2N-4

アカウント情報の信頼度を考慮した Twitter アカウント検索システムの設計と実装

券田　孝晴 †　　西山 裕之 †

† 東京理科大学理工学研究科

1　はじめに

近年，Twitter[1] と呼ばれる Web サービスが注目を集めている．Twitter では，「つぶやき」と呼ばれる短い文章を投稿することで，ユーザは情報を発信したり，コミュニケーションを取ることができる．特に，「フォロー」というユーザ登録機能により，他のユーザのつぶやきを自由に閲覧できるという特徴がある．

Twitter は，その手軽さにより，様々な目的で利用されている．日記やメッセージ交換などの投稿は，コミュニケーション目的の SNS のような使い方といえる．また，ニュースや更新情報についての投稿は，情報発信を目的とした，ブログのような使い方である．さらに，最近ではフォロー機能を利用し，有用なつぶやきの収集を目的とした，情報収集としての使い方もある．

しかし，有用なつぶやき情報を収集するための問題が多く存在する．1 つは，Twitter が提供しているアカウント検索機能は，アカウント ID 検索のみであるため，膨大な Twitter のユーザ数を考慮すると，非力であること．また，アカウント検索に関する研究 [2] があるが，本研究では，有用なユーザを発見するという点を重視しているため，この研究とは目的が異なる．

2 つは，アカウント自体が信頼できる情報なのか確認する手段がほぼないことである．Twitter には，企業や団体などの「公式アカウント」が存在する．これら公式アカウントは，より情報発信意識が高いため，有用なつぶやきを投稿することが多い．しかし，2010 年 1 月時点で，認証済みアカウント機能以外に，アカウントが公式であることを確認する手段は存在しない．また，アカウントの信頼性を確認することができないことにより，なりすましによるつぶやきという新たな問題 [3] も発生している．

3 つは，有用なつぶやきを投稿しているかはそのアカウントのつぶやきを一つ一つ確認する必要があることである．さらに，コミュニケーション利用を目的としたつぶやきは大半が意味を成さないので，情報収集の際の大量なノイズとなっている．

そこで，本研究では，より有用なつぶやきを投稿するより公式のアカウントを発見することを目的とする．そのために，公式的信頼度，内容的信頼度という 2 つの信頼度を測る．公式的信頼度とは，アカウントがより公式のものなのかという信頼度を示し，内容的信頼度とは，つぶやきがより有用なものなのかという信頼度を示す．2 つの信頼度を算出することによって，アカウントに評価付けを行い，ユーザに提示する検索システムを構築し，実装する．

Twitter account search system in consideration of trust information.
Takaharu Kenda†, Hiroyuki Nishiyama†
†Graduate School of Sci.and Tech, Tokyo Univesity of Science

2　設計方針

本研究では，アカウントの信頼度を考慮した Twitter アカウント検索システムを構築するために，Google 検索の詳細検索 (site:twitter.com) を用いる．新しい検索システムを創るのではなく，検索の結果に信頼度などの評価付けを行い，情報提供（ソーティング）するシステムを構築する．本システムの機能は，ブラウザ処理とサーバ処理の 2 部で構成される．本システムの概要を図 1 に示す．次章よりこれらについて詳細に述べる．

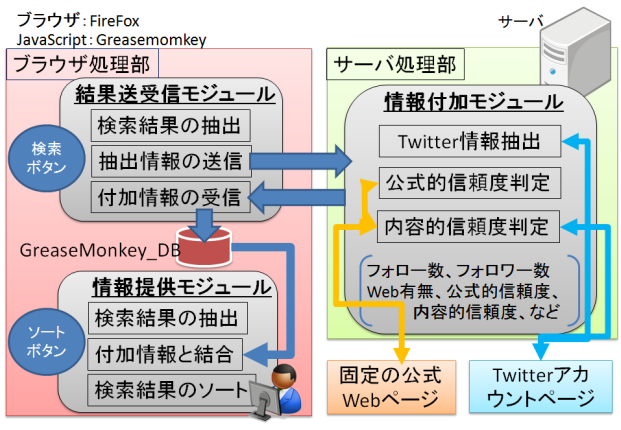

図 1: システム概要図

また，本システムは，ユーザにとって使いやすいインターフェースにするため，ユーザインタフェース用のブラウザには拡張性の高い Firefox を用い，Firefox アドオンである Greasemonkey を利用している．インターフェース画面を図 2 に示す．本システムを利用することにより，様々な条件（信頼度，フォロワー数など）で，アカウントを検索することで，公式のアカウ

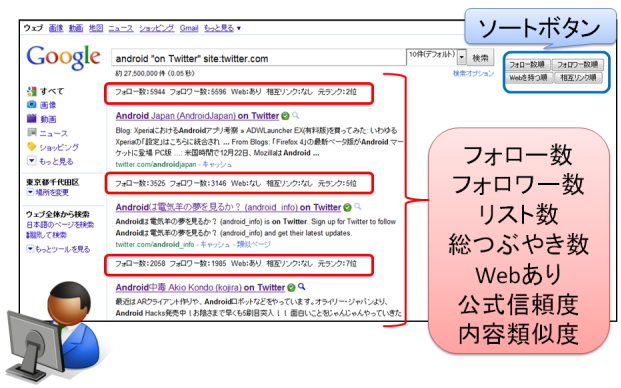

図 2: インターフェース画面

ントから，有用なつぶやきを投稿するアカウントなどを効率的に検索することが可能となる．

3 ブラウザ処理部

ブラウザ処理部では，2つのモジュールが動作している．この2つのモジュールは，ボタン操作によって独立に動作する．1つは，ブラウザの検索ボタンが押下されたときに動作する結果送受信モジュールである．これは，情報の橋渡しの役割を持つ．Google 検索の詳細検索を行い，任意の結果数に対して Twitter アカウントの URL を抽出する．それらをサーバ処理部に送信する．そして付加情報の受信を待ち，情報を受け取ったらそれらを GreaseMonkey の DB に書き込む．2つは，ブラウザのソートボタンのどれかが押下されたときに動作する情報提供モジュールである．これは，ユーザに情報を提示するインターフェースの役割を持つ．結果送受信モジュールと同様に，始めは Google 検索の詳細検索を行う．そして，検索結果と GreaseMonkey の DB に保存された付加情報を結合させ，ユーザが行いたいソート条件に従って結果を提示する．

4 サーバ処理部

4.1 Twitter 情報抽出

Twitter 情報取得では，Twitter アカウントページに対して各々情報抽出を行う．Twitter アカウントページの HTML を解析することにより，アカウントの全ての詳細項目，及び，そのアカウントのつぶやきを最大 100 件取得する．取得するアカウントの詳細項目は表 1 に，また，取得するつぶやきの詳細項目は表 2 に示す．

表 1: アカウントデータ詳細

ユーザ ID	スクリーン名	ユーザ名	現在地
Web（URL）	自己紹介文	認証済みアカウントの有無	
フォロー数	フォロワー数	リスト数	全ツイート数

表 2: つぶやきデータ詳細（※が付いたものは複数項目）

テキスト本文	つぶやき ID	作成日時	ソース
返信の有無	※つぶやきに付加された URL		※ハッシュタグ
※つぶやきに返信したユーザ ID，スクリーン名，ユーザ名			

4.2 公式的信頼度判定

公式アカウントならば，公式固有の Web ページを持っていると考えられる．これを基に，本研究では公式的信頼度を測る．以下に処理の流れを示す．

1. Twitter アカウントの詳細情報「Web」に記されているアドレスからその Twitter アカウントページにリンクが張られているか確認する
2. 「Web」に記されているアドレスが，アカウント詳細情報に関係した公式のページであるか確認する
3. アカウント詳細情報から抽出した名前（文字列）がどのくらい有名なのかを 3 段階で評価する．

1 と 2 の処理で，公式アカウントかどうかの判定を行っている．1 と 2 で確認できたならば，3 の処理を行い，1 から 3 の信頼度を算出する．また，それ以外の場合，信頼度は 0 と算出する．2 の処理は，まず，アカウント名と自己紹介文に共通する語句を抽出し，その語句と「公式」というキーワードを加えた 2 単語で AND 検索をする．検索は Google 検索エンジンを利用した．その検索の結果，トップ 3 にランクされる URL とアカウントの詳細情報「Web」に記されているアドレスとマッチング処理を行い，公式のページかどうか判定する．3 の処理は，2 の処理で抽出した語句が，官公庁・地方自治体や東証一部上場の有名企業か，それ以外の企業か，または一般の公式の団体か，を企業名データベースより判断し 3 段階評価している．

4.3 内容的信頼度判定

有用なつぶやきを投稿するアカウントならば，公式 Web ページに掲載されている確実な情報をつぶやきで発信していると考えられる．これを基に，本研究では，Twitter アカウントのつぶやきと公式 Web ページの内容（公式情報）の類似度を求め，それを内容的信頼度とする．しかし，つぶやきは 1 章で述べたように情報発信目的だけでなくコミュニケーション目的のつぶやきも多い．そこで，コミュニケーション目的のつぶやき（RT や@を含んだつぶやき）を除外した，つぶやきを対象とする．

内容の類似度は，ベクトル空間モデルによるコサイン相関値を用いる [4]．文書 D とすると，全てのつぶやきを D_i，全ての公式情報を D_j と表す．N 個の有効語 w の出現頻度による D_i の文書ベクトル $\vec{d_i}$ は，次式で定義される．

$$\vec{d_i} = [w_1(D_i), w_2(D_i), \cdots, w_n(D_i)] \tag{1}$$

取得した全てのつぶやきに形態素解析を行う．そして，名詞，固有名詞を取り出し，有効語とする．

従って，2 つの文書 D_i と D_j の類似度 $S(D_i, D_j)$ は次式で表す．

$$S(D_i, D_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{\|\vec{d_i}\| \cdot \|\vec{d_j}\|} \tag{2}$$

5 おわりに

本研究では，Twitter アカウントにおける 2 つの信頼度，公式的信頼度と内容的信頼度を算出し，アカウントに評価付けを行い，ユーザに提示する検索システムを構築した．2 つの信頼度によって，有用なアカウントつまり公式のアカウントや有用なつぶやきを投稿するアカウントを検索することが可能になった．これにより，情報収集のための Twitter における問題点を解決した．今後の展望として，つぶやき自体の検索や Twiiter 特有の機能であるフォロー関係を利用したアカウント検索などが考えられる．

参考文献

[1] Twitter：http://twitter.com/ (2006)

[2] J. J. Jianshu Weng, Ee-peng Lim and Q. He：”Twitterrank: Finding topic-sensitive in uential twitterers”, Web Information and Data Management (2010).

[3] 松坂志：Twitter なりすまし問題と対策，独立行政法人情報処理推進機構（IPA）

[4] R,Baeza-Yates. and B,Ribeiro-Neto：Modern Information Retrieval. ACM Press,(1999).