3N-09

関係データ解析による Folksonomy 上のユーザモデリング

北澤 拓也，杉山 雅英（会津大学）

1. はじめに　Web ページや画像など多様なコンテンツに対してユーザが自由にタグ付けを行うことのできるシステムを Folksonomy と呼び，これまで推薦や知識発見を目標とした試みが多数なされてきた [1, 2]．しかしそれらはスパース性やタグの表記ゆれによるデータの扱いづらさ故に前処理など煩雑な手続きを要する．そこで本研究では，関係データ解析による単純かつ直感的なアプローチをとる．具体的には，無限関係モデル (IRM: Infinite Relational Model) [3] に基づいて Folksonomy 上のユーザ・コンテンツ間の関係をクラスタリングし，その結果からユーザの特徴付けを行う．

2. 無限関係モデルに基づくクラスタリング　IRM は関係データのクラスタリングに用いられるノンパラメトリックベイズモデルである．例として集合 S_1 と S_2 からなり，黒を 1，白を 0 とする 2 値の関係行列で表現されたデータを IRM に基づいてクラスタリングした様子を図 1 に示す．

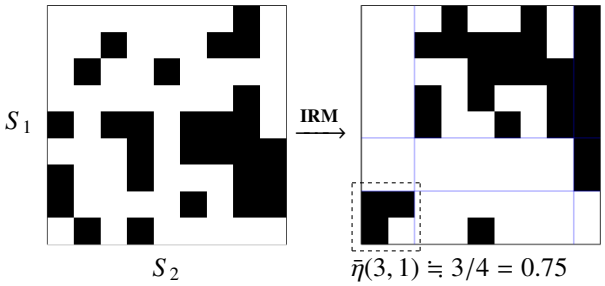

図 1　IRM に基づくクラスタリングの概念図

S_1 上の第 k クラスタ ($k = 1, 2, ..., N_1$) と S_2 上の第 ℓ クラスタ ($\ell = 1, 2, ..., N_2$) が成す矩形領域の 1 の存在確率の期待値 $\bar{\eta}(k, \ell)$ は IRM のパラメータ β (≥ 0) を用いて式 (1) で与えられる．なお $m(k, \ell)$，$\bar{m}(k, \ell)$ はそれぞれ，矩形領域内の 1 の数と 0 の数を示す．

$$\bar{\eta}(k, \ell) = \frac{m(k, \ell) + \beta}{m(k, \ell) + \bar{m}(k, \ell) + 2\beta} \tag{1}$$

本研究では Folksonomy 上のユーザ，コンテンツ，タグの集合をそれぞれ U，C，T と定義し，U の i 番目のユーザと C の j 番目のコンテンツの間に関係がある[1]場合に 1，そうでないときに 0 をとる関係行列 R に対して IRM に基づいたクラスタリングを行う．

∗ *User Modeling through Relational Clustering on Folksonomy,* T. Kitazawa, M. Sugiyama (The University of Aizu).

3. 実データのクラスタリング [4]　2. で述べた定義の下，我々は日本最大級のソーシャルブックマークサービス『はてなブックマーク』より収集したデータ[2]に対して IRM に基づくクラスタリングを行った．1,017 人のユーザ集合を U，7,000 件の Web ページ集合を C として行なった実験では，163 個のユーザクラスタ ($N_1 = 163$) および 8 個の Web ページクラスタ ($N_2 = 8$) が得られた．その一部を図 2 に示す．ここで U_k および C_ℓ はそれぞれ第 k ユーザクラスタ，第 ℓ Web ページ（コンテンツ）クラスタを意味し，T_ℓ は C_ℓ 内の Web ページに付与されたタグとその頻度のペアを要素に持つ集合を指す．

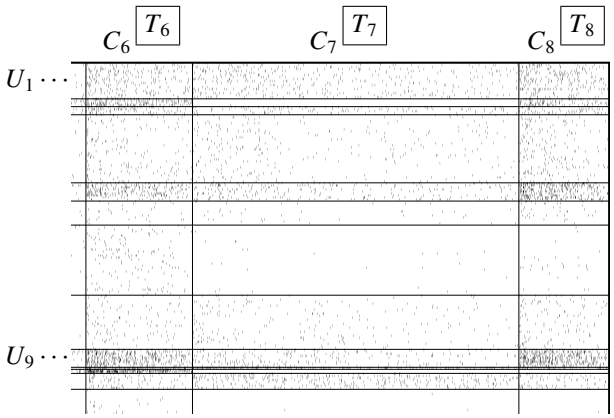

図 2　実データを用いた $U \times C$ のクラスタリング結果

図 2 中の U_9 と C_8 が交差する箇所は比較的濃く，1 が密集したブロックである．これは $\bar{\eta}(9, 8)$ が高い値になることを示しており，U_9 と C_8 の間に比較的強い関係があると解釈できる．式 (1) で計算すると $\bar{\eta}(9, 7) \fallingdotseq 0.02$ であるのに対し $\bar{\eta}(9, 8) \fallingdotseq 0.14$ となる．

4. Folksonomy 上のユーザモデリング　本節では $U \times C$ の関係データのクラスタリング結果を利用したユーザモデリング手法を提案し，3. のデータで実験を行う．

4.1 タグの重み付けによるユーザクラスタの特徴付け

3. の例のように $\bar{\eta}(k, \ell)$ が高い値を示す場合，タグ集合 T_ℓ 内のタグがユーザクラスタ U_k の特徴を表していると仮定する．そして tf-idf 重み付けを応用した手続きによって T_ℓ 内のタグの重み付けを行う．

[1] はてなブックマーク (http://b.hatena.ne.jp/) などのソーシャルブックマークサービスでは “ユーザが Web ページをブックマークしている”，写真共有サービスの Flickr (https://www.flickr.com/) では “ユーザが写真をお気に入りに登録している” などと解釈．

[2] 2014 年 6 月 4 日から 5 日にかけて収集したデータ．

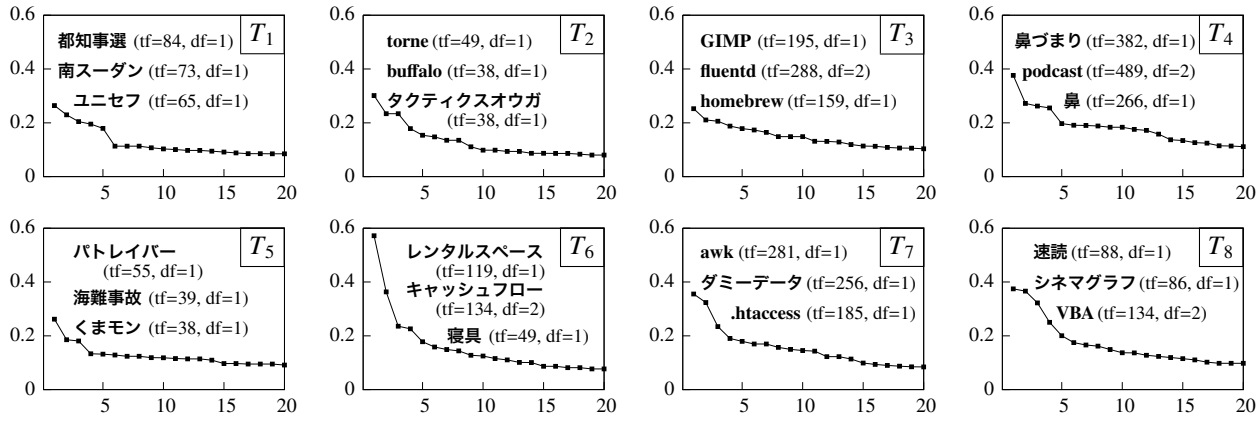

図 3　タグ集合 T_ℓ 上の Top-20 タグの重み $w_{t,\ell}$ と Top-3 タグ

以下，T_ℓ およびタグ $t \in T$ をそれぞれ tf-idf 重み付けにおける文書と単語に対応させる．すなわち $\mathrm{tf}_{t,\ell}$ は T_ℓ 内のタグ t の頻度，df_t はタグ t を含む T_ℓ の数を表す．そしてコンテンツクラスタ数 N_2 を用いて $\mathrm{idf}_t = \max\{0, \log((N_2 - \mathrm{df}_t)/\mathrm{df}_t)\}$ と定義し [5]，T_ℓ 上のタグ t の重み $w_{t,\ell}$ を式 (2) で求める．このとき $w_0 = \sqrt{\sum_{t \in T_\ell} w_{t,\ell}^2}$ で正規化し，異なる T_ℓ 間での重みの比較を可能にした．

$$w_{t,\ell} = \frac{\mathrm{tf}_{t,\ell} \cdot \mathrm{idf}_t}{w_0} \tag{2}$$

さらに式 (3) で正規化係数 $\bar{\eta}_0 = \sum_{\ell=1}^{N_2} \bar{\eta}(k,\ell)$ を用いて U_k におけるタグ t の重み $w_k(t)$ を計算する．

$$w_k(t) = \sum_{\ell=1}^{N_2} \frac{\bar{\eta}(k,\ell)}{\bar{\eta}_0} \cdot w_{t,\ell} \tag{3}$$

最終的に $w_k(t)$ の大きなタグがユーザクラスタ U_k を特徴付けているタグであるとみなす．

4.2 実データにおけるユーザのモデリングおよび考察

3. のクラスタリング結果を用いて提案手法によるユーザモデリングを試みた．はじめに式 (2) で計算されたタグ集合 T_ℓ ($N_2 = 8$) 上のタグの重みのうち Top-20 タグの重みを図 3 に示す．ここで横軸はタグの重み順位，縦軸は重みを表し，枠内には Top-3 タグを重みの降順で tf 値および df 値と共に記載した．

Top-3 タグに着目すると T_1 は “ニュース”，T_3 や T_7 は “IT” といった特徴が分かり，一方で T_8 のように雑多な集合も存在する．なお，idf 値をより一般的な定義 $\mathrm{idf}_t = \log(N_2/\mathrm{df}_t)$ で計算すると一時的な流行によるタグ[3]が複数のタグ集合で大きな重みとなり，その後のユーザモデリングの結果に強く影響することが確認された．

[3] df 値が中程度かつ極端に大きな tf 値を持つタグ．

図 4 は 3. で着目したユーザクラスタ U_9 に対するタグ t の重み $w_9(t)$ を式 (3) で計算した結果である．主に図 3 の T_8 で見られた雑多なタグの重みが大きく，U_9 に属するユーザは幅広い Web ページに関心があると推測できる．さらに w_9(VBA) が大きいことからプログラミングへの興味も読み取れる．

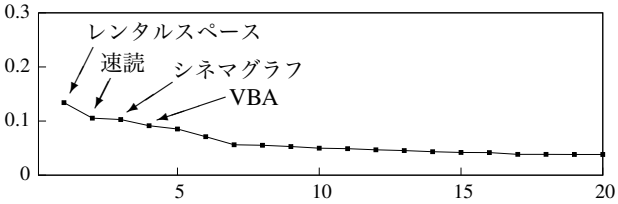

図 4　U_9 に対するタグ t の重み $w_9(t)$（Top-20 タグ）

3. のデータ収集から半年後，検証のため U_9 に属するユーザの最新ブックマークを再度取得した．その結果プログラミングに関連する Web ページが数件ブックマークされている一方で，残るほぼすべての Web ページが大衆的な話題を扱ったものであることが確認された．したがって提案手法がユーザクラスタの特徴を適切に捉えていたと言える．しかしユーザモデリング結果との顕著な関連はその他には無く，重み付け方法の改良やタグに対する前・後処理の必要性を示唆する結果となった．

5. むすび　本研究では関係データ解析による Folksonomy 上のユーザモデリング手法を提案し，実データを用いてその有用性と課題を明らかにした．更なる展望として関係データの時間変化やユーザとタグの関係 $U \times T$ を考慮することが挙げられる．また，多様な Folksonomy サービスを対象とした比較実験も検討する．

参考文献

[1] S. Niwa, et al., *Proc. of ITNG*, pp.388-393, (Apr. 2006). [2] A. Shepitsen, et al., *Proc. of RecSys2008*, pp.259-266 (Oct. 2008). [3] C. Kemp, et al., *Proc. of AAAI2006*, pp.381-388 (July 2006). [4] T. Kitazawa, M. Sugiyama, *Proc. of ECEI2014*, 2A05 (Aug. 2014). [5] C. D. Manning, et al., “Introduction to Information Retrieval,” Cambridge Univ. Press (2008).