E-049

Folksonomyにおけるタグ語の意味的階層関係の抽出

Extracting Semantic Hierarchy of Folksonomy Tag Words

杉本 徹† 五十嵐 幹‡
Toru Sugimoto Motoi Igarashi

1．まえがき

現在の Web 上には大量の情報が氾濫し，多くのコンテンツが埋没してしまっている．情報の多さや多様さが原因となって，権威的な体系化も困難となっている．この問題に対してコンテンツにメタデータを付与し，埋没を防ごうとする試みがあり，その具体例として，ユーザが自分の分類のためにコンテンツに対して自由にタグ付けをした情報をメタデータとして利用する Folksonomy が挙げられる．

Folksonomy ではタグ付けする言葉をユーザが自由に決められるという手軽さと柔軟性がある一方，その自由さから，タグ語の揺れや詳細度の違いによって検索の際の再現率が下がったり，タグ付けする際にどんな語をタグとして付けたらよいか迷ったりすることがある．

本研究ではユーザによって付けられたタグ語の相互関係のモデル化，特に意味的な階層関係の自動的な抽出を行うために，コンテンツに付与されたタグ語の共起関係を利用する方法を検討する．そしてその結果を用いてコンテンツをより有効に活用することを目指す．

2．研究の手順

本研究では，ソーシャルブックマークサービス「はてなブックマーク」（http://b.hatena.ne.jp/）に蓄積された情報を題材として利用し，以下の手順で研究を進める．

① 「はてなブックマーク」のタグ検索を用いて，タグ付けされた URL（以下コンテンツ）を収集する．
② はてなブックマークエントリー情報取得 API を用いてコンテンツに付与されたタグ情報を取得する．
③ コンテンツごとに集めたタグを集計する．
④ 集計結果から，2つのタグ語間の共起確率を求める．
⑤ 共起確率の分布からタグ語間の階層関係を推測する．
⑥ 推測したタグ語間の上位－下位関係と人手で判定した上位－下位関係を比較し，本手法の評価を行う．

2.1 コンテンツの収集

本研究では，787 万件（06 年 12 月段階）という十分な量のコンテンツがあり，コンテンツに対して誰にでもタグ付けが許されている，タグ情報が API で取得できる，また日本語であるといった点からソーシャルブックマークサービス「はてなブックマーク」を題材として扱う．

「はてなブックマーク」でタグ検索を行った結果を表示する Web ページの内容から，対象となるタグ語が付与されたブックマークの URL を抽出し，「はてなブックマークエントリー情報取得 API」を用いることでコンテンツに付与されたタグ列の情報を JSON データ型で取得した．なお，本研究では実験のサンプルとなるタグ語として「Java」というタグ語に共起したタグ語を中心に 26 種類を選定し用いることにした．

† 芝浦工業大学 工学部 情報工学科
‡ 北陸銀行

2.2 収集したタグの集計

タグの組み合わせ方や，表現に用いるタグの選び方はユーザによって異なる．しかし，同じコンテンツに付与されたタグには強い関連性があると推測される．このことからタグ語の相互関係を明らかにする上でその共起関係を用いることにした．共起関係とは，2 つのタグが同じコンテンツ内に同時に出現する関係のことである．

1 つのコンテンツに付与されたタグは複数のユーザにより複数回付与されていたとしても 1 回として集計し，対象としたタグ a を含むコンテンツすべてで共起タグ b の出現回数を求め，そこからタグ a，b の共起確率

$$P(\mathrm{b}\mid\mathrm{a}) = \frac{\text{タグ}a\text{、}b\text{を両方含むコンテンツの数}}{\text{タグ}a\text{を含むコンテンツの数}}$$

を求めた．またこの過程で，表記揺れやノイズを手動で統合・排除する処理を行った．

2.3 タグ語の階層関係の抽出

タグが付与されたコンテンツの集合の包含関係を用いて，タグ語間の意味的な階層関係を抽出する．ここで意味的階層関係としては，「一般－特殊」関係だけでなく「全体－部分」関係なども含めた広い意味の上位－下位関係を想定している．

下位語，例えば「Java」というタグを持つコンテンツに上位語である「programming」というタグが共起する可能性は高いが，逆は成り立たない．このようなタグ付けの実態を利用して上位－下位関係を推定する．

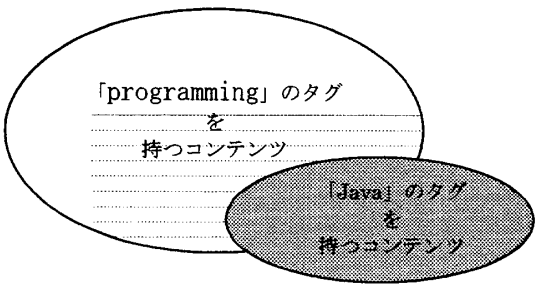

図 1：階層関係にあるタグ語を持つコンテンツの分布例

タグ語 a と b が意味的な階層関係にあり，a が b の上位語であるならば，理想的には P(a|b)=1 となるはずである．実際にはタグ付けが各々のユーザに任されるために，本来付けられるべきタグの付与漏れが避けられず，そのタグが付けられるべきコンテンツの一部分にしか付与されていない．そのため P(a|b) は必ずしも 1 にはならないが，それに近い大きな値になるはずである．

一方，タグ語 b は a の下位語であるから，逆向きの条件付き確率 P(b|a) は小さい値になる．仮に P(a|b) と P(b|a) の値がともに大きく，かつ近いものは同階層にある類義語と呼ぶべきものになると考えられる．

そこで，P(a|b) の大きさ，および P(a|b) と P(b|a) の比に着目して階層関係の判断をすることにした．すなわち，あるタグ語のペアに対し，これらの尺度が決められた閾値以上でありかつ P(a|b) > P(b|a) である場合にタグ語 a はタグ語 b の上位語であると判定する．サンプルとして選んだ 26 個のタグ語を組み合わせて得られる全 325 通りのタグ語ペアに対し，それらが上位語－下位語の関係にあるかどうかの判定を行った．

3. 評価

本手法の評価は，前節の最後に述べた全 325 通りのタグ語ペアに対する本手法に基づく判定結果と，人間による判断結果を比較することにより行った．人間による判定は，情報系の専門知識を持つ大学生 4 人が独立に行い，そのうち 3 人以上の見解が一致した 37 組のタグ語ペアについて上位－下位関係にあると認定した．

条件付き確率 P(a|b) の大きさを階層関係判定の尺度とし閾値を変えていった際に，人間による判定を正解とした適合率，再現率を求めた結果を図 2 に示す．また，P(a|b) が大きい値となるタグ語ペアのリストを表 1 に示す．

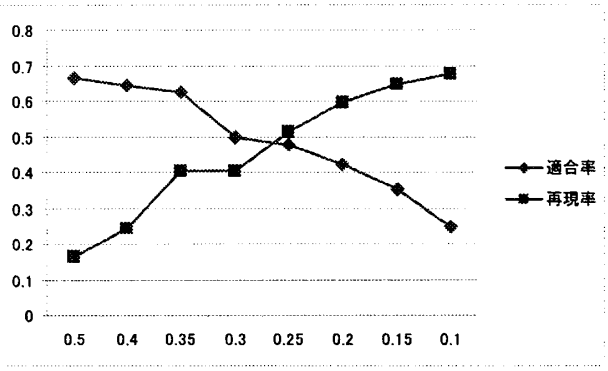

図 2: P(a|b) の大きさに基づく階層関係抽出の評価

表 1: P(a|b) の値が大きいタグ語ペア（上位 15 件）

（上位）→（下位）	P(a|b)	評価
java → j2ee	0.939	○
programming→オブジェクト指向	0.746	○
programming → c	0.601	○
java → eclipse	0.555	○
design → css	0.545	○
web → css	0.541	×
mac → apple	0.530	×
programming → 開発	0.515	×
software → アプリケーション	0.513	○
javascript → library	0.463	×
software → windows	0.461	○
開発 → オブジェクト指向	0.435	○
web → ブログ	0.425	○
computer → hardware	0.410	×
web → design	0.398	×

図 2 によれば，P(a|b) > 0.35 かつ P(a|b) > P(b|a) という判定条件にした場合に，適合率が約 6 割，再現率が約 4 割という値になる．そこで適合率を下げている原因を分析したところ，次の 3 つの原因があることが分かった．

① 人間による判断の揺れ，曖昧さ
表 1 における「web → css」や「computer → hardware」のように，意味の捉え方により階層関係にあるとも無いとも判断できそうなペアが存在する．

② タグ語の人気度の差による，判定される関係の逆転
表 1 における「mac → apple」や「programming → 開発」のように，上位－下位が逆向きに判定されてしまう．その原因は，多くのユーザによって好んで付けられるタグ語（この例では，mac や programming）とそうでないタグ語が存在することによる．

③ タグ語の状況限定的な使用
表 1 における「javascript → library」や「web → design」は，「library」，「design」という語をそれぞれ「javascript の library」，「web の design」と考えれば，上位－下位関係にあると見なすことも可能である．

これらの中で②に関しては，タグ語自体の人気度を加味した判定を行えるように改良する方法を検討する必要があるが，それを除けばおおむね良い抽出結果が得られたのではないかと考えている．

なお，本研究においてタグ語間の階層関係を抽出するためのもう 1 つの尺度として検討していた P(a|b) と P(b|a) の比の大きさに関しては，実験の結果十分な適合率が得られないことが分かった．

4. 結論

同一のブックマークに付与された 2 つのタグ間の共起関係の累積から語の意味の包含関係を推測し，上位語－下位語というような語の階層関係の抽出を試みた．精度に関しては改善すべき点も残されているが，Folksonomy においてユーザが自由に付けたタグ語の出現分布から流行のキーワード間に存在する意味的関係を自動抽出する可能性を示すことができたと考えている．

このようにして抽出した意味的関係は，タグ検索の精度向上や，ユーザがタグ付けする際の候補語提示に応用できると思われる．検索質問を自動的に拡張する際の拡張の仕方や，ユーザに提示するタグ語の分類表示など，単に類義語を利用する場合に比べてきめの細かい選択が行える可能性があると考えている．

参考文献

[1] 丹羽智史，土肥拓生，本位田真一：Folksonomy の 3 部グラフ構造を利用したタグクラスタリング，人工知能学会　セマンティックウェブとオントロジー研究会，2006.

[2] Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su : Towards the Semantic Web: Collaborative Tag Suggestions, WWW 2006.

[3] Patrick Schmitz : Inducing Ontology from Flickr Tags, WWW 2006.