# Folksonomy の機械化：Blog 記事へのマルチタグ付与

大倉 務　　清田 陽司　　中川 裕志
東京大学 理学部 情報科学科　　東京大学 情報基盤センター　　東京大学 情報基盤センター
ohkura@is.s.u-tokyo.ac.jp　　kiyota@r.dl.itc.u-tokyo.ac.jp　　nakagawa@dl.itc.u-tokyo.ac.jp

## 1　はじめに

### 1.1　Folksonomy とは

最近、Folksonomy[1] と呼ばれる新しい整理法が注目を集めている。Folksonomy とは、個々のユーザーが自身の視点でタグ（自由に選んだキーワード）をアイテムに付与することで、システム全体としてみれば多様なタグが個々のアイテムに付与されるというもので、ユーザーの語彙や視点が分類に反映されるという特徴をもつ（図 1）。Flickr[2] や del.icio.us[3] といったサービスで採用され、これらのサービスの人気の主な要因といわれている。

まず、Folksonomy を用いた典型的なサービスについて述べる。ユーザーは何らかのアイテム（写真や URL など）をサービスに登録し、他のユーザーと共有する。そして、ユーザーは自分のアイテムだけでなく他人のアイテムにもタグを付け、自分の好きな形で整理することができる。各々が自分の整理のために付けたタグは共有されるので、全体としてみると個々のアイテムにはたくさん視点を反映したタグが付く。

Folksonomy の特筆すべき特徴は、ユーザーによる分類という点である。現在、専門家による分類や著作者による分類が一般的だが、Folksonomy では一般ユーザーの多様な視点を反映した分類を行える。一人一人の個人に適切な分類能力があるとは考えにくいが、個人が集まり、たくさんの視点を組み合わせることで、分類の精度は高まる。タグはユーザーの視点でユーザーの語彙から選ばれたものであるため、その分類基準や表現は一般ユーザーの理解しやすいものとなる。

また、語彙やその指し示す概念は時間が経つにつれて変化するが、従来の固定されたカテゴリ体系による分類では語彙やその概念の変化への対応は遅れがちである。Folksonomy では、ユーザーが常に最新の語彙

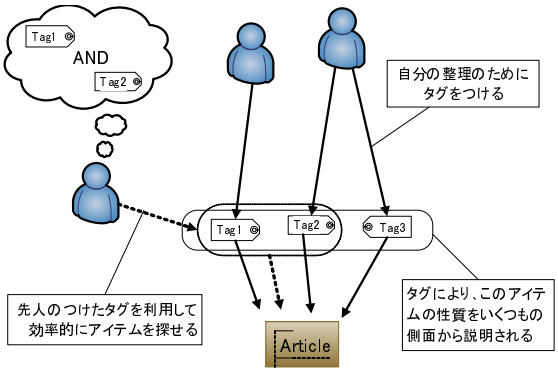



図 1: Folksonomy

や概念に基づいた新しいタグをつけていくので、これらの変化に俊敏に対応できる。

このように、Folksonomy は

1. タグはユーザーの語彙から選ばれる
2. 分類には多数の個人の視点が反映される
3. 語彙やその指し示す概念の時間的変化に適応する

という特徴を持つ。

結果として、Folksonomy では閲覧による受動的な情報検索を実現できる。情報検索には主に下記の 2 種類の方法がある。

- クエリを用い、その語を含む文書を探す（能動的）
- 閲覧しながら、興味のある物を探す（受動的）

クエリによる方法は現在のインターネットでは一般的な方法である。インターネットの創成期には閲覧による方法[4] も一般的であったが、現在では衰退している。しかし、この 2 種は利用できる場面が異なり、ユーザーは両者揃った時に情報を最も有効活用できる。

Folksonomy では、タグは個々のユーザーが整理するために付与するものであるので、通常、分類対象を記述する説明的な単語が選ばれる。このため、タグを利用した閲覧システムは、興味のあるものを探すのに有用である。

[1] "folk" と "taxonomy" の複合語
[2] http://www.flickr.com/
[3] http://del.icio.us/
[4] Yahoo Directory などのディレクトリ型検索エンジン

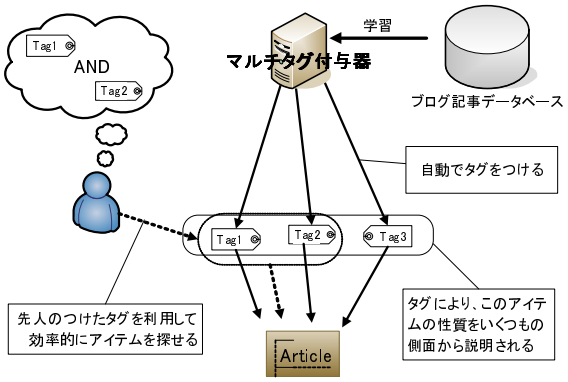



図 2: 機械化された Folksonomy

### 1.2 Folksonomy の機械化

Folksonomy のスケーラビリティは専門家によるそれよりもはるかに高いが、人手に依る以上、膨大な情報にはやはり対応できない。膨大な情報に対応するためには、Folksonomy を機械化する必要がある。この際、上記の Folksonomy の 3 つの特徴を失わないようにしなくてはならない。

上記の特徴 1 を満たすため、タグ名はブログのカテゴリ名を元にした。多くのブログサービスやブログツールでは、ユーザー自身がカテゴリ体系を決め、その体系で自分の記事を分類できる。このため、ブログで用いられているカテゴリ名はユーザーの語彙に基づいており、またユーザーが分類のためにつけた名前であるので、タグ名の候補となり得るものが多いと考えられる。

また、特徴 2 を満たすため、マルチタグを付与することとした。さらに、個々のタグは複数の著者のブログから得た記事群を用いて学習させた分類器を用いて付与することとした。

さらに、特徴 3 を満たすため、タグ付与器の学習は繰り返し（毎日 1 回など）行うこととした。対象とするカテゴリ数が多いため、用いる機械学習アルゴリズムは適用だけでなく学習も高速であることが必要となる。

また、今回はタグを付与する対象としてもブログ記事を利用した（図 2）。これは、日本国内に限っても、去年 1 年間に新しく投稿されたブログ記事は 1 億件に達するとみられ、人手による Folksonomy では網羅できない規模であり、Folksonomy の自動化を試みる対象として適切だと考えたためである。

図 3 にブログ記事を対象とした、機械的な Folksonomy システムの概観を示した。ブログ記事を対象としたマルチタグ付与システムであり、1 つ 1 つの記事に複数のタグを付与する。付与するタグや、そのタグの

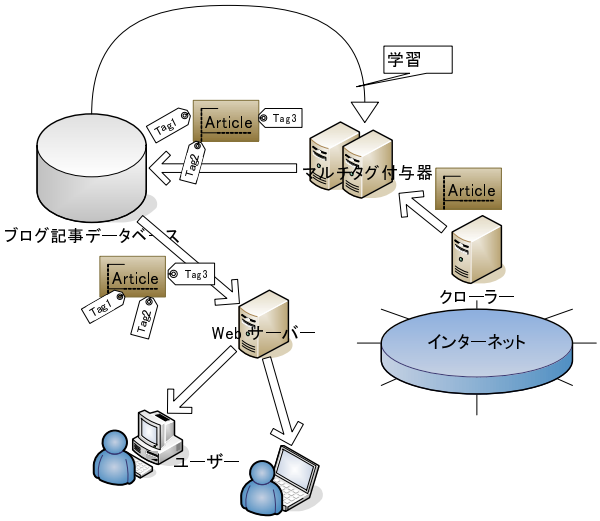



図 3: システムの概観

概念は収集した多数のブログ記事から抽出する。特定のタグを付与するかどうかを判断するタグ付与器について 2 節で、その後の 3 節ではタグとして適切な単語を選ぶ処理について述べる。

## 2 タグ付与器

特定のタグを付与するかどうかを判断するタグ付与器を、すべてのタグ候補について構成することで、マルチタグ付与器を構築する。タグ付与器は、2 クラス（タグを付与する・しない）の文書分類器である。できるだけ多くのデータを用いて学習させることで精度の向上を狙い、またユーザーの語彙や概念の変化に追従するため、タグ付与器は繰り返し学習させる必要がある。

### 2.1 文書分類手法の選択

まず、タグ付与器に用いる文書分類手法について検討する。タグ付与器に適した文書分類手法とは高速な分類が可能で、ノイズ耐性があり、調整が必要なパラメータが少なく、現実的なトレーニング時間で、分類時のメモリ消費量が少なく、高い分類精度をもつものである。広く用いられているいくつかの文書分類手法 (k-NearestNeighbor, Naive Bayes, AdaBoost, SVM) についてこれらの観点から総合的に比較し、SVM が最適であると判断した。

### 2.2 学習

素性には単語ベクトルを用いた（分かち書きには MeCab[5]を用いた）。さらに、周期的な事象が多いことを考慮し、日付と時間の情報も素性に加えた。各素

[5]http://mecab.sourceforge.jp/

表 1: カテゴリ名とそのタグとしての適切さ

| 良いタグ | 説明力不足 | 一般性欠如 |
|---|---|---|
| “音楽” | “その他” | “ブログルメ” |
| ‘映画” | “日常” | “清掃業者比較” |
| ‘和菓子” | “つぶやき” | “翼は空へ” |
| “FFXI” | “友達” | “お題目次” |

性の重み付けには tf-idf を用いた（日時情報は単語 2 語分の重みとした）。日時情報を素性として用いたところ、用いない場合に比べて 1%強の分類精度向上がみられた。精度の向上は、特にニュースや音楽、旅行関連で顕著であった。

学習データとしては、タグをカテゴリ名として持つ記事の集合を Positive Example, ランダムに選んだ記事を Negative Example として用いた。

### 2.3 SVM 出力の確率値への変換

マルチタグ付与において、False Positive と True Negative のエラーのコストには差がある。特定の記事に対して付与されるべきタグは、全体から見て極めて少数である。このため、誤ったタグを付与してしまう問題は、付与し損ねるミスよりも重大である。そこで、Platt[1] の手法を用いて SVM の出力を確率値に変換し、90%以上の確率で付与すべきであると判断されたもののみタグを付与することとした。

## 3 タグ候補集合の選択

マルチタグ付与を行うにあたっては、付与の候補となるタグの集合を決める必要がある。タグ候補集合をブログのカテゴリ名から一般的かつ説明的なものを選び出すことで構成する方法を述べる。

タグはユーザーの語彙に基づいたものであるべきである。ここで、ブログのカテゴリ名に注目した。ほとんどのブログでは、そのカテゴリ体系は著者自身によって構築されており、カテゴリ名は一般ユーザーの語彙に基づいてつけられてるといってよい。

さらにタグは説明的でかつ一定程度の人に共有されるものでなければならない。タグとして適切なカテゴリ名とそうでないカテゴリ名の例を表 1 に挙げた。

まず、カテゴリ名が一般的な単語であるかどうかは、そのカテゴリ名が利用されているブログの数で判断した。複数の人が同じ名前を用いていれば問題ないと考えられるが、一人で複数のブログを運営している例もある。これを考慮し、5 個以上のブログで共通して用いられている単語は一般的であると判断した。

もしカテゴリ名に説明力があればこのカテゴリ名をタグとして用いた場合のタグ付与器の精度は自然と高精度に、なければ精度は低くなるはずである（具体的なデータは次節の実験で示す）

タグかとして利用するかどうか判断は、必要最小限のデータが集まった時点でタグと認識すると共に、精度が低いものを間違ってタグとして利用してしまうことは避けたい。ここで、どのカテゴリ名にも分類器の精度を正確に測定するのに十分なブログ記事があるわけではないため、少ないサンプルから真の分類精度を推定する必要がある。サンプルの記事 1 つを正しく分類するという事象を確率事象だとみなせば、この事象の生起確率は分類精度に等しい。このため、分類に成功した例の数は二項分布となる。二項分布がサンプル数が少なくなければ正規分布で近似できることを用いれば、99.5%の信頼度で

$$\frac{c}{n} - 2.58 \times \sqrt{\frac{c}{n}(1.0 - \frac{c}{n})/n} \leq p \qquad (1)$$

という式が満たされる。そこで式 1 が閾値を超えるならばこのカテゴリ名はタグ名として十分な説明力があると判断した。

## 4 実験

### 4.1 データ

Web から収集した 2005 年の 4/13〜12/1 の間に投稿された日本語ブログ記事を実験データとして用いた。RSS データで dc:subject 属性で記述されている物をカテゴリ名とみなし、また実験時間を短縮するため、“未分類”・“日記”・“weblog”・“ニュース”・空白の 5 カテゴリを除いた 2,460,374 記事を対象とした。

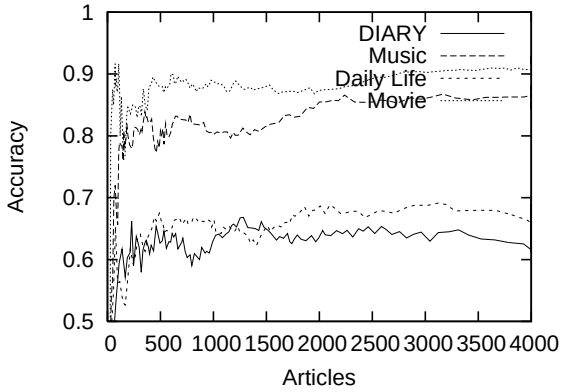
### 4.2 カテゴリによる分類精度の違い

いくつかのカテゴリ名について学習に用いる記事数と分類精度の関係を調べた（図 4）。ここでは 2.3 節で述べたバイアスのかかっていない SVM 出力で記事を分類した（SVM の実装として TinySVM[6]を利用した）。ここで、精度は学習記事数の $\frac{1}{3}$ の記事（学習とは異なる記事）で測定した。この図より、説明力のある語 (日記, 日常生活) の精度は高い水準で、説明力のない語 (音楽, 映画) は低い水準に収束することが分かる。

### 4.3 タグ選択

一定数以上の記事をもつカテゴリ名に対し前述のアルゴリズムを適用し、人手で作成したタグ候補集合と

[6]http://chasen.org/~taku/software/TinySVM/




DIARY:“日記” Music:“音楽” Daily Life:“日常生活” Movie:“映画”

図 4: カテゴリごとの分類精度の違い

表 2: 候補タグ集合の選定の評価

| 記事数 | tp | tn | fp | fn | 精度 | 適合率 |
|---|---|---|---|---|---|---|
| >= 100 | 91 | 3 | 113 | 135 | 66.0% | 96.8% |
| >= 200 | 76 | 3 | 67 | 107 | 72.3% | 96.2% |
| >= 400 | 41 | 2 | 24 | 62 | 79.8% | 95.3% |
| >= 800 | 27 | 1 | 5 | 34 | 91.0% | 96.4% |
| >= 1600 | 11 | 1 | 0 | 16 | 96.4% | 91.6% |

tp : システムも人間もタグとして適切と判断したカテゴリ数
tn : システムは適切、人間は不適と判断したカテゴリ数
fp : システムは不適、人間は適切と判断したカテゴリ数
fn : システムも人間もタグとして不適と判断したカテゴリ数

比較する実験を行った。図 4(と他のカテゴリでの同様の実験) の結果を考慮し、閾値は 75%とした。表 2 の結果より、前述の要件である高い適合率を達成できていることが分かる。

800 件以上記事があるカテゴリについての唯一の間違い（tn）は「お知らせ」カテゴリである。このカテゴリ名は一般的で説明力が高いとはいえないが、ブログカテゴリでの「お知らせ」はブログサイトの方針説明等に限られており、これが誤った判断の原因となっていると考えられる。

### 4.4 タグ付与

4.3 節で選んだタグ (800 記事以上) を、実際に付与した際の再現率と適合率を測定した。再現率は、ブログ記事の著者が分類したカテゴリと同一のものを付与できるかで測定した。また、適合率は抽出結果を人手で評価した。共に評価は 1 カテゴリあたり 20 記事前後で行った。

図 5 の結果より、多くのタグにおいて高い適合率を実現しているが、再現率は低いものもあることが分かる。また、図中の適合率が 50%を下回った唯一のタグは「野球」で、サッカー関連の記事が多数含まれていた。これは、似た単語が用語が多いことが原因だと考えられる。

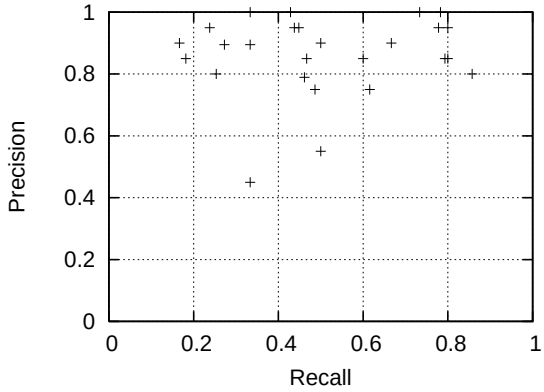



図 5: タグ付与の適合率 (Precison) と再現率 (Recall)

## 5 議論

本手法はユーザーによるタグ付けを置き換えるものであるので、既存の Folksonomy サービスの有用な表示法と組み合わせることが可能である。また、本手法は 1 つ 1 つのタグ付与器が完全に独立しているため、並列化が容易である（本論文中の実験の多くも 4 プロセッサで並列して行った)。

本手法の問題点としては、似た用語が使われがちな概念間の識別が難しいという問題がある。この問題の解決は今後の課題である。

現在、SVM の問題点である長い学習時間に対応するため、3 節で述べた精度予測法を応用して不要と思われる学習をスキップする試みを行っている。今後はスカッシング等の他の学習高速化手法と組み合わせることを考えている。

## 6 まとめ

本論文では，受動的な情報検索の新しい形として Folksonomy の機械化を提案した。さらに、その実現のためのシステムの構成、タグ付与器の構成、タグ候補の選び方について述べ、それら有効性を実験的に検証した。

## 参考文献

[1] John C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.