D-012

フォークソノミーにおけるタグの意味的関係分析に関する一考察

A Study on Analysis of Tag Semantic Relationship in Folksonomy

岸端 佑季† 雲居 玄道‡ 後藤正幸†† 東基衛††
Yuki Kishibata Gendo Kumoi Masayuki Goto Motoei Azuma

1．まえがき

近年，"del.cious.us[1]"，"はてなブックマーク[2]"などで注目を集めているソーシャルブックマークサービス(SBS)において，Web ページの分類形態として個々のユーザがタグ付けを行うフォークソノミーが採用されている．

フォークソノミーは，少数の専門家が時間をかけ Web ページを分類する必要がないため，ブログやニュース記事などの鮮度の高い情報の分類に適している．その一方で，個々のユーザが自由にタグを付与することができるため，ユーザごとの表記の差異や様々な抽象度を持ったタグが混在してしまい，ユーザがタグを指定してページを検索する際に，検索が困難になるといった問題がある．

この問題の原因として，検索システムがタグを文字列の並びで処理することで，タグ同士における関連性，同義性の判定ができないことが挙げられる．よって，タグ同士の同義性，関連性といった関係を自動で推定することが求められている．

Xian ら[3]は，タグ同士の関係を推定するため，関連性のあるタグ同士は同一，もしくは類似したページに付与されやすく，似た嗜好をもつユーザに利用されやすいというフォークソノミー特有の性質から，ユーザ，Web ページ，タグは互いに関連性があると考え，それらの関係性を PLSI（Probabilistic Latent Semantic Indexing）モデルを導入して確率モデル化した．しかし，定式化したモデルでは，タグ同士の関連性は表現できるが，同義性までは表現することはできていない．

丹波ら[4]はフォークソノミーの性質として，同一ページに付与されるタグ同士は同義語である可能性が高く，同一ユーザが利用するタグ同士は同義語ではない可能性が高いと提唱している．すなわち，ページ間，ユーザ間でのタグ共起の差異を分析することで，タグの意味的関係を把握できることを示唆している．

本研究では，Xian らの研究をもとに，丹羽ら[4]が提唱したフォークソノミーの性質を利用し，タグ同士の同義性を表現することができる確率モデルを提案した．

2．従来研究

2.1 PLSI モデル

本研究で利用する PLSI モデルについて説明する．

PLSI[5]は，情報検索のための確率モデルとして Hofmann によって提唱された．PLSI では，文書 d_i (i=1, 2, …, I) と単語 w_j (j=1, 2, …, J)に対して，これらのデータが意味的な隠れクラス変数 c_m (m=1, 2, …, M)を介したとき，文書 d_i における w_j の出現確率は，

$$P(d_i, w_j) = \sum_{m=1}^{M} p(d_i \mid c_m) p(w_j \mid c_m) p(c_m) \quad (1)$$

で与えられることを仮定する．

(1)式から，EM アルゴリズムにより，$p(d_i \mid c_m)$, $p(w_j \mid c_m)$, $p(c_m)$ を最尤推定することができる．

推定した $p(w_j \mid c_m)$，$p(c_m)$ から，ベイズの定理を利用することにより，単語 w_j のクラス変数 c_m への帰属確率 $p(c_m \mid w_j)$ は

$$p(c_m \mid w_j) = \frac{p(w_j \mid c_m) p(c_m)}{p(w_j)} \quad (2)$$

と，単語のもつ意味を確率的に表現することができ，帰属確率 $p(c_m \mid w_j)$ の分布の類似性が意味の類似性を表すことを提案している．

2.2 Xian ら[3]の研究

フォークソノミーの性質として，似た嗜好をもつユーザは，類似したページにタグを付与する傾向があり，付与されるタグも，そのようなユーザ間では関連性のあるものが多い．また，類似したページには似た嗜好をもつユーザがタグを付与することが多く，類似したページ間では付与されるタグも関連性のあるものが多い．

この性質から，Xian らはユーザ，タグ，Web ページは互いに隠れクラス変数を介して関連性があると考えた．

そこで，SBS 上のユーザ，Web ページ，タグのそれぞれの集合を $\boldsymbol{U}$, $\boldsymbol{R}$, $\boldsymbol{T}$, 隠れクラス変数 c_m (m=1, 2, …, M)の下でのユーザ $u_i \in \boldsymbol{U}$, ページ $r_j \in \boldsymbol{R}$, タグ $t_k \in \boldsymbol{T}$ の生起確率をそれぞれ $p(u_i \mid c_m)$, $p(r_j \mid c_m)$, $p(t_k \mid c_m)$, クラス変数 c_m の生起確率を $p(c_m)$ とすると，クラス変数 c_m を介した u_i, r_j, t_k の同時出現確率 $P(u_i, r_j, t_k)$ は，PLSI モデルを導入することにより

$$P(u_i, r_j, t_k) = \sum_{m=1}^{M} p(u_i \mid c_m) p(r_j \mid c_m) p(t_k \mid c_m) p(c_m) \quad (3)$$

として，u_i, r_j, t_k の関係性を表現する確率モデルとして定式化することができる．そして，最尤推定により，タグ t_k のクラス変数 c_m への帰属確率 $p(c_m \mid t_k)$ を求め，帰属確率の分布の類似度からタグ同士の関連性を推定することができる．

3．提案手法

Xian らの確率モデルから得られる帰属確率 $p(c_m \mid t_k)$ の分布の類似度では，タグ同士の関連性を推定することはできるが，同義性は推定することができない．一方，丹波ら[4]に示されているように，ページ間，ユーザ間でのタグ共起の差異を分析することで，関連性，同義性などのタグの意味的関係を把握できる可能性がある．

そこで本研究では，新たにタグ同士の同義性を推定するために，ページとタグの関連性を表現する確率モデル

† 早稲田大学創造理工学研究科
‡ 早稲田大学理工学研究所
†† 早稲田大学創造理工学部

とユーザとタグの関連性を表現する確率モデルの２つのモデルを推定し，それらの差異を分析することで，タグ間の意味的関係を把握する方法を提案する．

なお，本章で，数式の中で使われる記号，変数は前節 2.2 と同じものとする．

3.1 ページとタグの関係性を表現するモデル

丹波らはフォークソノミーの性質として，同一ページに付与されているタグ同士は，同じ文章の特徴を表しているため，同義語か関連性がある可能性が高いと提唱している．

この性質を考慮して，ページとタグは隠れクラス変数を介して関連性があると考えられる．そこで，ページ r_j におけるタグ t_k の出現確率 $p(r_j,t_k)$ を，PLSI モデルを導入することにより

$$p(r_j,t_k)=\sum_{m=1}^{M} p_R(r_j\mid c_m)p_R(t_k\mid c_m)p_R(c_m) \qquad (4)$$

として r_j と t_k の関係性を表現する確率モデルを定式化することができる．

次に，(4)式から EM アルゴリズムを用いて，最尤推定された $p_R(t_k\mid c_m)$，$p_R(c_m)$ から，ベイズの定理を利用して，タグ t_k のクラス変数 c_m への帰属確率 $p_R(c_m\mid t_k)$ を求める．

$$P_R(c_m\mid t_k)=\frac{P_R(t_k\mid c_m)P_R(c_m)}{\sum_m P_R(t_k\mid c_m)P_R(c_m)} \qquad (5)$$

ページとタグの関連性を表わす PLSI モデルから算出された帰属確率 $p_R(c_m\mid t_k)$ の分布の類似性を測ることによって，タグ同士の関連性を推定することができると考えられる．ここで計算される帰属確率分布は，任意のページにおけるタグの出現確率をもとに算出された分布なので，類似した帰属確率分布をもつタグ同士は，“同一のページに付与されている”という意味での関連性が高いペアであるといえる．

そこで，この(5)式によって計算される帰属確率分布に基づくタグ t_k と $t_{k'}$ との類似度を $Sym_R(t_k, t_{k'})$と定義する．ここで，確率分布間の類似性（非類似性）を測るための指標として，確率分布間の距離指標である KL(Kullback-Leibler)-Divergence[6]，JS(Jensen-Shannon)-Divergence[7]，Skew-Divergence[8]や，ベクトル間の距離尺度である内積，コサイン尺度などを考えることが可能である．

$Sym_R(t_k, t_{k'})$の値が大きいタグ同士は同義語であるか関連性が強い可能性が高いといえる．

3.2 ユーザとタグの関係性を表現するモデル

同義語のタグが生じるのは，異なるユーザ間でどのタグを使うかに関する同意が取れていないことに起因している．よって，丹波ら[3]はフォークソノミーの性質として，同一ユーザに利用されているタグ同士は，同義語である可能性は低いが，関連性がある可能性が高いと提唱している．

この性質を考慮して，ユーザ u_i とタグ t_n は隠れクラス変数を介した関連性に着目する．そこで，ユーザ u_i におけるタグ t_n の出現確率 $p(u_i,t_n)$ を，PLSI モデルを導入することによって

$$p(u_i,t_n)=\sum_{m=1}^{M} p_U(u_i\mid c_m)p_U(t_n\mid c_m)p_U(c_m) \qquad (6)$$

として u_i と t_n の関係性を表現する確率モデルを定式化することができる．

次に，(6)式から最尤推定された $p_U(u_i\mid c_m)$，$p_U(c_m)$ から，タグ t_n のクラス変数 c_m への帰属確率 $p_U(c_m\mid t_n)$ を求める．

$$P_U(c_m\mid t_k)=\frac{P_U(t_k\mid c_m)P_U(c_m)}{\sum_m P_U(t_k\mid c_m)P_U(c_m)} \qquad (7)$$

算出された帰属確率 $p_U(c_m\mid t_n)$ の分布の類似度からタグ同士の類似性を推定することができると考えられる．この帰属確率分布は，任意のユーザにおけるタグの出現確率をもとに算出された分布なので，類似した帰属確率分布をもつタグ同士は“同一のユーザに利用されている”という意味での関連性が高いペアであるといえる．

(7)式によって計算される帰属確率分布に基づくタグ t_k と $t_{k'}$ との類似度を $Sym_U(t_k, t_{k'})$と定義する．$Sym_U(t_k, t_{k'})$の値の大きいタグ同士は関連性は高いものの，同義語である可能性が低いと考えられる．

3.3 同義性の推定

3.1，3.2節で定義されたタグ t_k, $t_{k'}$間の類似度 $Sym_R(t_k, t_{k'})$ と $Sym_U(t_k, t_{k'})$を用いて，タグ同士の同義性を測る指標 $Syn(t_k, t_{k'})$を次のように定義する．

$$Syn(t_k,t_{k'})=\frac{Sym_R(t_k,t_{k'})}{Sym_U(t_k,t_{k'})} \qquad (8)$$

ここで，しきい値 α（αは非負の定数）により $Syn(t_k, t_{k'})>\alpha$ であれば，タグ t_k とタグ t_h は同義語であり，$Syn(t_k, t_{k'})<\alpha$ であれば同義語ではないと推定する．

4. まとめ

本研究ではフォークソノミーにおけるタグの同義性の新しい推定法を提案した．この手法は，丹羽ら[4]らが提唱したページ-タグ間，ユーザ-タグ間の関係性を Xian ら[3]による PLSI モデルに適用し，確率モデルによってタグ同士の意味的関係を分析することが可能である．

参考文献

[1]del.cious.us http://del.icio.us.

[2]はてなブックマーク http://b.hatena.ne.jp

[3]Xian Wu 他, “Exploring Social Annotations for the Semantic Web”, In WWW, pp417-426, (2006).

[4]丹波智史, 土肥拓生他, “Folksonomy の 3 部グラフ構造を用いたタグクラスタリング”, 人工知能学会研究会資料, (2006).

[5] Thomas Hofmann ,“Probabilistic Latent Semantic Indexing”, Proc. of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR 99) pp.50-57, (1999).

[6] S.Kullbak,R.A.Leibler ,“On Information and Sufficiency”,Annnals of ,Mathematical Statistics,vol22,no.1,pp.79-86(1951).

[7] Jianhua Lin, “Divergence measures based on the shannon entropy”, IEEE Transactions on Information Theory,37(1),pp.145-151,(1991).

[8]Lillian Lee. “On the Effectiveness of the Skew Divergence for Statistical Language Analysis”, Artificial Intelligence and Statistics 2001, pp65-72(2001).