4T-3

ソーシャルブックマークにおけるユーザとタグの関連度抽出法の検討*

夏目　大伍†　　入部　百合絵‡　　桂田　浩一†　　新田　恒雄†
†豊橋技術科学大学大学院工学研究科
‡豊橋技術科学大学情報メディア基盤センター

1. はじめに

近年，Web 上に溢れる大量のデータからユーザの特徴や特性などを分析し，その結果をシステムに活用する研究が盛んに行われている．そこで，ユーザの嗜好を抽出する研究が行われているが，それらの情報を抽出するためのデータを十分に獲得することは難しい．

本研究では，近年普及しているソーシャルブックマークに着目し，そこで蓄積されたデータからユーザの嗜好を獲得することを目指す．ソーシャルブックマークでは Web ページをブックマークする際に，Folksonomy の考えに基づき自由にタグが付与される結果，大量のタグデータが存在している[1]．また，ブックマークするということは自然な行為であり，特に Web ブラウザのブックマークの場合，ブックマーク内のフォルダに付けた名前などはユーザの関心や興味などを表現したものであると言われている[2]．ソーシャルブックマークの場合にも，ユーザが付与したタグは，上記ブックマーク内のフォルダ名と同じ働きを持つものだと考えられる．

本研究は，ユーザが付与したタグを分析することでユーザの嗜好や関心を抽出することを目的しており，その手段としてユーザとタグの関連度抽出法を検討する．

2. Folksonomy とソーシャルブックマーク

近年，Folksonomy を利用したサービスであるソーシャルブックマークが国内外で急速に普及し始めている．以下でこの特徴を簡単に述べる．

2.1 Folksonomy

Folksonomy とは従来の Taxonomy に代わる新しい分類手法である．Taxonomy はあらかじめ決められたキーワードを使用して分類するという伝統的な分類方法である．例としては，図書館司書による本の分類や Yahoo ディレクトリに基づく Web ページの分類が挙げられる．これに対して Folksonomy では，多数のユーザが「タグ」と呼ばれる自由に付与できるキーワードを用いて分類を行う．「タグ」は個々のユーザにとって，そのページの分類や性質を表すキーワードであり，ユーザの価値観に基づいて自由にキーワードを選ぶことができる．タグは 1 つの Web ページに対して複数付与することができる．

2.2 ソーシャルブックマーク

ソーシャルブックマークは Web サイト上で不特定多数のユーザとブックマークを共有するサービスである．Folksonomy に基づき，ブックマークした Web ページに対して自由にタグを付与することができる．また，ブックマークした Web ページやタグにより，ユーザ同士のブックマーク情報を関連付けることが可能である．これにより，多数のユーザがリンクされるので，大量のデータを取得することができる．ソーシャルブックマークの代表的なサービスとしては，「del.icio.us」[3]や「はてなブックマーク」[4]が挙げられる．

3. 関連度の抽出

3.1 データの取得

日本最大のソーシャルブックマークである「はてなブックマーク」では，登録されたユーザのブックマークデータが RSS フィードとして公開されている．そこで本研究では，はてなブックマークの人気エントリーをブックマークしているユーザの RSS フィードを取得し，その内容を解析することで Web ページのタイトル，URL，付与されたタグを抽出した．この方法により，6,625 人分のデータを取得した．ユーザによりブックマークされた Web ページの総数は 89,685 ページで，使用されたタグは 19,846 種類あった．使用人数が少ないタグはノイズの可能性が高いため，本研究では使用人数が 5 人以上のタグのみ用いた．このため，実際に使用したタグは 2,770 種類である．

3.2 ユーザの付与したタグへの重み付け

ユーザを特徴付けるために，ユーザが付与したタグに対し重み付けを行い，その算出には TF-IDF を用いた．TF-IDF は，主に文書中に出現する単語に対して重み付けを行う方法であり，文書を特徴付ける単語ほど大きな重み算出される．

本研究では，ユーザ毎のタグ一覧(タグ名とその使用回数を一覧としたもの)に対して TF-IDF を適用した．TF を算出する式を(1)式に，IDF を算出する式を(2)式に示す．

$$TF(U,T)=\frac{w(U,T)}{\sum_{T_i \in TAGS} w(U,T_i)} \quad (1)$$

$$IDF(T)=\log \frac{\sum_{U_j \in USERS}\sum_{T_i \in TAGS} w(U_j,T_i)}{\sum_{U_j \in USERS} w(U_j,T)} \quad (2)$$

ここで，$w(U,T)$ はユーザ U がタグ T を付与した回数である．(1)式はユーザ U が付与した全てのタグに対してタグ T の占める割合を，(2)式は取得した全データにおけるタグ T の希少性を示すものである．なお，TF-IDF は TF と IDF の積で求める．

3.3 タグ間の関連度の算出

ユーザとタグの関連を考えた場合，ユーザが付与したタグ以外にもユーザと関連するタグは存在すると考えられる．例えば，ユーザが付与したタグと同じ意味を持つタグや，ユーザが付与したタグから連想されるようなタグ(例：「Java」と「Eclipse」)は，ユーザとの関連度が高いと予想される．そこで，タグ間の関連度を求めるために以下の手法を用いた．

3.3.1 共起行列の作成

タグ間の関連度を抽出するために，ユーザが付与したタグを Web ページ毎に整理し，タグ同士の共起回数から共起ベクトル及び共起行列を作成した．作成した行列は，タグ同士の関連を求める行列のため，n 次正方行列となる．この行列を図 1 に示す．なお，a_{ij} の値はタグ t_i におけるタグ k_j の共起回数である．また，各行は注目しているタグ，各列はそれらと共起したタグを表す．

3.3.2 コサイン尺度を用いたタグ間の関連度の算出

コサイン尺度を用いてタグ間の関連度の算出を行った．関連度を算出する式を(3)式に示す．

$$rel(t_i,t_j)=\frac{\mathbf{t_i}\cdot\mathbf{t_j}}{\left|\mathbf{t_i}\right|\left|\mathbf{t_j}\right|} \quad (3)$$

ここで，$\mathbf{t_i}$ はタグ t_i の共起ベクトル，$\mathbf{t_j}$ はタグ t_j の共起ベクトルとする．関連度は 0 から 1 の間の値で表される．

---

*Study on methods to extract user - tag relations in social bookmark
D.Natsume†，Y.Iribe‡，K.Katsurada†，T.Nitta†
† Graduate School of Engineering, Toyohashi Univ. of Tech.
‡ Information and Media Center, Toyohashi Univ. of Tech.

3.3.3 Simpson 係数を用いたタグ間の関連度の算出

コサイン尺度以外の指標として Simpson 係数を用いてタグ間の関連度を算出した．Simpson 係数は，関係の強さを図る指標の一つで，付与された Web ページが少ないタグをベースとした共起の強さを表している．欠点としては，あるタグが付与された Web ページが極端に少ない場合，非常に高い値が算出されることである．そのため，閾値などを設定することにより．それらを排除する必要がある．Simpson 係数を算出する式を(4)式に示す．

$$S(t_1,t_2)=\begin{cases}\dfrac{|T_1\cap T_2|}{\min(|T_1|,|T_2|)} & T_1>k \text{ and } T_2>k\text{の場合}\\ 0 & \text{それ以外}\end{cases} \quad (4)$$

なお，T_1 はタグ t_1 が付与された Web ページ数，T_2 はタグ t_2 が付与された付与された Web ページ数とする．閾値 k は 9 とした．

4. 結果と考察

はてなブックマークでは，パソコンや Web 関係の Web ページが頻繁にブックマークされる傾向があるので，パソコンや Web に関連したタグの使用数が多くなる．このため，TF-IDF による重み付けを行った結果，多くのユーザのパソコン，Web 関係のタグの重みが高くなっている．本研究では，使用人数が少ないタグをデータから除外しているため, IDF の値が極端に大きくならないことも原因の一つであると考えられる．

タグ間の関連度に関しては，コサイン尺度を用いた場合と Simpson 係数を用いた場合では求めた結果に違いが存在する．「Music」タグの関連タグを算出する際にコサイン尺度を用いた場合の結果を表 1 に, Simpson 係数を用いた場合の結果を表 2 に示す．

ソーシャルブックマークを分析する際には，同じような意味で表記の異なるタグが複数存在するために問題となる[5]．コサイン尺度を用いた場合，このような類似タグが関連度の高いものとして上位に現れる傾向が確認された．はてなブックマークでは，同一の Web ページに対して同じ意味を持った表記の異なるタグが複数付与されるため，必然的にこれらのタグ同士は共起回数が増える．このため，コサイン尺度を用いて関連度を算出した場合は，同じ意味で表記の異なるタグの関連度が高くなるものと考えられる．

Simpson 係数を用いた場合は，コサイン尺度を用いた場合よりも，注目しているタグから連想されるタグが上位に現れる傾向が強い．Simpson 係数では，分母に min をとっているので使用回数がそれほど多くないタグでも，関連度が高くなる可能性がある．例えば，ゲーム機本体とゲームソフトについて考える．ゲームソフト名がタグとして付与されている Web ページに高頻度でゲーム機本体の名前がタグとして付与されているならば，ゲームソフトからみてゲーム機本体は関連度が高くなると推測される．このことから，Simpson 係数では連想語のようなタグが上位に出現すると推察される．

$$\begin{array}{c} \\ \text{注目しているタグ} \end{array}\begin{array}{c} \overbrace{}^{\text{共起したタグ}} \\ k_1\ \ k_2\ \ \cdots\ \ k_n \end{array}\begin{array}{c} \\ \begin{cases} t_1 \\ t_2 \\ \vdots \\ t_l \end{cases} \end{array}\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{l1} & a_{l2} & \cdots & a_{ln} \end{pmatrix}$$

図 1　共起行列

表 1　「Music」についてコサイン尺度を用いた場合の結果例

音	0.84679
音楽	0.8216
audio	0.73939
ROCK	0.71674
音楽配信	0.71396
ファミコン	0.69041
guitar	0.67647

表 2　「Music」について Simpson 係数を用いた場合の結果例

lyrics	0.8
Perfume	0.733333
歌	0.583333
ROCK	0.583333
mix	0.545455
jazz	0.5
Last.fm	0.458333

ユーザの嗜好を抽出する場合，ユーザが付与したタグ以外にも候補となるタグが多い方が，嗜好を幅広く推定できる可能性がある．従って，コサイン尺度と Simpson 係数を適切に用いることで，ユーザに関わる有益な情報の獲得が期待される．

5. まとめ

本報告では，タグへの重み付け及びタグ間の関連度の算出を行うために複数の抽出方法を試行し，各手法の特徴や問題点について考察を行った．

今後の課題は，被験者による評価を行うことにより，タグ間の関連度を算出する最も有効な手法について検討していくことである．さらに，その結果に基づき，ユーザとタグの関連度を求めていく予定である．

参考文献

[1]丹羽智史，他：“Folksonomy マイニングに基づく Web ページ推薦システム”，情報処理学会論文誌 Vol.47 No.5，pp.1382-1392，2006.

[2]濱崎雅弘，他：“Bookmark からの共通話題ネットワークの発見手法の提案とその評価”，人工知能学会論文誌 17 巻 3 号 SP-D，pp.276-284，2002.

[3]del.icio.us：http://del.icio.us/

[4]はてなブックマーク：http://b.hatena.ne.jp/

[5] Scott A. Golder ， Bernardo A. Huberman :“The Structure of Collaborative Tagging Systems”，HP Labs Technical Report，http://www.hpl.hp.com/research/idl/papers/tags/，2006.