3E1-1

ソーシャルブックマークサービスを利用したタグ付け自動化システム開発に関する一考察

A Development of an Auto Tagging System on Social Bookmark Services

小野裕作[*1] 當間愛晃[*2] 遠藤聡志[*2]
Yusaku ONO Naruaki TOMA Satoshi ENDO

[*1]琉球大学大学院理工学研究科情報工学専攻
Information Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus

[*2]琉球大学工学部情報工学科
Department of Information Engineering, Faculty of Engineering, University of the Ryukyus

Social Bookmark Service is useful. Traditional bookmarks have problems about difficluty to manage and sharing data on several environment. To use the service, these provlems are dissolved.

But there are still problems such as the problem of tagging. The tagging is a classification method in Social Bookmark Services.

The purpose of this research is the development of an auto tagging system. This system resolve the problems of tagging to tag bookmark instead of users.

1. はじめに

1.1 ソーシャルブックマークサービス

ソーシャルブックマークサービス（Social Bookmark Service, 以下 SBS）とは、ネットワーク上にブックマークを保存し他のユーザーと共有する Web サービスである。日本国内でメジャーな SBS に、はてなブックマーク [1] や del.icio.us[2] がある。SBS を使うことのメリットとして、以下の三つが挙げられる。

- **異なるマシン間で同一のブックマークを利用できる**
  SBS ではブックマークがネットワーク上に保存されている。そのため、Web ブラウザがあればマシンに関わらず同一のブックマークを利用できる。家や職場などで複数のマシンを使用している場合、この特徴は非常に有用である。

- **ブックマークの共有により有用な Web ページを発見できる**
  SBS 上では、各ユーザーのブックマークが公開されている。また、多くのサービスではブックマークされた数も同様に公開されている。ユーザーは公開されたブックマーク集から現在注目を浴びている Web ページや有用な Web ページを発見することができる。

- **タグを利用することができる**
  タグとはコンテンツの内容を表すキーワードである。これを利用することで、ユーザーはブックマークの管理や分類に関して多くのメリットを受けることができる。詳細については 1.2 節で述べる。

1.2 タグ

タグとはコンテンツの内容を表すキーワードである。SBS ではタグを用いてブックマークの分類を行う。タグによる分類の特徴として以下の二つが挙げられる。

- **複数項目への分類**
  タグを利用することにより、ユーザーはブックマークを複数の項目へ分類することができる。これにより、従来と比較してより柔軟な分類を行うことができる。

- **論理演算による検索**
  タグをキーとして、AND 演算などを行うことができる。これにより、ブックマークの検索を容易に行うことができる。

1.3 タグの問題点

前述したように、タグを利用することでユーザーは多くのメリットを享受することができる。しかし、タグによる新たな問題点も発生している [3]。それを以下に示す。

- **表記揺れ**
  表記揺れとは、同様の意味を持つタグが複数存在することを指す（例：ブログ、blog、weblog）。ユーザーが名称の基準を決めていない場合、もしくはスペルミスにより表記揺れが発生する可能性がある。

  表記揺れが存在することにより、一つの事柄についての情報が複数タグに分散してしまう。結果として、ブックマークの検索や再利用が困難になるという問題がある。

- **多義語**
  多くの意味を持つ単語のことを多義語という。タグとして多義語が利用される場合、そうでない場合と比較して検索や再利用が困難になるという問題がある。例えば、「Opera」というタグは歌劇としてのオペラと Web ブラウザの Opera の両方に付与される可能性がある。Opera をキーとして検索を行う場合、これら両方に関する Web ページが候補として上がってしまう。

  ただし、他のタグとの組み合わせや AND 演算などを利用することでこの問題はある程度軽減することが可能と考えられる。

連絡先: 小野裕作，琉球大学理工学研究科情報工学専攻，
yono@eva.ie.u-ryukyu.ac.jp

- **着眼点のばらつき**
  ユーザーによって、どのようにタグ付けを行うのかはそれぞれ異なる。

  例として、液晶 TV に関する Web ページへのタグ付けを考えてみる。最も考えやすいタグとしては、「TV」「液晶」などがある。広義的なタグについては「家電」「電化製品」などが考えられる。より詳細なタグ付けする場合、「メーカー名」「商品名」などが考えられる。それぞれの Web ページに対して、どの程度広く、どの程度詳細にタグ付けをするかというのはそれぞれのユーザーに託される。

  タグの付けられ方によっては、適切な検索が行われず、必要な情報が得られない場合が考えられる。

- **再利用を考慮したタグ付けの困難さ**
  ブックマークにタグ付けをする最大の理由は、再利用性の向上である。そのためには、再利用しやすいようなタグ付けを行う必要がある。しかし、各々のコンテンツをどのようにタグ付けするかについては、様々な可能性が考えられる。人間によってタグ付けが行われる以上、常に一定の基準で適切なタグ付けを行うのは難しい場合がある。

- **ブックマークの移行**
  SBS を利用することで、より容易にブックマークを管理することができる。そのため、従来は Web ブラウザで管理していたブックマークを SBS へ移行したいという場合がある。

  しかし、タグによる分類が行える Web ブラウザはほとんど存在しない。そのためユーザーは新たにタグ付けを行う必要がある。ブックマークの数が増えれば増えるほど、この作業は非常に煩雑になる。

1.4 提案システム

本研究では、1.3 節で述べた問題のうち、表記揺れとブックマークの移行に注目した。これらの問題を解消する手段として、2 章に述べる機械学習を用いたタグ付け自動化システムを提案する。タグ付け自動化システムとは、ブックマークに対してユーザーの代わりに適切なタグを提示するシステムである。

2. 機械学習

本研究では、自動タグ付けを実現する手段として機械学習を利用する。

まず、タグ付けを 2 クラス分類問題（タグを付けるか、付けないか）として定義する。複数タグのタグ付けに対応するために、タグごとに分類器を作成する。

2.1 採用手法

機械学習にはいくつもの手法があるが、本研究では 2 クラス分類問題において評価の高い Support Vector Machine（以下、SVM）を採用した [4]。

SVM ではデータの分類を行うために分離超平面を学習する（図 1）。図 1 中の丸と三角はそれぞれのクラスに属する事例を表している。また、白抜きのものはもう一方のクラスに対して最も距離が近い事例で、Support Vector と呼ばれる。このとき、2 クラス間の距離が最大になるような分離超平面を学習することで、最も汎化能力が高い分離超平面となる。汎化能力とは、未知のデータに対して適切な分類を行うことができる能力

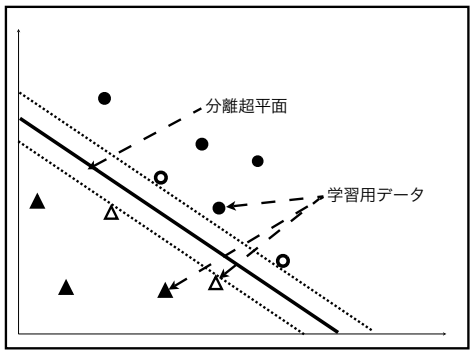

図 1: SVM による 2 次元データ分類

である。また、このようにクラス間の距離を最大にすることをマージン最大化という。

SVM を採用した理由として、以下の二つがある。

- **汎化能力が高い**
  前述したとおり、SVM は汎化能力が高い。そのため、少ないデータ数でも適切な学習を行えると期待できる。

- **高次元の事例に対しての認識精度が高い**
  本研究では、特徴量として Web ページ中の名詞を利用する（2.2 節）。その結果、事例ごとの次元数は数千や数万など、非常に大きい数値になる。SVM はこういった事例に対しての認識精度が高い。

2.2 特徴量

本研究では、Web ページを特徴付けるものは Web ページ中に出現する名詞であると仮定した。そのため、特徴量として Web ページ中の名詞を利用する。重みとして名詞が出現したかどうか（0、1）を利用する。

2.2.1 特徴量抽出

特徴量を抽出する手順を以下に述べる（図 2）。

1. Web ページから名詞を抽出し、リスト化する。名詞にはそれぞれ固有の値が割り振られる。

2. リストを利用して、事例を配列へと変換する。

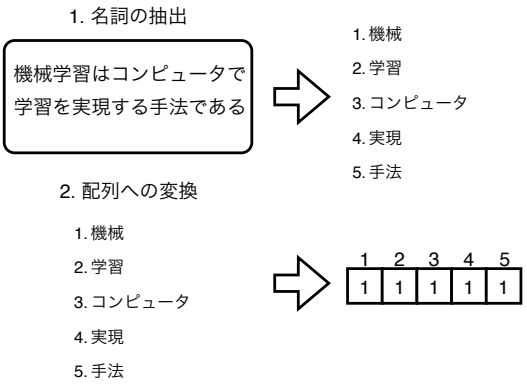

図 2: Web ページからの特徴量抽出

2.2.2 名詞抽出

本研究では、日本語 Web ページと英語 Web ページに対するタグ付け自動化を想定している。そこで、それぞれの言語に対応した形態素解析器が必要となる。日本語形態素解析器としては MeCab[5]、英語形態素解析器としては Tree Tagger[6] をそれぞれ利用した。

日本語については、「数」、「代名詞」、「非自立」、「記号」を除く名詞を利用した。英語については、「NN（単数形普通名詞）」、「NNS（複数形普通名詞）」、「NP（単数形固有名詞）」、「NPS（複数形固有名詞）」の四つの品詞を用いた [7]。

3. タグ付け自動化システム

提案システムの詳細について述べる。

3.1 システムの構成

提案システムは、クライアントとサーバーの二つから構成される（図 3）。

クライアントはユーザーインターフェースとして機能する。より簡単に利用できるように、Firefox の拡張機能として現在開発を進めている。Firefox の拡張機能を選択した理由として、ブラウザと高い親和性がある、Firefox が多くの OS で利用できることなどがある。

サーバーはクライアントからのリクエストを受け取り、自動タグ付けや学習などの処理を行う。

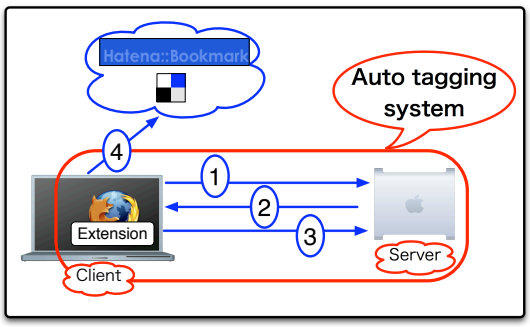

図 3: タグ付け自動化システムの構成

3.2 主な機能

ユーザーが利用できる機能を以下に挙げる。

- **事前学習**
  ユーザーが提案システムを利用し始める際に用いる機能である。これは、以下の 2 パターンを用意している。
  - **ユーザーが新規に SBS を利用し始める場合**
    ユーザーに興味のあるキーワードを入力してもらい、そのキーワードに関連したブックマークをサービス上から抽出する。抽出したブックマークを学習用データとして利用する。del.icio.us を始めとした多くの SBS では、タグによるブックマークの検索が実装されている。これを利用することで、キーワードを基にしたブックマークの抽出を行うことができる。
  - **ユーザーが既に SBS を利用している場合**
    ユーザーが所有しているブックマーク情報を学習用データとして用いる。ただし、そのブックマーク情報が学習用データとして十分な量でない場合が考えられる。この場合には、ユーザーが利用しているタグを基にして、SBS からのブックマークの抽出を行うことでデータ量を補うこととする。

- **自動タグ付け**
  ブックマークを保存する際に、自動的にタグを付ける機能である。システムによって付けられたタグがユーザーの希望に即したものでない場合は、ユーザーによる修正が可能である。修正されたタグは保存され、後述するフィードバック学習に利用される。

- **フィードバック学習**
  自動タグ付け時にユーザーから修正されたデータを学習用データとして学習を行う。これを繰り返すことで、よりユーザーの嗜好や傾向にあったタグ付けを行うことができる。この学習は一定期間ごとに行われる。

4. 実験

提案システムの妥当性を評価するために、SBS のデータを利用した実験を行った。この実験についての詳細を示す。

4.1 目的

以下の二つがこの実験の目的である。

- 本研究において定義した特徴量（2.2 節）によって適切な学習を行えることを確認する。
- タグごとに学習難易度の違いがあるのかを確認する。

4.2 手順

当実験は以下の手順で行われた。

1. **データの取得**
   del.icio.us からタグごとにブックマーク（URL とタグの組）を取得した。なお、元となるタグは同じく del.icio.us からランダムに抽出した。

2. **データの分割**
   取得したデータを学習用データとテスト用データの二つに分割した。このとき、タグごとに正事例（タグを付けるべき事例）と負事例（タグを付けないべき事例）の二つが必要となる。各タグの負事例としては、該当タグ以外のタグの正事例を用いた。

3. **学習**
   学習用データを用いて、タグ付け器の構築を行った。

4. **タグ付け**
   構築したタグ付け器を利用してテスト用データに対するタグ付けを行った。この際の分類精度を評価基準とする。

データ数などの詳細については表 1 に示す。

4.3 結果

実験結果を図 4 に示す。この図は、タグごとのテスト用データに対する分類精度を表している。縦軸は分類精度、横軸はタグの種類を示している。横軸については、分類精度をキーとしたソートを行った。

多くのタグについては良好な結果を得たが、一部のタグについては精度が低いことがわかる。

表 1: 実験環境

タグ総数	84
タグごとの総データ数	400
タグごとの訓練用データ数	200
タグごとのテスト用データ数	200
SVM ライブラリ	LibSVM
日本語形態素解析器	MeCab
英語形態素解析器	Tree Tagger

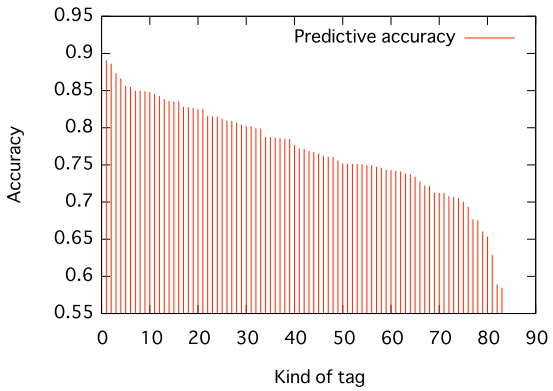

図 4: 実験結果

4.3.1 タグの利用方法に関する調査

実験結果において精度が低かったタグに関して、どのような Web ページに対してタグ付けされているのか、その利用方法を調査した。調査結果を表 2 に示す。

表 2: タグの利用方法

タグ	利用方法
mouse	動物のネズミに関するページ、 コンピュータマウスに関するページなど
type	font type に関するページ、 typing に関するページ、 type に関するページなど
scripting	スクリプト言語に関するページ。 (例 shell script, VBScript, AppleScript, Perl)
rescue	救助活動や遭難などに関するページ、 コンピュータに関するページなど (例 故障した HDD からデータを取り出す)

表 2 から、精度が低いタグは、複数の話題に対して利用される傾向があることがわかる。mouse について言えば、多義語の問題とも関連性があるといえる。

4.4 考察

実験結果から、名詞は特徴量として有効であることがわかった。しかしタグによって判別精度にばらつきがあることから、更なる改良の余地があると考えられる。

また、複数の話題が一つのタグで扱われている場合には学習が難しくなる。そのため、混在する複数の話題を認識し、分割する手法が必要であると考える。

今回、訓練用データとテスト用データをそれぞれ 200 ずつ用意して実験を行った。しかし、タグによっては必要データ数が集められず、実験に利用できない場合もあった。このようにデータ数が集められないタグの特徴の一つとして、日本語によるタグが挙げられる。del.icio.us 上では日本語タグが少ない傾向があり、これが原因だと思われる。そこで、日本語タグと同義語の英語タグを利用することで、データの取得がより容易になると考えられる。

5. まとめ

本研究では、SBS の利用を支援することを目的として、タグ付け自動化システムの開発を行った。自動的なタグ付けを実現するために、判別手法の検討とタグ付け器の構築を行った。また、ユーザーが容易にシステムを利用できるように、ユーザーインターフェースとして Firefox の拡張機能を開発した。

提案した特徴量の妥当性を検証するために評価実験を行った。その結果、特徴量や前処理に関して改良の必要があることが判明した。

今後の課題として、サーバープログラムを含めたシステムの公開がある。そのため、現在ドキュメントの整備などを行っている。また、Firefox 以外のブラウザでも利用できるようなクライアントについても開発する必要がある。

今後の展望として、ブックマーク管理を容易にするインターフェースと、ブックマーク管理支援システムの開発を考えている。また、従来のカテゴリやタグに置き換わる新たなブックマーク分類方法についても検討や提案を行っていく。

参考文献

[1] はてなブックマーク - ソーシャルブックマーク
http://b.hatena.ne.jp/

[2] del.icio.us
http://del.icio.us

[3] Scott A. Golder and Bernardo A. Huberman. (2005).
“The Structure of Colaborative Tagging Systems”.
Information Dynamics Lab, HP Labs.

[4] Corinna Cortes and Vladimir Vapnik. (1995).
“Support-vector networks”.
Machine Learning, 20(3):273-297

[5] MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.sourceforge.net/

[6] TreeTagger
http://www.ims.uni-stuttgart.de/projekte/corplex
/TreeTagger/

[7] Mitchell P. Marcus and Beatrice Santorini and Mary Ann Marcinkiewicz. (1994).
“Building a Large Annotated Corpus of English: The Penn Treebank”.
Computational Linguistics, 19:313-330