E-001

ソーシャルブックマークを用いたウェブサイトの意味内容の抽出

Extracting Semantic Content from Website using Social Bookmark

長谷川 直広† 黄 潤和‡
Naohiro Hasegawa Runhe Huang

1. まえがき

ソーシャルブックマークと呼ばれる，インターネット上で複数のユーザがウェブサイトのブックマークを共有できるサービスに関する研究が注目されている．ソーシャルブックマークでは，ブックマークをする際にウェブサイトに対して好きなタグを複数付与することができ，ユーザは他人と共有したり，過去にブックマークしたウェブサイトから検索を容易にしたりするためにタグを付与する．付与されるタグは一般にウェブサイトのカテゴリや内容を表すことが多く，またユーザの趣味趣向を反映していることから様々な方面へ応用できる有益な情報である．

ソーシャルブックマークに関連する研究として，ユーザのウェブサイトの閲覧履歴とソーシャルブックマーク情報を用いたタグの推薦[1]があるが，ウェブサイトの特徴語を抽出する際に，ウェブサイトのコンテンツから代表的な単語を tf-idf 法を用いているが，ウェブサイトの形態は様々であり，ブログやウェブ記事などの文章が多いウェブサイトでは有効であるが，その他の場合では正しく内容を抽出することができない．また，ユーザが SNS 上で取り上げたウェブサイトをユーザの趣味趣向を反映した情報とし，ユーザプロファイルの作成する研究[2]では，ユーザプロファイルの作成に，ユーザがブックマークしたウェブサイトのタグを特徴量とし用いている．しかし， 現状すべてのウェブサイトに対してタグが付与されておらず，付与されている場合でも，ブックマーク数が少なく，特徴量として用いるには不十分である場合が存在する．

本研究では、抽出したウェブサイトの意味内容を情報量として扱うことで，ソーシャルブックマークを用いた研究において，ウェブサイトに付与されるタグを正しく推定する方法を検討する．ウェブサイトの意味内容を抽出する手法として，ソーシャルブックマーク上でウェブサイトに付与されたタグ群から表記のゆれを取り除いて正規化したタグをコンテンツの内容を表すものとし、付与されたタグとウェブサイトのコンテンツから抽出した代表的な単語群を教師データとし，分類器を用いてウェブサイトに付与されるタグから意味内容を推定する．

2. タグ分類器

ウェブサイトにタグを付与するために，分類器を用いる．一般的に文書分類では，ナイーブベイズやサポートベクトルマシーンなどが用いられるが，本研究では，ナイーブベイズを用いる．ナーブベイズは，テキストから得られる単語が互いに独立であるという単純な仮定とベイズの定理に基づいた分類手法であり，設計や仮定が単純であるが、学習や推論などにおいて高速で動作し，特に文書分類において高い精度を誇ると知られている．

† 法政大学大学院 情報科学研究科

‡ 法政大学 情報科学部

3. 教師データ作成

教師データに用いるデータセットを国内最大のソーシャルブックマークサービスのはてなブックマーク[3]から取得する．はてなブックマークで用いられるタグは，ブログや Wikipedia などの他のタグ付与を行うシステムと異なり，複数のユーザが複数のタグを自由に付与するため，タグの表記の揺れの問題が存在する．さらに，“あとで読む”や日付など，ウェブサイトの内容を表すタグではないものを多く付与されている．また，ウェブサイトを代表する単語を特徴量とし教師データとする．

3.1 分類器で用いるタグの選択

ユーザがウェブサイトの内容を表すタグを付与する際に，本質的な意味は同じだが表現が若干異なる単語が複数発生する表記揺れと呼ばれる問題がある．しかし，ソーシャルブックマークでは，表記揺れが発生する単語群は，ウェブサイトに同時に付与される可能性が高い．そのため，ある２つのタグが同じウェブサイトに付与される関係性を調べ，閾値以上の単語同士を１つの単語に統合することで，表記揺れの問題に対応する．

タグ a に対するタグ b の関係性は，タグ a を含むウェブサイト数を n(a)，タグ b を含むウェブサイト数を n(b)，そしてタグ a とタグ b を両方とも含むウェブサイト数を n(a∩b)とし，以下の式で求める．

$$P_{ab} = \frac{n(a \cap b)}{n(a)}$$

Folksonomy におけるタグの意味的階層関係の抽出[4]では，$P_{ab} \fallingdotseq 1$ となるとき，タグ a とタグ b の意味的な階層関係にあり，タグ a が タグ b の上位語となると述べられている．本研究では，タグ同士の階層関係は扱わず，同じ意味のタグの統合を目的とし，P_{ab} と P_{ba} がそれぞれ一定の値以上である２つのタグをより利用頻度が高い方のタグに統合する．図 1 のプログラミングタグと programming タグの様に，日本語と英語という表現の大きく違いがあるが，プログラミング関連のウェブサイトに，共に付与される確率が非常に高く，同じ意味の単語であると推定できる．

最後に，ウェブサイトを表す内容として不適切なタグが多数存在する．基本的に，はてなブックマークでは，後に任意のタグを含むウェブサイトを検索しやすくするために，ウェブサイトのジャンルや内容を表すタグを付与することが多い．しかし，興味のあるウェブサイトを見つけたが，そのときは時間がなく，あとで読み返したいウェブサイトに付与する，あとで読む（みる）や，ブックマーク数の閾値を表す threshN(N には 25 や 200 などが入る)や，myhotentry など汎用的で，ウェブサイトの内容を表現していないタグを除去する必要がある．

3.2 特徴量の生成

ウェブサイトのコンテンツは，現在 HTML5，CSS3，JavaScript などの多彩な表現力をもつ技術により様々な形態を持っており，素の HTML テキストのままでは，ノイズデータが多く存在するため、HTML タグや HTML に組み込まれている CSS，Javascript コードを除去する必要がある．本研究では，形態素解析ライブラリである MeCab を用いて名詞，そのなかでも一般名と固有名詞のみを抜き出し、記事の単語群とする．単語を抜き出した後，それぞれの単語の特徴量を tf-idf 法により求め，その値が大きいものの上位 10 の単語とその特徴量を学習に用いる．

4. 実験

4.1 データ

はてなブックマークの記事検索ページから，プログラミング言語の中でもスクリプト言語と呼ばれる，PHP，Perl，Python，Ruby の 4 つを検索ワードとし，ブックマーク数が 100 件以上のものをそれぞれ最大 400 件取得するプログラムを作成し，実行した結果 1365 件収集した．その後，ウェブサイトに付与されているタグを取得することができるはてなブックマーク API を用いて，先ほど取得した全てのウェブサイトに付与されているタグを取得した．10 以下しか付与されていないタグは，ウェブサイトの内容を表すタグとして不適切であるとし，そのタグを除去した．また、はてなブックマークで独自に使われているタグに関しては，タグの判断が困難であり，数がそれほど多くないことから，この処理は自動で行わずに手動で行った．また，タグを統合する条件を $P_{ab}>0.45$ かつ $P_{ba}>0.45$ とした．

4.2 実験方法

すべてのタグにそれぞれ，自身のタグとその他のタグであることを示す’other’の 2 つのタグ分類器を作成する．学習には，収集したすべてのウェブサイトにおいて，自身のタグを含む場合は，自身のダグとウェブサイトの特徴語を，そうでない場合は，’others’タグとウェブサイトの特徴語を分類器に学習させる． 学習が完了した後，分類精度を確認するために，交差検定を用いる．交差検定とは，すべてのテストデータを N 個に分割し，その中の 1 つをテストデータとし，残りの N-1 個を教師データとする評価方法である．

4.3 実験結果

取得した全タグ数は 9973 件で，タグが付与されるウェブサイトの数が 10 より大きいタグが全体の約 10%の 992 タグであった．さらに，統合後のタグ数は 810 となった．タグの統合例を表 1 に示す．

表 1. タグの統合の例

design，デザイン，webdesign，css
paas，cloudcomputing，ホスティング，hosting
aws，ec2，amazon ec2，ec
高速化，チューニング，tuning，パフォーマンス，performance
統計，機械学習，自然言語処理，nlp
book，本，書籍，books
macosx，mac os x，osx

表 2 . タグ付与の結果

タグ	正解率	適合率	再現率	F 値
php	0.659341	0.742424	0.777778	0.759690
ruby	0.673993	0.560000	0.673077	0.611354
python	0.736264	0.488636	0.614286	0.544304
perl	0.772894	0.547368	0.732394	0.626506

表 3. 分類器によって付与されたタグの例

実際に付与されているタグ	分類器が付与したタグ
mac ： 49	web
coda ： 49	web 制作
web 制作 ： 35	ツール
コーディング ： 25	tutorial
dreamweaver ： 21	アプリ
ipad ： 14	wordpress
ツール ： 13	開発環境
アプリ ： 12	iphone

表 2 にタグ付与の結果の中で，代表的なタグについて示す． 正解率は約 70%で，適合率＜再現率という結果となった．適合率＜再現率となるのは，正解のタグを高い確率で付与できているが，不適切なタグの付与がより多くなっている場合である．つまり取りこぼしが多い状況である．つまり，本研究における分類器は，ウェブサイトに関係のないタグを付与している確率が若干高い．

表 3 に Mac で動作する web 制作に特化したテキストエディタに関する記事に付与されているタグの例を示す．web，web 制作，ツール，アプリは正確に分類していることが分かる．また，tutorial や開発環境，iphone など実際には付与されていないが，ウェブサイトの内容を表すタグに関連するタグを付与しており，recall 値の上昇の原因となっていると思われる．しかし，ウェブサイトの内容を表すタグとして適切なタグが付与されていることから，分類器の精度があまり良い結果となったが，ウェブサイトの内容を表すタグを推薦できていると言える．

5. 結論

本研究では，ソーシャルブックマークで付与されるタグからウェブサイトの内容を表しているタグを選択し，それぞれに分類器を作成した．そして，分類器の評価とウェブサイトに付与されているタグとの比較を行った． 付与されたタグは，内容を表すものとして十分であり，このシステムを用いることで，未知のテキストから内容を抽出できることを示した．

参考文献

[1] 阿部 佑樹, 糸川 剛, 北須賀輝明, 有次 正義. Web の閲覧履歴を情報源としたソーシャルブックマークにおけるタグ推薦の提案. DEIM Forum, 2011.
[2] 福島 良典，大澤 幸生. ソーシャルメディアを利用したセレンディピティな情報推薦. 人工知能学会誌, 2012.
[3] はてなブックマーク http://b.hatena.ne.jp/
[4] 杉本 徹，五十嵐 幹, Folksonomy におけるタグの意味的階層関係の抽出，FIT，2008