Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

GLoBEs: ソーシャルメディアにおける位置情報を用いたグループ検出機構

蛭田 慎也[†1] 間 博人[†2]
森 雅智[†2] 徳田 英幸[†2,†3]

近年，位置情報取得機能付き端末の普及により，ソーシャルメディアにおいて位置情報を含んだ発言を気軽に共有できるようになった．ソーシャルメディアにおいて，興味を持つ場所や趣味が似ているユーザを検出し，広告や情報配信を行う需要は大きく，今後ますます拡大すると考えられる．しかし，位置情報の取得方法やプライバシの問題などがあり，位置情報を用いてユーザ間の関係を推定する手法は確立されていない．そこで本研究では，これらの課題を解決するため，ソーシャルメディアにおける位置情報付きの発言を利用した類似ユーザの推定手法 GLoBEs (Grouping Algorithm Based on Location of Micro-Blog Entries) を提案する．本稿において，ソーシャルメディアの発言を取得してユーザ間の類似度を計算するモジュールを実装し，評価を行った．本研究は，既存の類似ユーザ推定手法が用いていた独自ネットワークの構築や特殊な端末を必要としないため，既存のソーシャルメディアにおける全世界のユーザ情報を活用することが可能になる．

GLoBEs: A Spatio-Temporal Grouping Algorithm using Micro-Blog Entries

Shinya Hiruta,[†1] Hiroto Aida,[†2] Masato Mori[†2]
and Hideyuki Tokuda[†2,†3]

Recently, social media such as Twitter and Facebook attract public attention and many users can now easily share the information. Furthermore, spread of devices having location acquisition technologies makes it possible to acquire coordinates at many places. From such a background, location sharing applications (LSA) on social networking service (SNS) become widely accepted. In this paper, We propose GLoBEs (Grouping Algorithm Based on Location of Micro-Blog Entries) to measure similarities among users by mining location of micro-blog entries. In order to evaluate GLoBEs, we compared the measured similarity and actual user's relationships on Twitter.

1. はじめに

近年，位置情報共有アプリケーションが普及している．位置情報共有アプリケーションとは，ユーザの所持している端末から位置情報を取得し，その情報をユーザ同士で共有したり，ユーザにサービスを提供したりするアプリケーションである．このような位置情報共有アプリケーションが普及した背景の一つには，位置情報取得機能付きの携帯電話や，iPhone[1)] や Android[2)] などのスマートフォンの普及がある．また，近年 Facebook[3)] や Twitter[4)] などのソーシャルメディアの利用者も急速に増加している．これらのソーシャルメディアの流行により，情報を公開し，共有するという行為をユーザがより気軽に行うことができるようになった．このような傾向は，位置情報共有アプリケーションにも大きな影響を与えている．近年の位置情報共有アプリケーションの多くが，ネットワーク上での人と人とのつながりを重視し，位置情報を他人と共有する行為に重点を置いている．

ソーシャルメディアにおいては，ユーザ同士の関係からなるソーシャルキャピタル が重要となる．ユーザにとって，同じような興味や趣味，生活範囲を持つユーザの情報は，自分にとっても重要である可能性が高いと考えられるからである．一見，大衆が無秩序に情報を垂れ流しているようでも，類似している情報を適切にまとめることで，情報の有用性は向上する．したがって，ソーシャルメディアにおいて興味や趣味の似ているユーザを推定することが，今後ますます重要になっていくと考える．このような，興味や行動パターンが似ているユーザのことを，本稿では「類似ユーザ」と定義する．また，ソーシャルメディアにおけるオンライン上でのつながりの数と，ユーザが訪れた物理的な位置には正の相関関係があるという既存研究がある[5)]．実世界でのユーザの位置情報を取得して行動パターンや趣味が似ているユーザを見つけ出し，オンラインでのソーシャルネットワークへフィードバックすることができれば，同じ地域に住むユーザのグループに対して地域限定の広告やクーポンを配信するなどの新たなサービスの可能性が生まれると考えられる．

既存の類似ユーザ推定手法としては，ネットワーク上での人と人のつながりであるソー

†1 慶應義塾大学 総合政策学部
Faculty of Policy Management, Keio University
†2 慶應義塾大学大学院 政策・メディア研究科
Graduate School of Media and Governance, Keio University
†3 慶應義塾大学 環境情報学部
Faculty of Environment and Information Studies, Keio University

Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

シャルグラフを用いて類似ユーザを推定するリンクマイニング が一般的である．しかし，既存の手法はオンライン上のつながりのみを判断基準としているため，実際の行動範囲や興味を持つ場所が似ているユーザは検出できないという問題がある．一方，既存研究では GPS ロガーを用いてユーザの移動軌跡を常に記録し分析する手法 により，実世界の行動パターンを基準に類似ユーザを推定する手法が提案されている．しかし，ユーザの意図と関係なく常に現在位置を記録してしまうため，プライバシが全く考慮されていない．また，特殊な端末をユーザに持ち歩かせ，都度データを回収するのは非常に手間がかかり，現実的に適用可能ではない．

このような問題を解決するため，本研究では Twitter などのソーシャルメディアにユーザが自ら公開した位置情報を用いて類似ユーザを推定する手法を考案する．従来の手法では，連続した膨大な GPS ログからユーザが立ち寄った場所や興味を持った場所を推定する必要があった．しかし，ソーシャルメディアにおいてはユーザが自ら興味を持った場所の位置情報や感想などを発言している．このような発言情報を用いれば，移動軌跡のみを用いる手法に比べてユーザの行動パターンや興味を正確に推定できると考えられる．また，ユーザが自らの意志で公開した位置情報のみを取得するため，常に位置情報を取得する手法と比べてユーザのプライバシを侵害するおそれが少なくなる．解析したグループ情報においては，公開された情報のみを用いているとしてもプライバシ保護の観点からは繊細な情報のため，適切な相手にのみ公開する運用を行うことが望ましい．

本研究では，実世界における位置情報に基づいた類似ユーザを検出し，ソーシャルメディアにおいて新たなつながりを提示することを目指す．以下に，本稿の構成を述べる．第 2 節では既存の類似ユーザ検出手法と問題点について述べ，第 3 節では本稿で提案する GLoBEs の設計を述べる．第 4 節では GLoBEs の実装について述べ，第 5 節では各モジュールの評価を行う．そして，第 6 節で本稿のまとめを述べる．

2. 類似ユーザ検出における既存研究

本節では，類似ユーザ検出における既存研究について分類し，GLoBEs と比較する．位置情報共有に関する研究は多く行われており[6]，ソーシャルメディアにおけるオンライン上でのつながりの数と，ユーザが訪れた物理的な位置には正の相関関係があると報告されている[5]．

Twitter[4] や Facebook[3] などのソーシャルメディアにおいては，ユーザ間のつながりを示すソーシャルグラフを解析する手法[7] を応用し，類似ユーザの推定が行われていると考えられる．このような手法は，オンライン上のつながりのみを判断基準としているため，実世界の行動範囲や興味を持つ場所が似ているユーザを検出することは不可能である．

一方，GPS ロガーなどを用いてユーザの移動軌跡を常に記録し，分析する手法も研究されている[8]．この手法では実世界の行動パターンを基準に類似ユーザを推定可能であるが，ユーザの意図と関係なく常に移動軌跡を記録してしまうため，プライバシが全く考慮されていないという問題がある．また，膨大な移動軌跡ログからどの場所が重要だったのかを解析することが大きな課題である．

また，ユーザが位置情報サービスを利用する際のプライバシ問題については，偽の位置情報を含む複数のクエリを投げてユーザの本当の位置情報を隠蔽する手法[9] など，数多くの手法が研究されている[10]．

3. ソーシャルメディアにおける位置情報を用いたグループ検出機構の提案

本節では，ソーシャルメディアに投稿された位置情報付きの発言を収集することによりユーザ間の類似度を推定する GLoBEs（Grouping Algorithm Based on Location of Micro-Blog Entries）を提案する．GLoBEs は，実世界の行動パターンや興味を持つ場所が似ているユーザを検出し，各ユーザの類似度を算出する．ユーザ間の類似度が閾値以上のユーザを，類似ユーザと定義する．また，あるユーザを基準にした際，基準ユーザから見た類似ユーザをグループとして取得し，グループ情報をアプリケーションから利用可能にする．

3.1 想 定 環 境

GLoBEs の想定する情報取得対象と，推定したグループ情報の利用者を以下に挙げる．GLoBEs はユーザが興味を持った場所の位置情報を入力し，類似したユーザをグループ情報として出力する．

(1) **ユーザの位置情報データ**
ユーザが興味を持った場所の位置情報を取得するため，ソーシャルメディアに投稿された位置情報付きの発言を対象とする．今回，発言内容の文字列は解析に使用しないが，Twitter などのサービスを利用することで，既存のソーシャルネットワーク基盤を活用することが可能になる．

(2) **グループ情報の利用者**
位置情報共有アプリケーションの開発者を対象とする．GLoBEs が判定したグループ情報は，直接エンドユーザに提供されるものではない．位置情報共有アプリケーションを拡張するために，GLoBEs が提供するグループ情報を利用することを想定する．

Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

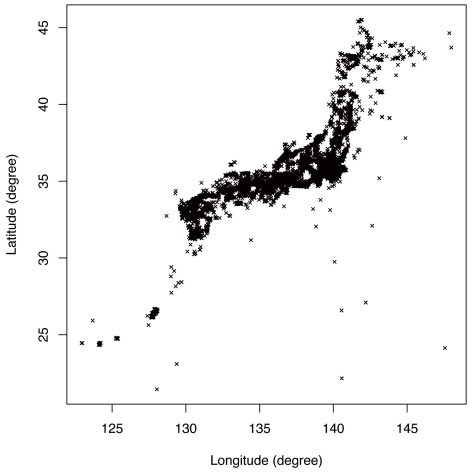

図 1 事前実験の位置情報をプロットした結果
Fig. 1 Result of pre experiment

3.2 事前実験

事前実験として，Twitter において位置情報付き発言がどの程度の頻度で発信されているか，それらが取得可能であるかを検証した．Twitter の StreamingAPI(statuses/filter) を利用し，日本列島周辺に位置情報が設定されたパブリックな発言を対象としてデータを取得した．実験期間は 2010 年 12 月 4 日から 2010 年 12 月 10 日の 7 日間で，110,419 件の位置情報付き発言を取得した．**図 1** に，取得した発言の位置情報をプロットした結果を示す．

事前実験の結果から，日本列島のほぼすべての場所において，位置情報付きの発言が行われていることが分かった．また，図の右下に，何も無い海上に位置情報が設定された発言が記録されている．これは，ランダムに位置情報を変えながら発言を繰り返す BOT や，海上を震源とする地震の位置情報を発言する BOT などによるものであった．このようなノイズとなる発言は，全体の 0.3%ほど観測された．さらに，この期間におけるユニークユーザ数は 16,256 人であった．それぞれのユーザあたりの発言数は（μ=6.8, σ=22.0）であり，ばらつきが大きいことが分かった．

表 1 ユーザ類似度判定に必要な項目
Table 1 Required information for calculating user similarity

項目	形式	必須/任意
発言時刻	年/月/日 時:分:秒	必須
発言ユーザ ID	数値	必須
発言位置情報	緯度, 経度	必須
発言内容	文字列	任意

これらの結果から，TwitterAPI を用いれば日本国内の多くの地域における位置情報付き発言を取得可能であることが検証された．

3.3 グループ判定アルゴリズム

ソーシャルメディアの発言を取得してからグループを判定するまでは，以下の手順で行う．

(1) 位置情報付き発言を取得する
(2) 位置情報をもとに発言のクラスタリングを行う
(3) ユーザ間の類似度を計算する
(4) グループ判定を行う

以降，それぞれの手順の詳細について述べる．

ソーシャルメディアにおける位置情報付き発言の取得

まず，ソーシャルメディアから位置情報付きの発言を取得する．取得において必要な項目を**表 1** にまとめた．これらの項目が取得可能なソーシャルメディアであれば，具体的なサービスを問わず適用可能である．

発言のクラスタリング

取得した位置情報付きの発言を，位置情報を基にクラスタリングを行う．クラスタリングを行う理由は，膨大な発言の中から，位置情報が近い組み合わせを効率的に取得するためである．クラスタリングを行わない場合，すべての発言の組み合わせに対して距離を計算する必要があり，膨大な計算量となるためアルゴリズムがスケールしなくなってしまうという問題がある．クラスタリングには，既存の複数のアルゴリズムが存在するが，今回は分類対象となる発言数が非常に多いため，計算量の少ない K-means 法を利用した．

図 2 では，事前実験のデータをクラスタ数 $K = 10$ でクラスタリングを行った結果を示す．色の境目がクラスタの境界であり，すべての発言は合計で 10 個のクラスタに分割されている．同じクラスタに属する発言は，行動範囲や興味を持つ場所が似ている発言であると解釈する．あるユーザを基準に考えたとき，まずはどのクラスタで何回発言しているかを計

Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

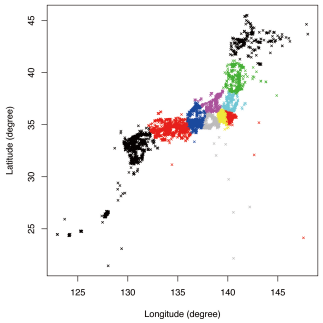

図 2 クラスタリング結果（$K = 10$ ）
Fig. 2 Clustering result ($K = 10$)

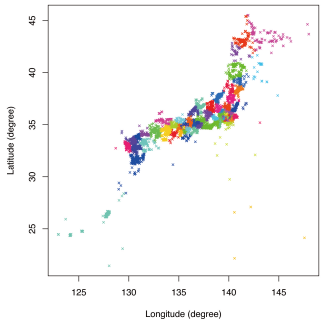

図 3 クラスタリング結果（$K = 100$）
Fig. 3 Clustering result ($K = 100$)

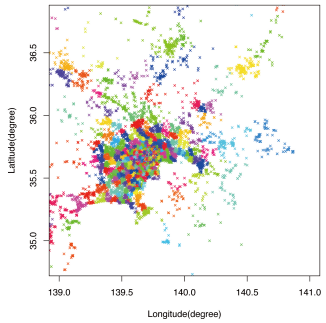

図 4 首都圏 （$K = 1000$）
Fig. 4 The Tokyo metropolitan area ($K = 1000$)

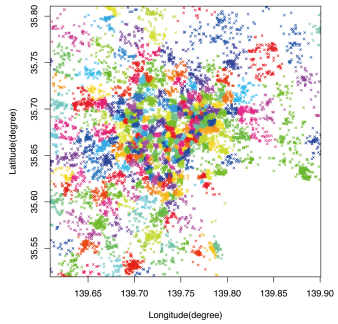

図 5 東京 23 区（$K = 1000$）
Fig. 5 The 23 wards of Tokyo ($K = 1000$)

算する．そして，他のユーザの中からクラスタを共有する発言が多いユーザほど，基準となるユーザからの類似度が高いと考えられる．

このようにして計算したクラスタは，クラスタの分割数 K が大きくなるほど，粒度が細かくなる．粒度が細かいクラスタを共有しているということは，普段使う駅，通っている学校や職場など，よりピンポイントに興味を持つ場所が似ていると考えられる．一方，粒度の粗いクラスタの共有を調べることで，東京から大阪によく出張している人など，移動する大きな地域が似ている人を判定することができる．そのため，GLoBEs では，クラスタの粒度を複数用意し，それぞれに重み付けを行うことで，より正確な類似度判定を可能にした．類似度計算における重みは，クラスタの粒度が細かくなるほど大きく設定する．クラスタ分割数は任意に設定可能であるが，今回は 10，100，1000 の三段階のクラスタを設定した．

$K = 10$ に設定すると，関東や東北，近畿などおおよそ日本の地方区分単位のクラスタリング結果となった．$K = 100$ では，**図 3** のように，おおよそ都道府県程度の大きさに加え，北海道や離島のような大きな面積を占める場所は，それぞれ複数に分割された結果となった．

図 4 では，$K = 1000$ でクラスタリングした結果を首都圏まで拡大した結果を示す．東京や神奈川の都心にかなりの発言が集中しておりクラスタが細かく分割され，千葉や群馬，栃木，茨城などの山間部においては発言がまばらで，クラスタの面積も大きくなっていることが分かる．これをさらに東京 23 区まで拡大したものが**図 5** である．これらの結果から，クラスタリングの分割数を大きくしていくと，都心など発言の密度が高い場所ほど，面積の小さいクラスタが多く発生すると考えられる．発言の密度が低い地域では，ある程度離れた場所で発言したユーザでも類似度が加算されやすくなるが，逆に発言の密度が高い地域においては，より近い場所で発言しないと類似度に影響しない．

類似度判定

計算したクラスタ情報を用いて，ユーザ間の類似度を計算する手法について述べる．ユーザ類似度は，すべてのユーザの組み合わせに対して設定される．例えば，ユーザ A とユーザ B が存在する場合，ユーザ A からユーザ B に対する類似度と，ユーザ B からユーザ A に対する類似度は異なるものとなる．また，ユーザ A からユーザ B に対する類似度の場合，類似度を計算する基準となるユーザ A を，基準ユーザと定義する．

クラスタは複数の粒度で計算されており，それぞれの粒度について計算した類似度に重み付けを行った加重平均を，全体の類似度とする．まずはひとつの粒度のクラスタにおける類似度計算の手順について，具体的な例を用いて説明する．

図 6 は，Taro と Hanako の 2 ユーザの位置情報付き発言にクラスタリング解析を行った例である．青丸が Taro の位置情報付き発言，赤丸が Hanako の位置情報付き発言で，近くの数字は発言 ID を表す．発言はそれぞれ K_1 から K_3 までの 3 クラスタに分類され，点線がクラスタの境界である．

Taro から Hanako に対する類似度を計算する手順について述べる．まず，ユーザがそれ

Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

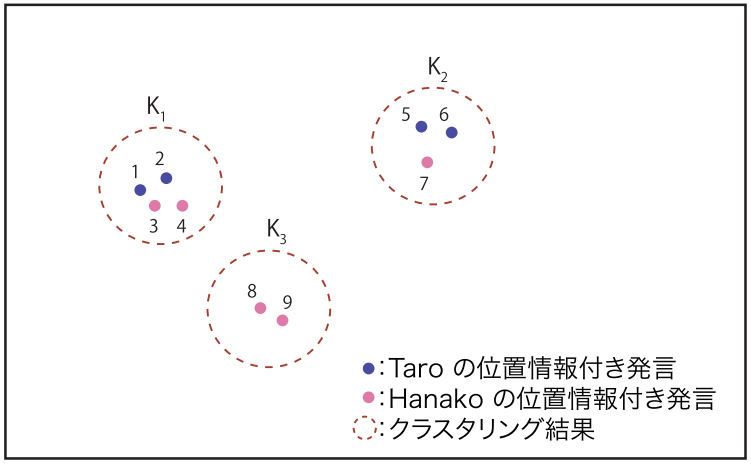

図 6 位置情報付き発言のクラスタリングを行った例
Fig. 6 Example of clustering analyzed data

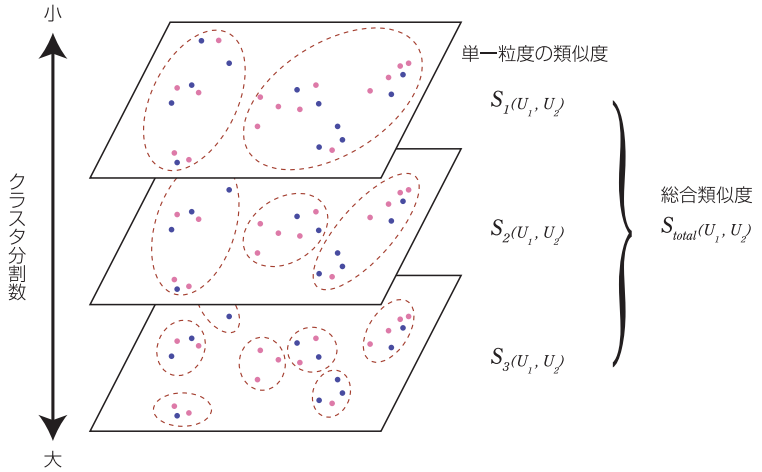

図 7 複数粒度のクラスタリングと総合類似度
Fig. 7 Layerd similarity calculation and total similarity

ぞれのクラスタで何回発言しているかをカウントする．

次に，それぞれの発言数について定義する．K_1 における Taro の発言数を $t_{(K_1,\ Taro)}$ とする．同様に，K_1 における Hanako の発言数を $t_{(K_1,\ Hanako)}$ とする．そして，K_1 において Taro と Hanako の発言数のうち少ない方を，発言共有数 $C_{K_1(Taro,\ Hanako)}$ とする．ここでは，$C_{K_1(Taro,\ Hanako)}$ は 2 となる．式 1 は，$C_{K_1(Taro,\ Hanako)}$ の定義を示す．

$$C_{K_1(Taro,\ Hanako)} = min(t_{(K_1,\ Taro)},\ t_{(K_1,\ Hanako)}) \tag{1}$$

また，すべてのクラスタにおける Taro と Hanako の発言共有数の合計を，合計発言共有数 $C_{total(Taro,\ Hanako)}$ とする．ここでは，$C_{total(Taro,\ Hanako)}$ は 3 となる．式 2 は，$C_{total(Taro,\ Hanako)}$ の定義であり，n はクラスタ分割数を示す．

$$C_{total(Taro,\ Hanako)} = \sum_{j=1}^{n} C_{K_j(Taro,\ Hanako)} \tag{2}$$

そして，Taro の全発言数を $T_{total(Taro)}$ とする．ここでは，$T_{total(Taro)}$ は 4 となる．$T_{total(Taro)}$ の定義を式 3 に示す．

$$T_{total(Taro)} = \sum_{j=1}^{n} t_{(K_j,\ Taro)} \tag{3}$$

最後に，合計発言共有数を基準ユーザである Taro の全発言数で除算した数を，Taro から Hanako に対する類似度 $S_{(Taro,\ Hanako)}$ とする．式 4 に，$S_{(Taro,\ Hanako)}$ の定義を示す．ここでは合計発言共有数は 3，Taro の全発言数は 4 のため，Taro から Hanako に対する類似度は 0.75 となる．

$$S_{(Taro,\ Hanako)} = \frac{C_{total(Taro,\ Hanako)}}{T_{total(Taro)}} \tag{4}$$

以上のアルゴリズムを一般化し，定義した式を述べる．基準ユーザ U_1 から対象となるユーザ U_2 に対する類似度を $S_{(U_1,\ U_2)}$ とする．式 5 は，$S_{(U_1,\ U_2)}$ を示す．

$$S_{(U_1,\ U_2)} = \frac{C_{total(U_1,\ U_2)}}{T_{total(U_1)}} \tag{5}$$

以上の手順で，単一粒度のクラスタにおける類似度 $S_{(U_1,\ U_2)}$ が計算される．クラスタリングは複数の分割数で行い，それぞれに重み付けをした結果を，総合類似度として計算する．図 7 に，複数の分割数によるクラスタリングと総合類似度の概要を示す．クラスタ分割

Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

数の小さい方から S_1，S_2 と割り当て，最大 S_m の粒度をもつ．m は最大粒度数を示す．

次に，すべての粒度の類似度に重み付けをした総合類似度 $S_{total(U_1,\ U_2)}$ を求める手順について述べる．それぞれの粒度で類似度を計算した結果を，$S_{1(U_1,\ U_2)}$，$S_{2(U_1,\ U_2)}$，… ，$S_{m(U_1,\ U_2)}$ とする．また，それぞれのクラスタ粒度に対する重みを w_1，w_2，… ，w_m とする．粒度の細かいクラスタを共有しているユーザほど類似度が高いと考えるため，分割数の多いクラスタほど重み w の値も大きく設定する．これらを加重平均した結果を式 6 に示す．

$$S_{total(U_1,\ U_2)} = \frac{w_1 \cdot S_{1(U_1,\ U_2)} + w_2 \cdot S_{2(U_1,\ U_2)} + \ldots + w_m \cdot S_{m(U_1,\ U_2)}}{w_1 + w_2 + \ldots + w_m} \qquad (6)$$

グループ判定

類似ユーザを基にしたグループ判定について述べる．U_1 を基準ユーザとして考えた場合，U_1 から見たすべての類似ユーザを，U_1 のグループであると定義する．

これを一般化すると以下のように定義される．全ユーザが n 人いる場合，U_{from} を基準ユーザとして考えると，$S_{total(U_{from},\ U_1)}$，$S_{total(U_{from},\ U_2)}$，… ，$S_{total(U_{from},\ U_n)}$ の類似度が存在し，それぞれ S_{th} を上回ったユーザを，U_{from} から見たグループであると判定する．

3.4 モジュール構成

GLoBEs は，発言取得モジュール，発言解析モジュール，クラスタリングモジュール，類似度計算モジュール，グループ情報データベース，グループ情報取得モジュールによって構成される．**図 8** にモジュール構成図を示す．

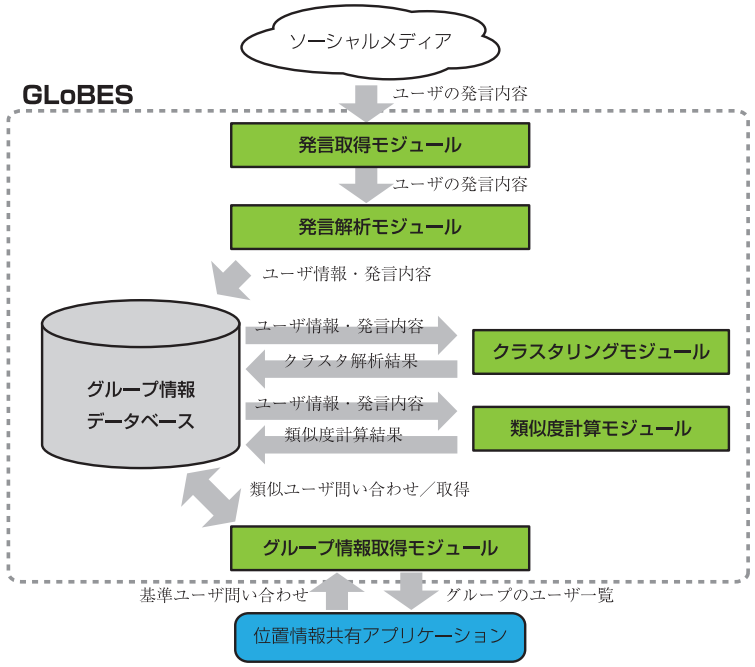

図 8 モジュール構成図
Fig. 8 Module configuration

4. GLoBEs の実装

本節では，GLoBEs が対象とするソーシャルメディアの選定と，実装環境について述べる．実装において使用したソフトウェアの構成と，各モジュールの実装言語をまとめる．

4.1 概　　要

まず，GLoBEs が対象とするソーシャルメディアの要件として，以下の二点を挙げる．

(1)　発言内容が API により取得可能なこと
(2)　発言に位置情報が付加されうること

1 番目は，ソーシャルメディアの発言を取得する必要があるため，API が公開されているソーシャルメディアを対象とする．2 番目の要件は，発言の位置情報を基にクラスタリング解析を行うため，それぞれの発言に緯度経度の位置情報が付加されていることを必要とする．

以上の要件を考慮し，今回対象としたソーシャルメディアについて述べる．発言を取得する API が公開されており，発言に位置情報を付加できるソーシャルメディアとしては，具体的には Twitter や GoogleBuzz が候補となった．今回は発言取得 API の充実や，サービス利用者の多さから Twitter を対象として実装を行った．

4.2 実 装 環 境

実装に使用したソフトウェア構成を**表 2** に示す．グループ情報取得モジュールでは Web ベースの API を提供するため，Web サーバとして実績のある Apache HTTP Server 2.2.9 を選定した．また，グループ情報データベースには RDBMS である MySQL 5.0.51a を選定した．各モジュールの実装には，スクリプト言語である PHP 5.2.6 を選定し，クラスタリングモジュールにおいては R 言語 ver. 2.7.1 を使用した．GLoBEs は常に発言を取得するため，長期間の安定稼働が必要とされる．そのため，これらのソフトウェアを実行させる OS として Debian GNU/Linux 5.0.7 を使用した．

Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

表 2 ソフトウェア構成
Table 2 Software configuration

要素	詳細
OS	Debian GNU/Linux 5.0.7(lenny)
Web サーバ	Apache HTTP Server 2.2.9
解析スクリプト	PHP ver. 5.2.6 with Suhosin-Patch 0.9.6.2
データベース	MySQL ver. 5.0.51a
クラスタリング解析	R ver. 2.7.1

表 3 ハードウェア構成
Table 3 Hardware configuration

要素	詳細
モデル	HP Pavilion Desktop PC v7780jp/CT
CPU	Intel Core2 Quad Q9550 @ 2.83GHz
メモリ	4GB PC2-6400（800MHz）
HDD	1000GB

5. 評　　価

GLoBEs の評価として，発言数とクラスタリング解析処理に必要な時間の関係と，GLoBEs の推定したグループ情報と Twitter におけるフォロー関係の 2 点を検証した．評価環境として用いたハードウェアの構成を**表 3** に示す．

5.1 発言数とクラスタリング解析処理に必要な時間

発言数とクラスタリング解析処理に必要な時間を比較することで，クラスタリングモジュールの評価を行う．評価方針としては，クラスタリング対象となる発言数とクラスタリング解析処理に必要な時間との関係を用い，評価結果を述べる．

評 価 方 法

クラスタリングモジュールは，クラスタリング対象となる発言数と，クラスタリング解析処理に必要な時間との関係を評価する．事前実験の結果，1 日の位置情報付き発言は日本周辺に限定するとおよそ 20000 件取得されることが分かった．クラスタリングモジュールは数日〜数週間に一度実行され，取得した一定期間の発言に対して解析を行う．そのため，解析対象となる発言数が増加しても計算可能であることが求められる．

評 価 結 果

2000 件，20000 件，200000 件の発言数に対し，それぞれ 10 回クラスタリング計算を試行し，実行にかかった時間と実行時間の標準偏差を**表 4** にまとめた．クラスタリング計算では，すべての発言を読み込み，$K = 10$，$K = 100$，$K = 1000$ の 3 つの粒度のクラスタリング結果を出力している．

表 4 発言数とクラスタリング解析時間における評価
Table 4 Evaluation result of clustering module

発言数	平均解析時間（秒）	標準偏差（秒）
2000	0.77	0.02
20000	5.05	0.29
200000	50.51	3.36

表 4 の結果から，約 1 日分の発言数である 20000 件のデータが，平均 5.05 秒で解析可能であることが分かる．また，計算量は $O(n)$ であり，問題無く定期的に解析を行うことができる．

5.2 グループ情報と Twitter におけるフォロー関係

GLoBEs の推定したグループ情報と Twitter におけるフォロー関係を比較することで，類似度計算モジュールの評価を行う．評価方針として，GLoBEs がグループとして判定したユーザを，Twitter 上でフォローしている割合を調べ，GLoBEs の計算した類似ユーザと Twitter におけるフォロー状況との関係性を評価する．

評 価 方 法

まず 2011 年 1 月 1 日 0:00:00〜2011 年 1 月 1 日 23:59:00 の 1 日間における 32,538 件の発言を取得し，クラスタリング解析を行った．次にグループ判定のパラメータとして，クラスタ分割数 10，100，1000 に対して $w_1 = 1$，$w_2 = 2$，$w_3 = 3$ の重み付けを行い，閾値 $S_{th} = 0.6$ に設定して類似度計算を行った．

さらに，期間内に発言した 5,484 人のユーザそれぞれについて，GLoBEs がグループであると判定したユーザが，実際にフォローしているユーザとどの程度一致しているかの割合を計算した．これを Following 一致率（$R_{following}$）とし，式 7 に 1 ユーザあたりの $R_{following}$ の計算式を示す．G はグループと判定したユーザ数であり，C はグループのうち実際にフォローしていた人数を示す．

$$R_{following} = \frac{C}{G} \tag{7}$$

すべてのユーザにおいて $R_{following}$ を計算し，平均と標準偏差を求める．

Vol.2011-MBL-57 No.40
Vol.2011-UBI-29 No.40
2011/3/8

評価結果

すべてのユーザにおける $R_{following}$ の結果は（μ=0.012，σ=0.065）であり，グループに含まれるユーザをフォローしている割合は1%程度であることが分かった．従って，GLoBEsの判定したグループ情報は，従来のソーシャルグラフをもとにした類似ユーザ推定とは大きく異なる結果を示している．定性的な評価として，グループして判定されたユーザは，プロフィールで公開している位置情報が比較的近かったり，FourSquare などのサービスを使用してチェックインする場所が類似しているなどの傾向が見られた．グループとして判定されたユーザの興味がどの程度似ているかを定量的に評価することは現段階では達成できておらず，今後の課題である．

6. ま と め

本稿では，ソーシャルメディアにおける位置情報付きの発言を利用した類似ユーザの推定手法 GLoBEs (Grouping Algorithm Based on Location of Micro-Blog Entries) を提案した．実世界の行動を基準とした類似ユーザ推定により，ソーシャルグラフのみを用いた従来の手法では発見できなかった新たなユーザ間の関係を明らかにした．今後の課題として，GLoBEs によるグループ情報と，実世界の行動パターンの類似性を定量的に評価する手法を検討し，より正確な類似ユーザ推定を行うことを目指す．また，ユーザのプライバシを適切に保護する機構を導入し，実際の位置情報共有アプリケーションに適用した際の有用性を実証していきたい．

参 考 文 献

1) Apple Inc.: iPhone, Website (2010). `http://www.apple.com/jp/iphone/`.
2) Open Handset Aliance: Android, Website (2010). `http://www.openhandsetalliance.com/`.
3) Facebook, Inc.: Facebook, Website (2010). `http://www.facebook.com/`.
4) Twitter, Inc.: Twitter, Website (2010). `http://twitter.com/`.
5) Cranshaw, J., Toch, E., Hong, J., Kittur, A. and Sadeh, N.: Bridging the gap between physical location and online social networks, *Proceedings of the 12th ACM international conference on Ubiquitous computing* (2010).
6) Tang, K.P., Lin, J., Hong, J.I., Siewiorek, D.P. and Sadeh, N.: Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing, *Proceedings of the 12th ACM international conference on Ubiquitous computing* (2010).
7) Getoor, L. and Diehl, C.P.: Link mining: a survey, *SIGKDD Explor. Newsl.*, Vol.7, pp.3–12 (2005).
8) Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W. and Ma, W.-Y.: Mining user similarity based on location history, *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems* (2008).
9) Shankar, P., Ganapathy, V. and Iftode, L.: Privately querying location-based services with SybilQuery, *Proceedings of the 11th international conference on Ubiquitous computing*, Ubicomp '09, New York, NY, USA, ACM, pp.31–40 (2009).
10) Brush, A.B., Krumm, J. and Scott, J.: Exploring end user preferences for location obfuscation, location-based services, and the value of location, *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, New York, NY, USA, ACM, pp.95–104 (2010).