

異種サービス間における信頼性の高いユーザおよび 情報検索コミュニケーション技術

平成 27 年 4 月 23 日受付

河 合 由起子*

要 旨

本研究では、異なるメディアとなる Twitter と Web を利用しているユーザどうしが、メディアの違いを気にすることなくリアルタイムにコミュニケーション可能なシステムの構築を目指す。具体的には、ツイートの発信場所とページの内容（場所名）を抽出することで、ツイートと Web ページを対応付ける。ツイートと対応付けることで、ページ上にツイートを提示できる。これにより、Web 閲覧ユーザは、ページを閲覧しながら内容に関わる場所で発信されたその場所に関するツイートを随時閲覧できる。また、Web 閲覧ユーザは発信したツイートに対して、メッセージを返信（発信）できる。発信されたメッセージは、ツイートユーザに対して、ツイートメッセージとして提示される。本論文では、位置情報に基づいたツイートと Web ページ関連付け手法ならびに Twitter と Web のシームレスなリアルタイム通信システムの構築手法について述べ、実装システムを検証する。

キーワード：検索，コミュニケーション，ビッグデータ分析，SNS 分析，Web マイニング

1. はじめに

本研究は、実空間と仮想空間の位置情報に基づき、各ユーザにとって信頼ある人および情報とを瞬時に同時発見し、情報を介して即時的に情報交換可能な検索・通信技術の研究開発を目的とする。本研究のチャレンジは、実空間と仮想空間の位置情報に基づくマッピングならびに信頼性の高い人と情報の同時発見、さらに携帯端末と PC 利用ユーザ間の情報を介したリアルタイムコミュニケーションの実現である。今年度は、情報検索・閲覧・交換時のユーザ行動のデータ実時間処理機構を研究開発した。具体的には、まず、SNS として Twitter を対象とし、日本および米国の位置情報付きツイートストリームデータを取得管理する。次に、ツイートデータと Web ページの関連付け手法を確立し、実装、評価検証を行い、実空間ユーザと仮想空間ユーザの関係性を発見する。更に、それらユーザ間のリアルタイム通信を実現した。

* 京都産業大学コンピューター理工学部

2. システム概要

本研究は、場所に関するツイート情報の取得ならびに、ツイートユーザと Web 閲覧ユーザが場所情報に基づきリアルタイムに通信可能なシステムの構築を目指す。

ツイートを発信すると、ツイート閲覧者だけでなく、ツイートの内容と発生場所の位置情報に基づき、関連する Web ページを検出し、それらを閲覧しているユーザのページ上にそのツイートがリアルタイムに提示される。Web 閲覧ユーザにとってはそれら提示されたツイートを閲覧することで、場所に関する現状把握の支援になる。各ツイートに対して返信する場合は、該当するツイートを選択することで、個別に返信できる。ただし、ツイートユーザは本サービスをフォローしている必要がある。

一方、Web 閲覧ユーザが本システムの入力ボックスを用いて情報発信すると、ページを閲覧している他の Web 閲覧ユーザにメッセージが送信される。また、本サービスをフォローしているツイートユーザに対してもツイートとして提示できる。ツイートによる返信は、前者の問合せ同様にツイートだけでなく Web ページ上にも提示され、全てのページ閲覧ユーザはこれらの問合せと返信を閲覧できる。なお、全てのメッセージは WebSocket サーバを経由するため匿名性が保たれる。具体事例として、ルーブル美術館にいるユーザがツイートを発信した場合に、そのツイートがルーブル美術館のページと関連付けられ、Web ブラウザに提示される。Web 閲覧ユーザは混雑具合やルーブル美術館の感想など状況に関する問合せができ、一方でツイートユーザはルーブル美術館の開館時間など、その時その場では確認しにくい情報に関してツイートによる問合せが可能になる。

2.1 位置情報に基づくツイートと Web ページ間リアルタイムメッセージ通信

本研究では、Web 閲覧ユーザとツイートユーザとを Web ページとツイートを通してリアルタイム通信可能にするため、リアルタイムに送信されるツイート（以下ストリーミングツイートする）ならびに Web ユーザがアクセスしている Web ページの URL を取得する。サーバはツイートユーザが発信したストリーミングツイートを取得し、位置情報に基づいて関連するページを取得し、対応付けの管理を行う。取得した関連ページに Web 閲覧ユーザがアクセスすると、対応するツイートを抽出し、ブラウザへ送信および提示する。なお、ツイートユーザは本サーバからメッセージを受信する際には、本サービスのアカウントのフォロワーとなっている必要があり、Web 閲覧ユーザは提案システムとなるアドオンを用いる必要がある。

Web 閲覧ユーザがアドオンの入力ボックスにメッセージを入力すると、サーバが受信し、同じページを閲覧しているユーザのブラウザへ送信する。また、Web ページに提示されている各ツイートにのみ直接返信も可能である。直接任意のツイートに送信する場合、サーバはブラウザからメッセージを受信し、ツイートしたツイートユーザ（フォロワー）へ送信する。全ての送受信はサーバを介するため、匿名性が保たれる。

2.2 ストリーミングツイートデータ取得

本論文では、位置情報に基づく問合せを目的としており、ページとツイートを位置情報に基づき関連付ける。まず、指定地域から重複を除いた緯度経度情報を含むストリーミングツイートを The Streaming APIs を用いて取得する。指定地域は、1度以上異なる南西および北東を指定することで、その2点に囲まれた矩形領域のストリーミングツイートを取得できる。次に、取得したストリーミングツイートの緯度経度情報から、Google Place APIを用いて、半径 d m の場所名を取得した。評価実験では、取得した場所名は関連する Web ページ取得の際に検索キーワードとして用いられることと、ツイート発信ユーザの移動も考慮し、 $d=5$ とした。また、ツイート内容を形態素解析し、名詞となる単語を取得する。以上より、ツイートユーザ id, アイコン画像 URL, 緯度, 経度, 場所名, ツイート内容, 単語集合, 取得時刻を一定時間管理する。

2.3 ツイートの緯度経度と内容に基づくツイート選別

前節より取得したストリーミングツイートに対して位置情報に基づいた内容判定を行い、ページと関連付ける。ツイートが発信された場所名と関連するかをツイートの内容から判定することで、ツイート発生場所と関係性の低いツイートの除去を行う。位置情報に基づいたツイート内容判定法は、一定範囲内の一定時間のツイートに多く出現する単語は関連性が高いと考え、場所名に対する特徴語として抽出する。この特徴語を多く含むツイートを場所名に関連するツイートとして選択する。まず、取得したツイート t の位置情報より、半径 d 内に存在する一定時間内のツイート n 個を取得する。次に、下記の式によりツイートの重要度を算出する。まず、ツイート t に出現する各単語 i のツイートに出現する頻度を抽出し、その平均値を算出する。また、特徴的な単語が出現しても単語数が多い場合は、ツイートの重要度が低下するため、シグモイド関数を用いることで、出現頻度の高い単語には、さらに重要度の重みを増やす方法を取ることにした。最後に、閾値以上のツイート t を位置情報に基づいたツイートとする。

2.4 Web ページの場所名抽出と場所名に基づく Web ページとツイートの対応付け

まず、Web 閲覧ユーザの閲覧している Web ページの URL を取得し、その Web ページのスニペットを取得する。次に、スニペットから出現頻度の高い単語を特徴語として抽出する。また、形態素解析よりその特徴語の中から地名を判別し、該当する単語をそのページの場所名とする。尚、複数地名が抽出された場合は全てを場所名とする。

前節より、ツイートユーザの位置情報付きツイートを The Streaming APIs を用いて取得し、緯度経度から場所名を取得した。さらに、場所に関連するツイートを選別した。ユーザが Web ページを閲覧すると、場所名から関連するツイートを検索し、Web 閲覧ユーザに提示する。ツイートユーザには、緯度経度情報から場所名を抽出し、その場所名と一致する Web ページを対応づける。なお、DB には取得したツイートおよび抽出した場所名を格納する。これらのツイートと Web ページを場

所名に基づき、対応付ける。

2.5 リアルタイム双方向通信

リアルタイム問い合わせシステムを構築する上で、Ajax や Comet, WebSocket といった様々な双方向通信手法が存在する。先行研究として、これまで我々は Web サーバと Web ブラウザ間の通信のための双方向通信として Ajax や Comet を用いてきたが、本研究では、より通信ロスの少ない WebSocket を用いる。ツイートユーザから Web 閲覧ユーザへの配信では、ツイートユーザの位置情報付きツイートを The Streaming APIs を用いて取得し、そのツイートの緯度経度情報から場所名の付与を行い、DB にツイートと場所名を格納する。そして、Web 閲覧ユーザが閲覧している Web ページの場所名から、場所名とマッチするツイートを DB に問い合わせ、該当するツイート情報を Web 閲覧ユーザに提示する。Web 閲覧ユーザからツイートユーザへの配信では、3.4 節より、Web 閲覧ユーザは Web ページ上に提示されたツイートに対してサーバを経由し、ツイートとしてメッセージを送信することが可能である。

3. 実装

本研究では、Web 閲覧ユーザとツイートユーザとを Web ページとツイートを通してリアルタイム通信可能なシステムの構築を目的としている。本章では、実装による評価としてプロトタイプを構築し、リアルタイム通信機能の評価ならびにページに提示されるツイートの評価実験を行う。

本プロトタイプでは、サーバは、Apache httpd 2.4, java, php5.5 を用いた。クライアント側は、Javascript を使い、Firefox 23.0.1 上で動作確認した。なお、2013 年 9 月 3 日から同年 12 月 14 日までに The Streaming APIs で日本全国のツイートデータを取得した。今回の検証では、上記のツイートデータから平日の 5 日間のツイートデータを使用した。なお、比較対象として、本手法で用いるデータと同じツイートにおいて、各々の場所名でハッシュタグ検索を行った場合の結果を用いる。ストリーミングツイート取得は、指定地域から重複を除いた緯度経度情報を含むストリーミングツイートを The Streaming APIs version 1.1 を用いて取得した。指定地域は、1 度以上異なる南西および北東を指定することで囲まれた矩形領域のストリーミングツイートを取得できる。評価実験では、関東地方を対象とし、南西 132.2, 29.9, 北東 146.1, 46.20 とした。取得したストリーミングツイートの緯度経度から、Google Place API version 3 を用いて、場所名を取得した。評価実験では、取得した場所名は関連する Web ページ取得の際の検索キーワードとして用いられることと、ツイート発信ユーザの移動を考慮して、事前実験より半径 5m 以内のものとした。

サーバとクライアント間のメッセージ通信は、ツイートユーザと Twitter サーバ、Twitter サーバと本システムとし、Twitter サーバと本システム間で通信することで、ツイートユーザに対する送受信のインターフェースは既存の Twitter サービスが利用可能となる。Web 閲覧ユーザと本システム間は WebSocket 通信を行い、アドオンを用いた質問応答インターフェースを構築した。

4. 考察

平日の5日間でツイートの発生位置から半径 d が 100m ~ 1km において、東京駅では 600m, その他の東京スカイツリー, 東京ディズニーランド, 羽田空港では 1km の各範囲内において, F 値が最も高い結果であった。このことより, 平日において, 人が密集している公共施設では, 本手法によるツイート選別は有効であることが確認できた。

また, F 値の低かった取得範囲に関しては, 提案手法でツイートを選別する場合, ツイートにおける特徴語として, 名詞となる単語のみを対象としており, 形容詞や顔文字など感想のツイートを対象としていないことが影響していたと考えられる。また, 対象ページに関しては, 今回は場所名を検索キーワードとした検索結果のページを対象としているが, ツイートを店舗ごとの小規模施設に細かく分類するため, 店舗などが密集している場所では, 周辺の店舗も対象とする必要があると考えられる。周辺店舗の名称の抽出は, 実験結果のツイートを検証した結果, ツイートから抽出された出現頻度の高い単語の利用が考えられるが, 今後, より多くの施設や駅等を密接度で分類し, 検証する必要がある。

5. 関連研究

近年, Twitter をテキストマイニングの対象とした研究は活発に行われており, Twitter に投稿されたツイートを分析することでイベントの検出や位置情報の取得を試みた研究も数多くある。

Arakawa ら [1] は位置情報ツイートから位置依存性の高い文字列を抽出する手法を述べている。位置情報ツイートから得たエリアを 100 キロ四方のグリッドに分割し, それぞれのグリッド内のツイート含有率を計算し, ツイート含有率がある閾値を超えたエリアを最終的に 1 キロ四方のグリッドまで走査することにより, 1 つのキーワードに対して複数の位置依存性を抽出することができる。この研究では, 位置情報とツイートのコンテンツを対応付けている。本研究でも, 位置情報とツイートのコンテンツ内容を関連付けているが, こちらは, 特定の場所や建造物を中心とした位置の重要性の高い文字列の抽出を行っている。また, Yamamoto らの Twitter に投稿された実生活情報から有用性の高いものを抽出し局面に応じた記事をユーザに提示するシステム [2][3] や, ツイートから地震や台風などのイベントの検出を試みた研究として榊らの研究 [4] がある。Twitter のタイムラインを監視しておくことでリアルタイムでイベントの検出を行い, 高い精度を得られた。これらの研究では, コンテンツベースでの抽出を行っているが, 本研究では, コンテンツ内容と緯度経度情報の関連付けを行っている。

また, Lee らの研究 [5] では, イベント検出対象となる地域をいくつかの小さな地域に分割し, ツイート数, ユーザ数, ユーザの移動状況の 3 点を分析し, その地域の通常時の状態を推測する。そして, 通常の状態とは異なった多くのツイートが投稿された場合, イベントが発生したとみなしている。

Nichols らの研究 [6] は, ツイートのコンテンツ内容の変化に注目しており, 更新の量の急増などでイベント内の重要な瞬間の識別を行う。Ribeiro らの研究 [7] では, ツイート内容を識別し緯度経度か

ら区域でのイベント発生を検出する。本研究では、ツイートの内容に着目し、特定の単語の出現頻度が高くなれば、イベントが発生したと見なしている。

位置情報付き画像ツイートをを用いてイベント検出を試みた研究として、Nakaji ら [8] はあらかじめ特定のキーワードや期間を設け、位置情報付きツイートを収集し、解析することで画像付きのイベント検出を試みた。Kaneko ら [9] はイベントのキーワードをシステムにより自動的に抽出することで多くのイベントを抽出することで未知のイベントのキーワードを得られるようにし、キーワードを用いて収集した画像を解析することで、ユーザが知らないイベントでも画像により視覚的にとらえることができるようにした。これらの研究は、それぞれ位置情報ベースとコンテンツベースで別々に取り扱っているが、本研究では、この2つを同時に取り扱う。Takemura ら [10] は、Twitter ユーザを、広く一般のユーザが興味を示す情報を発信するのか、一部のユーザのみが興味を示す情報を発信するのかの範囲を示すため、対象局所性と定義される指標を用いた分類を行う手法を提案している。本研究では、Twitter の位置情報と内容に基づいて発信されたツイートが発信された場所に関連しているかを判別する。

6. まとめ

本研究では、実空間で SNS 等を用いて情報発信しているツイートユーザと、異なる場所で関連する Web ページを閲覧しているユーザに対して、実空間の位置と Web コンテンツの内容とを対応付けることで、異なる場所で異なるメディアを利用しているユーザ間のリアルタイム双方向通信の実現を目指し、位置に基づくリアルタイム問合せシステムを構築した。評価実験ではプロトタイプを構築し、ツイートユーザ密度の異なる 4 地点において、各地点に対するツイートとページの関連性を検証した。実験結果より、東京スカイツリーや空港、東京駅等の密集地において、本手法が有効であることを確認できた。今回はシステムにおいての前実験として、Web 閲覧ユーザに適したツイートが提示されるかの検証を行った。今後は、実際にシステムを被験者に操作してもらい、情報取得に有用であったかといった、システムが Web 閲覧ユーザ及びツイートユーザに対して有効であるかの検証を行う予定である。

また、Web 閲覧ユーザの Web 上に必要な情報についてのアンケートを行った。アンケートの結果より、その場所で現在行われているイベント情報、その場所の今後の天気の提示やその場所へのアクセス情報といったものが、Web 閲覧ユーザにとって有益であることが確認できた。今後このような集約情報の提示の実装も行う予定である。

参考文献

- [1] Yutaka Arakawa, Shigeaki Tagashira and Akira Fukuda: Relationship Analysis between User's Contexts and Real InputWords through Twitter, IEEE Globecom 2010 Workshop on Ubiquitous Computing and Networks (UbiCoNet 2010), pp.1813-1817, 2010.

- [2] Shuhei Yamamoto and Tetsuji Satoh: Two Phase Extraction Method for Multi-label Classification of Real Life Tweets, In Proc. of the 15th International Conference on Information Integration and Web-based Applications Ψ & Services (iiWAS 2013), pp. 16-25, 2013.
- [3] Shuhei Yamamoto and Tetsuji Satoh: Two Phase Extraction Method for Extracting Real Life Tweets using LDA, In Proc. of the 15th Asia-Pacific Web Conference (APWeb 2013), Lecture Notes in Computer Science 7808, pp. 340-347, 2013.
- [4] Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, In Proc. of the International World Wide Web Conference (WWW 2010), pp. 851-860, 2010.
- [5] Ryong Lee and Kazutoshi Sumiya: Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection, In Proc. of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN 2010), pp. 1-10, 2010.
- [6] Jeffrey Nichols, Jalal Mahmud and Clemens Drews: Summarizing Sporting Events Using Twitter. In Proc. of the 2012 ACM International Conference on Intelligent User Interfaces (IUI 2012), pp. 189-198, 2012.
- [7] S. S. Ribeiro, C. A. Davis, D. R. R. Oliveira, W. Meira, T. S. Goncalves and G. L. Pappa: Traffic Observatory: A System to Detect and Locate Traffic Events and Conditions Using Twitter. In Proc. of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN 2012), pp. 5-11, 2012.
- [8] Yusuke Nakaji and Keiji Yanai: Visualization of Real World Events with Geotagged Tweet Photos. In Proc. of IEEE ICME Workshop on Social Media Computing (SMC 2012), pp. 272-277, 2012.
- [9] Takamu Kaneko and Keiji Yanai: Visual Event Mining from Geo-tweet Photos, IEEE ICME Workshop on Social Multimedia Research (SMMR 2013), pp. 1-6, 2013.
- [10] 山本 祐輔, 浅井 洋樹, 上田 高德, 秋岡 明香, 山名 早人: テレビ番組に対する意見をもつ Twitter ユーザのリアルタイム検出, 第5回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2013), C1-4, 2013.

A System of Search and Communication for high Credible User and Information between Web and SNS

Yukiko KAWAI

Abstract

We developed a system to support simultaneous communication between microbloggers (Twitter users) and Web users in real time through both the contents of microblogs (tweets) and Web pages. The system provides a question answering interface attached to Web pages, which allows Web users to chat with each other while browsing the same Web pages, or Twitter users whose tweets are presented with Web pages. To achieve this, the system extracts relationship between heterogeneous media on tweets and Web pages by generating queries based on posted information. Thus, the system can effectively present tweets from the Twitter users to help the Web users easily grasp the current situation in real time, and it also can effectively present messages from Web users to help Twitter users immediately obtain useful information or knowledge. In this paper, we discuss our prototype system and evaluate its effectiveness.

Keywords : Search, communication, Big data analysis, SNS analysis, Web mining