansports.com

プロ野球巨人や米大リーグのヤンキースで長距離打者として活躍し「ゴジラ」の愛称で親しまれた松井秀喜外野手（38）が現役引退を決意したことが 27 日，分かった．

まず，話題抽出部門においては，2 つの記事を比べた時，タイトル部中の名詞の一致により同じ話題として見なされ，グループ化される．一方，元記事の多種性判定部門では，元記事の文章やタイトル自体が異なるため，別記事と見なし，多種性があると判定される．

記事の同一性はタイトルと記事内容から以下のように判定する．なお記事はコピーされたものでも，ある程度変形，加工されて引用されている場合もあるので，完全な一致でなくても良いとしている．

第一段階としてタイトルで記事を比較する．タイトルの文字列が 15 文字以上一致する場合，同じ記事と見なす．第二段階は記事本体の文章で判断する．記事の文章を句読点で分割し，文の一覧を作る．2 つの記事の文の一覧を比較し，同一な文が 5 個以上現れれば，2 つの記事は同一であると判断する．

上記の処理の後，元記事の種類数をその話題が Tweet された数で割り，1000 を掛ける．これを話題グループの多種値と呼ぶ．グループに多種値を付けたデータベースを以下に示す．

group_id	title_count	topics_count	words	points
1	8	622	トンネル,事故,復旧,笹子,翌日,連続	12.86
2	9	634	これ,ネット,レコ,克久,大,意味,日本,服部,氏,現状,発言,話題	14.19
3	1	3	48,AKB,TweetBuzz,これ,レコード,前,大貫,委員,審査,意味,日本	333.33
4	6	297	4,1999,Chrome,Twitter,moon,since,ぷる,ソフトウェア,不正	20.2
5	1	107	10,PCMAX,SNS,以上,優良,出会い,創設,安心,実績,年	9.34
7	8	137	CHLionRagbaby,Twitpic,on,ちゃん,忘年会	58.39
8	7	326	au,スマ,ホ,復旧,社会,通信,障害	21.47
9	3	238	ジャンボ,以下,宝くじ,年末,当せん,番号,通り	12.61

図 3 多種値データベース

グループ番号と元記事の種類，話題の Tweet 数，代表話題の名詞リスト，多種値が格納されている．多種値が高いほど，話題の広がりに比例して，元記事の種類も多いことを示し，一方，多種値が低いほど，話題の広がり具合に反比例して，元記事の種類が少ないことを示している．

4. 評価

2012 年 11 月 1 日から 12 月 31 日の 2 ヶ月分のデータを用い検証を行った．話題性は 11 月 7 日の「Sony Music 楽曲が iTunes Store で配信開始」が最も高く，Tweet された数は 3956 件であった．元記事の種類も多くあり，多様値は 10.11 で，信頼性の高い話題と言える．逆に多種値の低い話題は 11 月 20 日の「再生 JAL の心意気/さかもと未明（漫画家）」という記事で Tweet された数は 968 件で多様値は 1.03 であった．ブログから派生した記事であり，実際にも信頼性は低いと考えられる．

5. まとめ

多種値の低さが実際の信頼性の低さを反映しているいくつかの事例は観測でき，信頼性判定の 1 つの目安とできそうであることが確認できた．しかし，元記事の同一性判定法など改良すべき点は多分にある．また，ここ数ヶ月で，アフィリエイト目的での Tweet が非常に増えてきたため，話題ではないアフィリエイトサイトが話題の上位に来てしまった．この対策も必要である．

参考文献

[1]宮部 真衣ら 「流言訂正技術に基づいた流言情報クラウドの提案」（第 4 回 楽天研究開発シンポジウム資料 2011）
[2]MdN Design Interactive: Web API PROGRAMMING
http://www.mdn.co.jp/webcre/API/vol101/
[3]Twitter Inc: Twitter Developers
https://dev.twitter.com/
[4]奈良先端科学技術大学：MeCab
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html