2H5-4

ライフログ利用履歴とユーザ群クラスタリングを用いた 非 Twitter ユーザのプロファイリング

Using a life log usage history and the user group clustering non-Twitter user profiling

望月佑樹　延原肇[*1]
Yuki Mochizuki Hajime Nobuhara

横石圭介[*2]
Keisuke Yokoishi

[*1] 筑波大学
University of Tsukuba

[*2] 株式会社サイバード
CYBIRD Co.Ltd

There are two types of users in the recommendation services, and one is the type of users who already belong to other SNS services (SNS users), and the other is the type of uses who did not use any SNS services (non-SNS users). In the case of SNS uses, we can extract the rich information from the SNS log for the profiling users. On the other hand, it is difficult for non-SNS users to get such the information from SNS log. The proposed method is to support the profiling non-SNS users based on the SNS users cluster information and the correspondence of log between the non-SNS users and SNS users.

1. はじめに

近年、インターネットをはじめとするインフラの整備、および高性能なスマートフォンの普及(普及率は約 8 割[1])により、人々は手軽に情報を獲得することが可能になっている。また、現行のスマートフォンには GPS 機能付きのものが多く存在しており、GoogleMap[2]や Swarm[3]などの GPS 機能を積極的に活用したアプリケーションも数多く登場している。例えば、Twitter[4]や Instagram[5]などに代表されるソーシャルメディアなどが挙げられ、様々なリソースから即時性の高い情報を得られるようになっている。同時に、これらのサービスから、利用ユーザに関する様々な情報を獲得、またユーザプロファイルへ応用することが可能になっている。Gunosy[6]では、Twitter や Facebook のアカウントでログインすることで、それらの SNS からユーザの情報を解析し、ユーザの興味や関心のある情報を自動で配信している。

あるコンテンツを推薦するサービスにおいて、ユーザの Twitter や Facebook といった SNS のアカウントを用いてログインさせれば、それぞれの SNS におけるプロファイルを利用し嗜好に合わせた推薦ができるようになっている。しかし、Twitter や Facebook などを利用していないユーザの場合、これらのアカウント情報が利用できないため、提供するサービス内で獲得した情報のみしか利用できない。

本研究では、あるコンテンツを推薦するサービスにおいて、他の SNS を利用していないユーザに対し、当該ユーザの過去の履歴を利用することで、SNS を利用しているユーザと同様のプロファイル情報を与えることを目的とする。今回は我々の研究室が企業と共同で開発を行っているコンテンツ推薦サービスである FourDiary をモデルケースとして利用し、SNS として Twitter を対象に考える。提案手法(図 1)ではまず、Twitter を利用しているユーザをフォロー数、フォロワー数などの関連情報によって構成されたベクトルでクラスタリングする。つぎに、FourDiary のユーザの非 Twitter ユーザを選択すれば、そのユーザの FourDiary の履歴と、クラスタリングされた Twitter ユーザの FourDiary の履歴を比較し、最も類似しているクラスタのセントロイドをユーザの特性を表すプロファイルとして付与する手法を提案する。

本論文では、第 2 章で Twitter を用いたプロファイリング手法についての関連研究について述べる。第 3 章で提案手法の詳細について説明する。第 4 章では評価実験と結果を示す。第 5 章ではまとめと今後の展望について論じる。

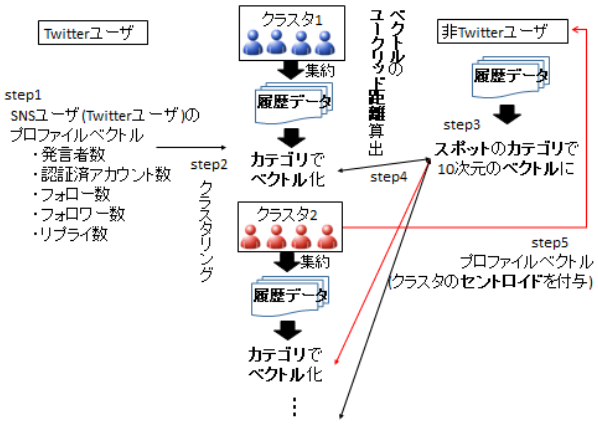

図 1　提案システムの概要

2. 関連研究

2.1 Twitter でのプロファイリングに関する研究

榊ら[7]は、Twitter のリスト機能に着目して、リスト機能をユーザに対するソーシャルブックマークと考え、ソーシャルブックマークを対象に用いられる手法により、ユーザの属性判別・特徴語の抽出を行い分析している。しかし、リスト機能はユーザ自身が自由にユーザの分類する機能であるため、分類の仕方によっては嗜好情報を表していない場合もあると考えられる。

向井ら[8]は、リツイート機能に着目している。リツイートを利用してユーザのプロファイリングを行った後、それをクラスタリングすることで協調フィルタリングと同様の効果の獲得とセレンディピティのある推薦を可能にしている。しかし、リツイートの投稿数は

望月佑樹
mail:mochizuki@cmu.iit.tsukuba.ac.jp
住所:茨城県つくば市天久保 3-16-4 リオ・グランデ 204

通常のつぶやきと比べると数がかなり少なく、ユーザ 1 人に対してプロファイリングに十分な数が得られているとは言えない。

2.2 本研究で取り扱う推薦サービス

著者らの所属する研究室では、すでに FourDiary[9]というレコメンデーション機能を有した推薦サービスを開発している。FourDiary は持ち歩くだけでその位置情報を取得し、自動で日々の行動を記録するライフログアプリケーションである。また、地域ごと・季節ごとに記録を振り返ることができ、日記のように扱うこともできる。また、FourDiary では記録された場所ごとに News、Spot、Twitter を図 2 のようにそれぞれ推薦する機能を有している。図 3 に推薦の全体の流れを示す。我々の研究室側では、レコメンデーションサーバー側の開発を行っている。アプリケーション側から送られてくるユーザの位置や時間などの情報と、ユーザの利用している Twitter・Facebook の情報をもとに、それぞれのモジュールでコンテンツを決定し推薦をする。

図 2 レコメンデーションの様子

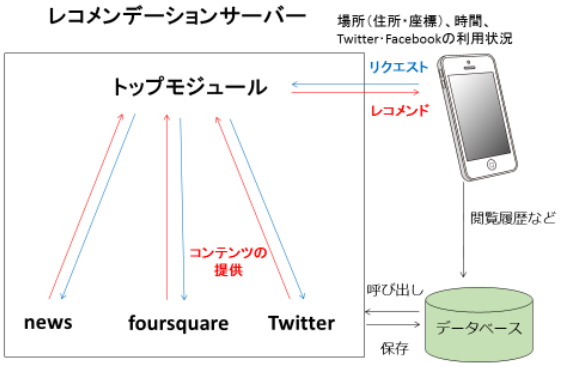

図 3 FourDiary の概要

3. 提案手法

本研究では、Twitter を使っていないユーザに対して、FourDiary 内の Twitter ユーザのプロファイル情報とアプリケーションから得られた履歴情報を利用することで、プロファイル情報を推定し付与することを目的とする。

提案手法のフローチャートを図 4 に示す。まず、Twitter を利用しているユーザにプロファイル情報をベクトルとして与える。このベクトルに対して k-means クラスタリングを適用することで、ユーザをプロファイル情報でクラスタに分割する。この際、クラスタの中心(セントロイド)を求めておく。次に、クラスタ毎に、クラスタに属するユーザの FourDiary における訪れた場所の履歴情報をまとめる。このクラスタ毎の履歴情報と非 Twitter ユーザの履歴情報を比較し、最も履歴が類似しているクラスタのセントロイドをユーザのプロファイル情報としてユーザに与える。

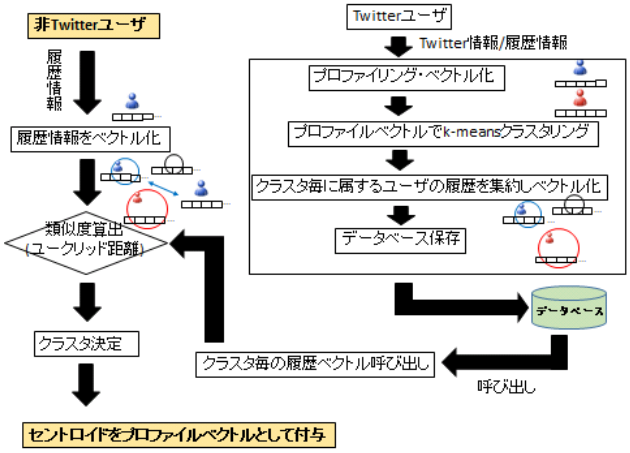

図 4 提案システムのフローチャート

3.1 k-means クラスタリング

Twitter ユーザのクラスタリングでは、TwitterAPI によって取得した各種タイムラインから、「発言者数」「認証済アカウント数」「フォロー数」「フォロワー数」「リプライ数」を要素とした 5 次元のプロファイルベクトルを作成する。クラスタリングには k-means 法を利用し、Twitter ユーザのデータ数を n とすると、その集合は

$$X = \{x_1, \ldots, x_n\} \tag{1}$$

で表され、ここで $x_i \in Z^{+5}$ である。これらを k 個のクラスタに分ける。このとき、データ集合はクラスタ $X_1, \ldots, X_k$ からなる集合

$$X = \{X_1, \ldots, X_k\} \tag{2}$$

と表せる。それぞれのクラスタの中心点(セントロイド)は

$$C = \{c_1, \ldots, c_k\} \tag{3}$$

であり、ここで $c_i \in R^5$ である。k-means 法では、データ集合 X を、

$$F(X, C) = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - c_j\|^2 \tag{4}$$

で定義される評価関数を最小化する。この最小化の手順は以下の流れによって行われる。

1. 各データ点 x_i に対してランダムにクラスタを割り振る。

2. 割り振ったデータをもとに各クラスタの中心点(セントロイド) c_i を以下の式によって計算する。

$$c_j = \frac{1}{|X_j|}\sum_{x \in X_j} x \tag{5}$$

X_j は j 番目のクラスタに属するデータ集合を表す。

3. 各データと各クラスタのセントロイド間の距離 $d(x,c)$ を計算する。

$$d(x,c) = \|x_i - c_j\|^2 \tag{6}$$

4. 各データ点を、セントロイドとの距離が最小となるクラスタに割り振りなおす。

2~4 を繰り返し、データの属するクラスタに変化がなかったとき、それをデータが属するクラスタとして決定する。

求めた各クラスタのセントロイドも同時に保存しておく。また、クラスタに属するユーザ数が 4 以下の場合は、そのクラスタを除外する。

3.2 非 Twitter ユーザとのマッチング

クラスタリング後に、そのクラスタに属するユーザの FourDiary 上での履歴をクラスタ毎に保存する。履歴には、ユーザが訪れた Foursquare のスポットが記録されている。履歴には Foursquare での n 個のカテゴリを利用する。クラスタに属するユーザの全履歴数を p、各カテゴリに属する履歴数をq_i とすると、履歴は

$$H_{\{j\},j=1,2,\ldots,k} = [\frac{q_1}{p}, \cdots, \frac{q_n}{p}] \tag{7}$$

という n 次元のベクトルで表せる。

非 Twitter ユーザについて、同様に FourDiary 上での履歴を 10 次元のベクトルにする。非 Twitter ユーザのベクトルの要素を $A = [a_1, \ldots, a_n]$、各クラスタのベクトルの要素を$H = [h_1, \ldots, h_n]$ とすると、ユークリッド距離は

$$D(A,H) = \sum_{i=1}^{n} \|a_i - h_i\|^2 \tag{8}$$

で表され、最もユークリッド距離の小さいクラスタを 1 つ決定する。そのクラスタのセントロイド(プロファイルベクトルの中央値)をユーザのプロファイルベクトルとして与える。

4. 評価実験

システムの有効性を検証するために、分割交差検証を行った。評価実験の手順を図 5 に示す。まず、事前に全 Twitter ユーザ 274 に対してクラスタリングを行い、クラスタに属するユーザとクラスタのセントロイドを求める。このとき、5 人未満のクラスタとそのクラスタに属するユーザを除いておく。次に、残った Twitter ユーザを 10 個に分割し、そのうちの 1 つをテストデータ集合、残りの 9 個をトレーニングデータ集合とする。トレーニングデータ集合を Twitter ユーザ、テストデータ集合を非 Twitter ユーザとして提案システムを適用し、テストデータに対しプロファイルベクトルを付与する。このプロファイルベクトルと事前に求めたクラスタのセントロイドのユークリッド距離を計算し、最も小さいクラスタに自身が属しているかを確認する。これを 10 回行い、正解率を求める。

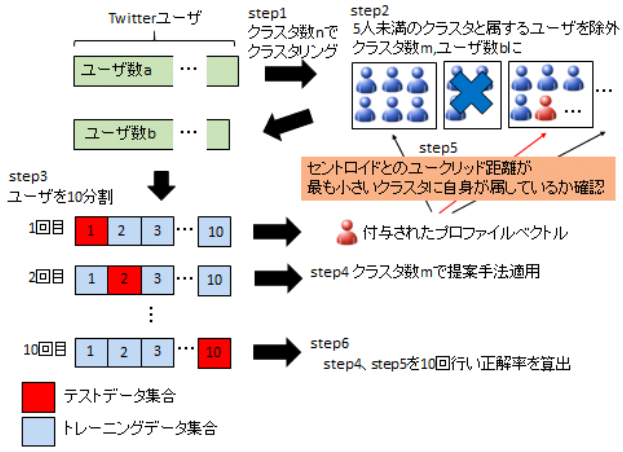

図 5 評価実験手順

表 1 クラスタ数と正解率

クラスタ数 n(m)	1 回目	2 回目	3 回目	4 回目	5 回目	6 回目	7 回目	平均	標準偏差	1/m
15(8)	0.118	0.092	0.072	0.096	0.101	0.140	0.113	0.104	0.020	0.125
20(11)	0.056	0.064	0.106	0.052	0.090	0.091	0.130	0.084	0.027	0.091
25(12)	0.091	0.060	0.064	0.078	0.047	0.095	0.052	0.070	0.017	0.083
30(16)	0.068	0.070	0.079	0.078	0.101	0.091	0.082	0.081	0.011	0.063

n:最初に指定するクラスタ数　m:5 人以上のクラスタ数

表 2 クラスタ数 30 のときのクラスタの履歴ベクトルとセントロイド

	履歴のベクトル										プロファイルベクトル(セントロイド)				
人数	夜の娯楽スポット	店舗&サービス	飲食店	カレッジ&大学	アウトドア&レクリエーション	旅行&交通	イベント	住宅	芸術&娯楽	専門&その他	発言者数	認証済ユーザ数	フォロー数	フォロワー数	リプライ数
51	0.054	0.272	0.512	0.003	0.043	0.027	0.000	0.004	0.069	0.016	37.216	8.392	115.392	97.000	27.647
12	0.024	**0.695**	0.227	0.000	0.018	**0.010**	0.000	0.000	**0.010**	0.016	67.833	8.500	455.917	591.250	15.500
8	0.000	0.588	0.214	0.000	0.046	0.042	0.000	0.000	0.084	0.025	88.375	14.875	1036.625	504.250	15.625
30	0.020	0.271	0.300	0.000	0.063	0.027	0.000	0.000	0.039	0.279	60.533	17.667	293.300	132.233	12.367
31	0.059	0.354	0.398	0.006	0.062	0.012	0.000	0.003	0.032	0.074	59.097	10.000	323.516	299.290	20.161
133	0.018	0.319	0.286	0.059	0.058	0.096	0.000	0.004	0.063	0.096	**8.955**	**3.158**	**14.150**	**6.534**	**4.632**
14	0.051	0.230	**0.611**	0.000	0.078	**0.012**	0.000	0.000	**0.016**	0.004	81.357	10.571	671.786	652.429	23.500
5	0.006	0.271	0.142	0.000	0.017	0.020	0.000	0.000	0.006	0.539	45.200	4.000	392.200	962.000	3.800
31	0.014	0.372	0.398	0.006	0.078	0.026	0.003	0.003	0.050	0.050	46.871	8.452	178.516	191.581	41.774
7	0.043	0.261	0.565	0.000	0.014	0.007	0.000	0.000	0.094	0.014	84.000	6.143	840.571	806.714	14.429
15	0.053	0.162	0.431	0.164	0.057	0.027	0.000	0.000	0.088	0.019	72.333	9.400	506.667	448.733	31.200
15	0.034	0.290	0.386	0.002	0.029	0.023	0.000	0.057	0.165	0.015	71.133	14.733	455.133	214.467	14.667
5	0.007	0.284	0.493	0.000	0.067	0.022	0.000	0.000	0.104	0.022	96.400	27.000	891.400	276.200	1.200
92	0.031	0.321	0.388	0.005	0.050	0.058	0.001	0.000	0.110	0.035	26.935	6.207	62.728	40.663	14.815
32	0.020	**0.235**	0.328	0.000	0.037	**0.173**	0.000	0.000	0.054	0.154	51.750	15.000	193.719	64.719	5.750
11	0.020	0.331	**0.267**	0.000	0.022	**0.213**	0.000	0.000	0.020	0.126	58.727	6.273	301.182	428.727	39.545

結果を表 1 に示す。正解率はあまり高くならず、1/m を超えているものは n=30 のときのみとなった。クラスタ数が大きくなるほど標準偏差と、正解率と 1/m の差が小さくなっていくことがわかる。正解したユーザとそのユーザの持つ履歴数に相関関係は見られなかった。

今回は SNS として性格的な特徴を取りやすい Twitter を利用したが、Facebook を利用することで具体的なユーザ情報(年齢・性別・住所・職業など)がわかるため、生活している地域やユーザの個人情報による分類が可能になると考えられる。また、FourDiary では訪れた場所のスポットの確定を自身でやる必要があるため、それを行わないユーザはカテゴリの数が少なくなってしまい、使い続けていくことでのカテゴリ数の向上も見込めない。

実験条件として、Twitter ユーザのプロファイルベクトルの取り方や、交差検証の分割数が適切でなかったことなどが問題点として考えられる。Twitter から取れるユーザ情報は今回利用したもの以外にもあるため、今回利用したものを含めて精査し、ベクトルの要素として相応しいか考えていく必要がある。また、今回はサンプル数の確保のために長く使っているユーザとあまり使っていないユーザをわけずに考えたが、履歴数の多いユーザのみで行うことで正解率は上昇すると考えられる。

また、表 2 に n=30 のときの各クラスタの詳細を示す。旅行&交通や芸術&娯楽の割合と店舗&サービスや飲食店の割合に多少の相関関係が見られ、行動範囲が狭い人と広い人である程度のグループわけができていることがわかる。また、人数の最も多いクラスタのプロファイルベクトルの要素の値がすべて小さいことから、アカウントはあるがあまり活用していないユーザが多いことがわかる。

5. おわりに

本論文では、Twitter ユーザから得られたプロファイル情報とサービス内で得られる履歴情報を用いて非 Twitter ユーザに対してプロファイル情報を与えるシステムを提案した。提案手法では、Twitter ユーザをプロファイルベクトルでクラスタリングし、クラスタ毎の履歴情報と非 Twitter ユーザの履歴情報を比較し、最も近いクラスタのセントロイドであるベクトルを非 Twitter ユーザに与える。評価実験を行った結果、ユーザに対して適切なプロファイル情報を与えることができたとはいえないことがわかった。また、クラスタリング後の各クラスタの履歴にはそれぞれ特徴が見られた。今後は、Twitter ユーザのプロファイルベクトルの取り方や履歴のベクトル化の方法、類似度の算出方法などを再考していくことを考えている。

参考文献

［1］ http://www.telecompaper.com/news/japan-smartphone-penetration-grows-to-77--1098138

［2］ https://itunes.apple.com/jp/app/google-maps/id585027354?mt=8

［3］ https://itunes.apple.com/jp/app/swarm-by-foursquare/id870161082?mt=8

［4］ https://twitter.com/

［5］ https://www.instagram.com/

［6］ https://gunosy.co.jp/

［7］ 榊剛史,松尾豊:ソーシャルブックマークとしての Twitter リスト機能の応用,The 24th Annual Conference of the Japanese Society for Artificial Intelligence(2010)

［8］ 向井友宏,黒澤義明,目良和也,竹澤寿幸:マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案,言語処理学会第 17 回年次大会発表論文集(2011)

［9］ https://www.fourdiary.com/