3M4-4

ドメイン別ユーザプロファイルの構築と情報推薦への応用

Building Domain Specific User Profiles from Twitter for Information Recommendation

鈴木 陽介　　尾崎 知伸
Yosuke Suzuki　Tomonobu Ozaki

日本大学文理学部
College of Humanities and Sciences, Nihon University

In recent years, information overload is recognized as one of major problems in information retrieval. To alleviate this problem, recommendation systems attract much attention. In this paper, we propose a framework for building domain-specific user profiles, which play a central role in recommendation systems. As a case study to confirm reliability of the proposal, we build user profiles in domain of cooking by using real world twitter datasets.

1. はじめに

近年，情報過多を背景に，ユーザの嗜好に合わせて有益な情報を提示する推薦システムの研究が盛んに行われている．推薦システムの代表的な手法として協調フィルタリング [Su 09] や内容ベースフィルタリング [神嶌 08] があげられる．これらの手法では，ユーザの嗜好プロファイルやアイテムの特徴プロファイルを基に推薦を実現しており，プロファイル構築の精度が推薦システムの能力に大きな影響を与える．

具体的な推薦では，推薦対象に対して何らかの視点が重要となる．例えば観光地を推薦する場合，観光地のジャンルやどのような景色を好むかに加え，食事や名産品など，観光に関する多様な視点を考慮して推薦する必要があるだろう．この様に目的や状況によってプロファイルを変更する，もしくは複合的に利用することで高精度な推薦システムの構築が可能であると考えられる．

一方，無数に考えられるドメインに対して，プロファイルを一つ一つ手作業で作成することは困難を極める．本研究ではこの問題を軽減するため，様々な情報が投稿されている Twitter に着目し，ドメインに特化したプロファイルの自動作成を試みる．Twitter とは日々の体験や感じたことなどを自由に投稿 (ツイート) できる SNS である．SNS 特有のリアルタイム性や情報伝搬能力の高さから，情報発信に利用しているユーザが多数存在する中で，特に企業や芸能人といったユーザは，自身の特徴的な分野に関する話題を多く投稿する傾向がある．

本研究では，あるドメインに対して特化した特徴的ユーザのツイートを利用したドメイン別ユーザプロファイル構築の自動化，およびその利用例として Twitter のユーザ推薦手法を提案する．

2. プロファイルの構築

本研究では，あるドメインに関するツイートを頻繁に投稿しているユーザを利用することで，多様なドメインにおけるユーザプロファイル構築の自動化を目的とする．提案手法の全体像を図 1 に示す．

連絡先: 尾崎知伸，日本大学 文理学部 情報科学科，〒 156-8550
東京都世田谷区桜上水 3-25-40，tozaki@chs.nihon-u.ac.jp

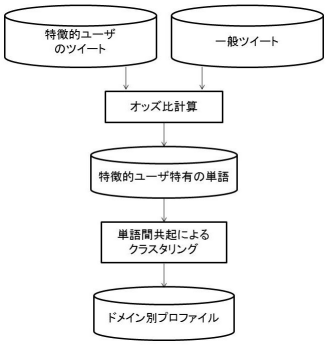

図 1: ドメイン別ユーザプロファイル構築手順

2.1 ドメイン別辞書の構築

ドメイン別ユーザプロファイル作成のために，ドメイン別辞書の構築を行う．一般的なツイートと，特徴的ユーザのツイートにおける単語の出現頻度を利用することでドメインに特化した単語を選定し，辞書を構築する．以下に具体的な手順を示す．

まず，TwitterAPI[*1] を利用してツイートを獲得する．その際，あるドメイン d に関するツイートを頻繁に投稿しているユーザらのツイート群を T^d，その他の一般的なユーザによる一般ツイート群を T^s とし，それぞれを分けて獲得する．次いで，一般ツイート群 T^s と特徴的ユーザらのツイート群 T^d にそれぞれ含まれる名詞，形容詞，副詞の集合を W^s, W^d とし，それら二つの単語群のオッズ比を用い，特徴的な語を抽出する．オッズ比とは，二つの集合において，ある事象がどちらの集合で起こりやすいかを表したものである．単語 $w \in W^d$ の T^s, T^d における出現確率をそれぞれ P_w^s, P_w^d とすると，単語 w のオッズ比は以下の式で表される．

$$OddsRatio(w) = \left(\frac{P_w^d}{1-P_w^d}\right) \Big/ \left(\frac{P_w^s}{1-P_w^s}\right)$$

ただし実際の計算では $P_w^s = 0$ となる可能性を考慮し，確率 P_w^d 及び P_w^s を計算する際に，実際の出現回数に 1 を加える補正を行う．オッズ比が $\theta\,(\geq 1)$ 以上であれば，単語 w が特徴的ユーザ特有の単語であると考えられる．さらに，出現確率 σ

*1 https://dev.twitter.com/

による制約を設け，ドメイン別辞書として，特徴的単語の集合

$$X^d = \{w \in W^d \,|\, P^d_w \geq \sigma, OddsRatio(w) \geq \theta\}$$

を抽出する．

2.2 単語クラスタ作成

一般にドメイン別辞書 X_d に含まれる単語数は非常に多いため，各単語の出現回数ベクトルをユーザプロファイルとしてしまうとプロファイルが過度に疎になってしまう．そこでドメイン別辞書の次元縮約を行う．

似たトピック，似た意味の単語は文字として異なっていたとしても，ユーザプロファイルを作成する際には同一視しても構わないという仮定のもと，次元縮約を行う．ここで，文字としては異なっているが意味としては似た単語の例として「ケータイ」と「スマホ」が考えられる．どちらの単語も「携帯電話」の意味で使われることから，文章中にケータイと一緒に出現する単語と，スマホと一緒に出現する単語は似ていると推測される．本研究ではこの特徴に着目し，ドメイン別辞書の次元縮約を試みる．

具体的には，下記に示す通り，ドメイン別辞書 X_d 中の単語 x_i を，特徴的ユーザらによるツイート群 T^d における単語 x_j との共起回数 $f_{co}(x_i, x_j)$ を用いてベクトル化した上でクラスタリングを適用し，次元縮約を実現する．

$$\vec{x_i} = (a_1, a_2, \cdots, a_{|X^d|})$$

where

$$a_j = \begin{cases} 0 & (i = j) \\ \frac{f_{co}(x_i, x_j)}{|T^d|} & (i \neq j) \end{cases}$$

クラスタリングには多くの手法があるが，本研究では，ユークリッド距離を用いたクラスタ数 $k = \lfloor |X^d| \times \alpha \rfloor \, (0 < \alpha \leq 1)$ の k-means 法を採用し，

$$C_i = \{x'_1, x'_2, \cdots, x'_{|C_i|}\}, \bigcup_{1 \leq i \leq k} C_i = X^d, \forall_{i \neq j} C_i \cap C_j = \emptyset$$

なる k 個の排他的単語クラスタ $C_i \, (1 \leq i \leq k)$ を獲得する．

2.3 ユーザプロファイル作成

単語クラスタを用いてドメイン別ユーザプロファイルを作成する．ユーザ u のツイート集合 T^u に含まれる名詞，形容詞，副詞の集合を W^u とし，ドメイン d における u のプロファイル $p(u, d)$ を，以下のベクトルとして獲得する．

$$p(u,d) = \left(\frac{freq(W^u, C_1)}{|T^u|}, \frac{freq(W^u, C_2)}{|T^u|}, \cdots, \frac{freq(W^u, C_k)}{|T^u|} \right)$$

ここで，$freq_{(w^u, C_i)}$ は，C_i 中の単語が出現する W^u 中のツイート数である．

3. ユーザ推薦

ドメイン別ユーザプロファイル構築の応用として，Twitter におけるユーザ推薦システムを想定し，2 つの推薦手法を提案する．第一の手法は，被推薦者自身のプロファイルと類似しているユーザを推薦する方法であり，第二の手法は，被推薦者のフレンドのプロファイルと類似しているユーザを推薦する方法である．また，推薦を行う際に利用するユーザプロファイル間の類似度にはコサイン類似度を利用する．以下では，ユーザ u_i と u_j の類似度を $cos(u_i, u_j)$ と表記する．

手法 1 被推薦者のプロファイルを利用した推薦

ツイートには投稿したユーザの嗜好・興味が現れると考え，被推薦者のプロファイルと類似しているユーザを推薦する．この手法では，被推薦者 u_i に対し，ユーザリスト $U \, (u_i \notin U)$ から類似度上位 N 人を推薦する．形式的には，u_i に対し，以下に示す $list^o_{u_i}$ を提示する．

$$list^o_{u_i} = \{u_j \in U \,|\, rank(u_i, u_j) < N\} \quad \text{where}$$
$$rank(u_i, u_j) = |\{u_k \in U \,|\, k \neq j, cos(u_i, u_k) > cos(u_i, u_j)\}|$$

手法 2 フレンドのプロファイルを利用した推薦

普段閲覧しているツイートはユーザにとって興味のある話題であると考え，被推薦者のフレンドとプロファイルが類似しているユーザを推薦する．形式的には，被推薦者 u_i に対し，そのフレンド F^i との類似度上位 N 人を推薦するため，以下のリスト $list^f_{u_i}$ を提示する．

$$list^f_{u_i} = \left\{u_j \in U \,|\, rank^f(u_i, u_j) < N\right\} \quad \text{where}$$
$$rank^f(u_i, u_j) = |\{u_k \in U \,|\, k \neq j, m(u_i, u_k) > m(u_i, u_j)\}|,$$
$$m(u_i, u_k) = \max_{u_f \in F^i} cos(u_f, u_k)$$

4. ケーススタディ

提案手法の妥当性を確認するため，料理ドメインを対象としたケーススタディを行った．

Twitter の情報を多視点からまとめいているツイナビ [*2] を参考に，料理に関するユーザであると分類されるユーザ 20 人を特徴的ユーザと判断し，そのツイートを (可能なかぎり) 全て獲得した．一方，2013 年 9 月から約 1 年間にわたりサンプルストリームを用いて獲得したツイートから，ランダムに約 800 万件をサンプリングし，一般ツイートとした．また，被推薦者として，特徴的ユーザのフォロワー約 4100 人を特定し，そのツイートとフレンドを獲得した．ツイート数は一人当たり約 2000 件であった．

実験では，ドメイン特有の単語として抽出するオッズ比に関するパラメタ θ を 5.0 と 1.0，また最低出現頻度パラメタ σ を 0.5%，クラスタ数を決定するパラメタ α を 10% と 100% と，いくつかのパラメタで評価を行った．詳細な結果は割愛するが，推薦の精度や多様性の面から，一定水準のプロファイルが得られ，またそれらが推薦に利用できることを確認した．

5. まとめと課題

本研究では，高精度な推薦に向けたドメイン別ユーザプロファイルの構築手法と，その応用としての Twitter のユーザ推薦手法を提案した．今後は，ドメイン別辞書の次元縮約方法の見直しやプロファイルの複合的な利用方法などに関して更なる検討を行っていく予定である．

参考文献

[Su 09] Su, X. and Khoshgoftaar, T. M.: A Survey of Collaborative Filtering Techniques, *Advances in Artificial Intelligence*, Vol. 2009, Article No. 4 (2009)

[神嶌 08] 神嶌敏弘：推薦システムのアルゴリズム (3)，人工知能学会誌, Vol. 23, No. 2, pp. 248–263 (2008)

*2 http://twinavi.jp/