ウェブデータからの行為抽出エンジンの開発
—「pingpong: 動く地図」プロジェクト—

1. 背景

Mixi、Facebook、Twitter といったソーシャルネットワークサービスやブログやマイクロブログといったコミュニケーション手段の普及により、誰でも気軽に情報を公開、共有できるようになっている近年、さまざまなユーザが、製品やイベントに対しての嗜好や意見、日常の行動などについて発信している。その結果、年々企業側が提供する製品に関する情報よりもこれらのコミュニケーション手段を通じてユーザが発信するコンテンツ（User Generated Content、以下 UGC）が、消費者の購買行動に影響を与えるようになってきていると言われている。このような背景の中 UGC から自動的に注目の話題を抽出したり、製品やブランドに関する評判を分析するサイトやシステムが開発、公開されている。

このようなさまざまなコミュニケーション手段を通じて発信される UGC から抽出可能な情報は、製品や出来事に対するユーザの意見、評判といったものに限らない。Twitter のサイトにおいて「いまなにしてる？」という問いかけに代表されるように、位置情報とともに UGC には多くの行為を含むデータが蓄積されている。これらのデータを収集し、さまざまな実空間での人々の行為が可視化することで、空間のレイアウトの持つ潜在的なアフォーダンスが顕在化され、新しいコミュニケーションやサービスが創出することが期待されている。

2. 目的

本プロジェクトでは、「ウェブデータからの行為抽出エンジンの開発」というテーマのもと「動く地図」システムの開発を行う。「動く地図」とは、人々が携帯端末で発信する情報をリアルタイムに取得し、地図と組み合わせることで「いま・ここ」で行われている行為を可視化するシステムである。このようなさまざま実空間での人々の行為を可視化し、店舗や施設の情報と結びつけることにより、空間の持つ潜在的な人とのインタラクションを顕在化され、新たなコミュニケーションおよびサービスを創出することを目的としている。

3. 開発の内容

「行為抽出エンジン：動く地図」システムを作成するにあたって、主に以下の 4 つの機能の実装を行った。システムのスナップショットを図 1 に示す。

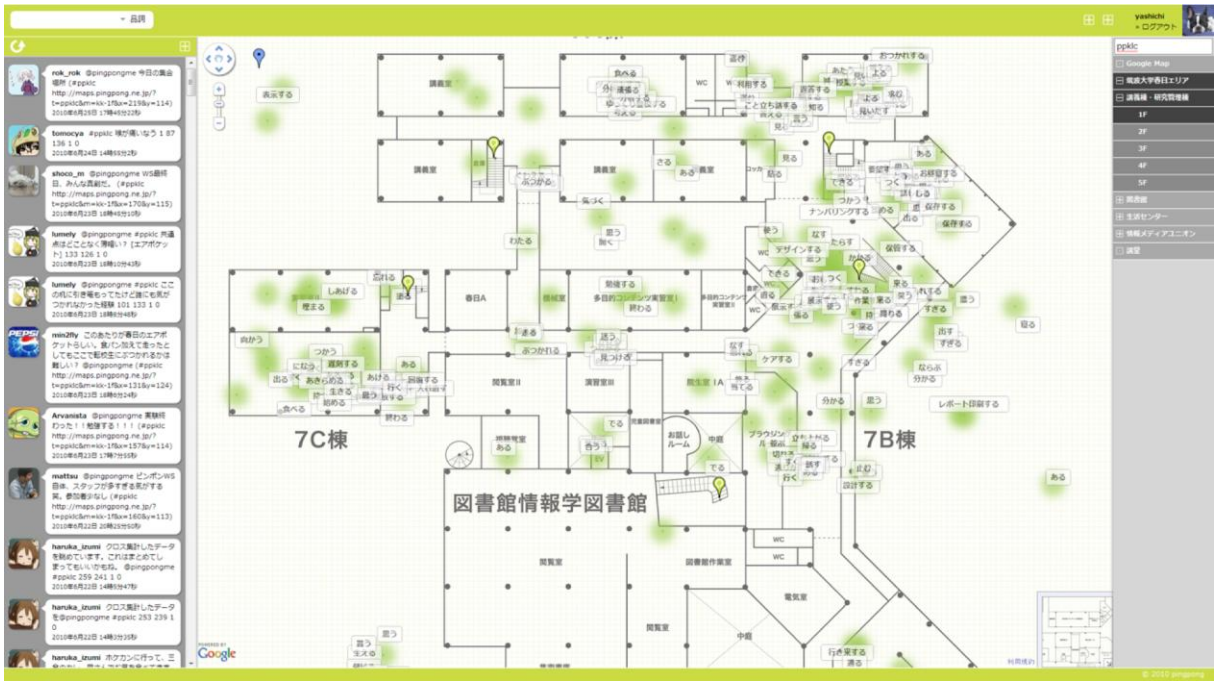

図1:「動く地図」システムのスナップショット

(1) Twitter からの行為情報の抽出

ウェブから Twitter が提供する API を通して、対象とする範囲の Tweet を収集し、自然言語処理を通じてテキストから動詞・目的語・主語のセットを一つの行為として抽出するエンジンの開発を行った。

(2) 建物内の行為情報を位置情報と共に取得するアプリケーションの開発

ユーザが携帯端末の GPS 機能を利用してテキストを投稿することにより、Twitter に投稿される情報からも位置情報を取得することは可能である。しかし、建物内の詳細な位置情報と共に、人々の行為を抽出することはできない。そこで、本プロジェクトでは、建物内においても詳細な位置情報と共に Twitter に投稿できるための専用アプリケーションの開発を行った。

(3) Twitter のハッシュタグ[1]と位置情報の対応付け

(2)で開発したアプリケーションによって投稿される Tweet は,ある特定の建物情報とハッシュタグを通じて関連付けられる仕組みになっている。この仕組みを使うことにより、GPS 付きの Tweet とシームレスに地図上に可視化することが可能となっている。

(4) 行為情報を地図上で可視化

(1)で収集し抽出した行為データの動詞部分をタグ化し、Google Map API を用い

[1] http://ja.wikipedia.org/wiki/Twitter のハッシュタグの欄を参照.

て可視化している。テキストから動詞部分のみを可視化することにより、数多くのTweet がある地点に投稿されても、煩雑になりすぎることなく地図上に表示することが可能となっている。

4. 従来の技術(または機能)との相違

本プロジェクトではUGCから、大規模データから人々の行為を抽出するエンジンの開発を行った。本プロジェクトでは，UGC の中特に，Twitter に代表されるマイクロブログを対象データとする．また人々の行為は、(主語，述語，目的語）の３つの要素として定義した。Twitterが提供するAPI を通じて投稿記事（以下，tweet）を取得し、自然言語処理技術を用いて各tweetから３つの要素を１ つの行為として抽出している。例えば，「インフルエンザにかかった」という tweet からは、(私，かかった，インフルエンザに）が行為として抽出される。

ある文章からの主語（以下，S)、述語（以下，V)、目的語（以下，O）の抽出に関する技術は自然言語処理の分野においてこれまで長年研究されているが、本研究の対象とするマイクロブログはその性質が次の２点で大きく異なる．

【非文法的／断片的な文章】

従来の自然言語処理の研究分野では、論文やニュースなど比較的フォーマルな文章が扱われてきた。一方、マイクロブログに投稿される文章は、文というよりも短い名詞句の連続という形で記述されることが多く、さらには，誤字など非文法的な表現がしばしば含まれる。このような非文法的かつ断片化されたテキストを扱う際には構文解析など深い処理が有効でない場合も多い。このようなテキストからロバストに行為を抽出するには、新たな抽出手法が必要不可欠であり、本プロジェクトではPMI（Pointwise Mutual Information）という指標を用いてこの解決を行っている。

【リアルタイム性】

従来のウェブページやブログに比べ、Twitterのようなマイクロブログはリアルタイム性が大きな特徴となっている。リアルタイムに投稿される情報を収集し解析することは、大量の情報をどのようにロバストに処理するかという技術的な課題と共に、人々が何に対して、どのように反応しているのかを眺める上で大変重要な要素となっている。本プロジェクトでは、Twitterのハッシュタグと位置情報を結びつけることにより、ユーザの興味の対象となるデータを出来るだけ扱うと共に、リアルタイムに扱う情報の量を制限することにより、この問題に対処している。

5. 期待される効果

開発した「動く地図」を利用することにより，以下のような一般へのサービスと分野の創出が期待できる。

（１） ショッピングモールなどの商業施設で本プロジェクトのシステムを導入することにより、例えば、非計画購買者（何を購入しようか，あらかじめ決めていな

い購買者）を動機づけ、購買行動を促進できる。
（２）　「動く地図」の持つ時間的コンテキストの活用により、利用者の行為に反応したリアルタイム性，即効性のある情報を入手できる。
（３）　地図に紐づいた情報の取得・提供により、行ったことのない空間の情報も、高い精度で手に入れることができる。

6. 普及（または活用）の見通し

（開発成果に関する利用者の具体的なイメージ[例えば、利用者数など]を可能な限り定量的に記載）

　開発した「動く地図」をウェブを通して公開していく。最終的には、ウェブ系の大企業と連携し、行為抽出エンジンを大規模に運用し、商業施設における空間レイアウト設計や、行動予測、実空間と結びついた広告サービスのアプリケーションとして提供することを目標としており、目標が達成されたときには多くのユーザに使われるシステムとして、情報産業への影響を与えることが可能であると考える。
　しかし、さまざまなウェブ上でのサービスが日々立ち上げられている今日、多くのユーザの目を集めることは難しくなってきており、実際にどのくらいのユーザに使われるかは実際にサービスを公開してからでないと分からない部分も多い。そこで、「動く地図」を、誰もが利用可能なサービスとして公開して発信・普及に努めると共に、大学や図書館といった公共空間への「動く地図」を導入するためのきっかけとして、ワークショップを活用していく予定である。

7. クリエータ名（所属）

岡　瑞起（東京大学　知の構造化センター）

（参考）関連URL
pingpong プロジェクト URL：http://www.pingpong.ne.jp/