3G8-3

ブログ記事への自動マルチタグ付与

Multi-Autotagging for Blog Entries

藤村　滋[*1]　藤村　考[*1]　奥田　英範[*1]
Shigeru Fujimura　Ko Fujimura　Hidenori Okuda

[*1]NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories

To utilize blog tags in classification or navigation, there are several considerable issues as follows: 40% of entries aren't tagged, there are many orthographic or synonymous tag variations, and not all tags are informative. We propose a method of multi-autotagging, based on k-NN, which is a contents-based classification method. It can also merge tags with the same meaning and identify informative tags. Experiments show the effectiveness of our methods.

1. はじめに

個々のユーザーによって設定された「タグ」という，コンテンツに関するメタデータを集約することで，タグを通して皆でコンテンツを分類する Folksonomy [1] と呼ばれる分類手法に注目が集まっている．Folksonomy は，Folks(人々) と Taxonomy(分類学) を組み合わせた造語であり,「皆で分類する」ことで，従来よりもより優れた分類になるという集合知 (wisdom of crowds)[2] の一形態と考えられる．Folksonomy では，個人が自由にタグを設定できることから，既定の分類体系と比べ万人の価値観の尺度に近くなることや，一つのコンテンツにマルチタグを許容することで，関連タグへのナビゲーションが可能になることが利点といわれている．Folksonomy を利用しているサービスの一例としては，写真共有の Flickr[*1]や，ソーシャルブックマークの del.icio.us[*2]などがあげられる．

一方で，十分に普及しているといっても過言でないブログにもタグ[*3]の機能はあるが，Folksonomy の魅力を引き出す様な決定的なサービスは生まれていない．現状では，Technorati 社のタグ検索[*4]をはじめとした，タグによる検索サービスに利用されている程度である．

そこで，ブロガーが設定したタグについて筆者らが分析した結果 [3]，まず，タグ未設定の記事が約 40% 程度あり，タグ未設定時に設定されるタグを考慮するとほぼ半数近くの記事はブロガーによってタグが設定されないことが分かった．次に，設定されているタグは，表記ゆれや類義語によって，同一概念のタグが乱立するという問題がある．さらにタグの有用性という点では,「日記」や「Weblog」といった内容を表さないタグが多く存在することが分かった．最後に，ブログのタグの多くはシングルタグであるということも Folksonomy という観点においては，不十分な点と考える．

そこで，ブログにおいてもタグを用いた魅力的なナビゲーションを可能にするためには，ブログ記事に対して適切なタグセットを作成し，その中から自動的にマルチタグを付与する技術が必要であると考えている．本稿では Folksonomy の概念に近い，事例ベースの k-NN 法によって，自動マルチタグ付与を行った．タグ付けに先立って，事前処理として表記ゆれ等により同一の概念を表すタグを自動的に統合する処理および,「日記」等の不適切なタグを識別する処理を加えることで精度の向上を図っている．また，簡単な評価実験の結果についても報告する．

以下，本稿の構成を示す．2 章では，自動マルチタグ付けの手法について述べる．3 章では，簡単な評価実験を行い，その結果と考察について述べる．4 章では，関連研究について述べる．最後に 5 章では，本稿のまとめと今後の課題についておわりにとして記す．

連絡先: 藤村　滋，NTT サイバーソリューション研究所，神奈川県横須賀市光の丘 1-1，fujimura.shigeru@lab.ntt.co.jp

*1　http://www.flickr.com/
*2　http://del.icio.us/
*3　カテゴリやジャンルとも呼ばれることもある．
*4　http://www.technorati.jp/tags/　2005 年 12 月 20 日に日本でも開始

2. ブログ記事への自動タグづけ

本章では，まず自動マルチタギングの手法について述べる．続いて，自動マルチタギングを行う前の 2 つの事前処理について説明を行う．ひとつは，表記ゆれや類義語によって内容的に類似したタグが乱立しているといった問題を解決するため，類似したタグを統合する処理についてである．もうひとつは,「Weblog」や「その他」などテキストの内容を表さないタグを判定する処理についてである．

2.1 k-NN を基にした自動マルチタギング

ブログ記事に自動でマルチタグを付与する方法としては，大きく分けて，記事中のキーワードをタグとして付与するトピック抽出による手法と既にタグが設定されている記事を学習データとして利用することで，記事に付与されるべきタグを推測する手法が考えられる．

本稿では，学習による手法であれば記事中に登場しない語句をタグとして付与できるという点をメリットと考え，学習による手法を選択することとした．学習に基づく手法としては，文書分類手法としてナイーブベイズ法，SVM 法，決定木学習による手法，また，本稿で用いた k-NN(k-Nearest Neighbor) 法など多数の手法がある．

ここで，本稿では k-NN 法を採用することとした．k-NN 法はそのアイデア自体は非常にシンプルなものであり，未知文書が入力されると，距離的に近い k 個の文書につけられたラベルを基に分類先を決定する事例ベースの手法である．ブログにおいては，距離的に近い k 個の文書とは類似度の高い k 個の記事，ラベルはタグと置き換えられる．

k-NN 法の特徴としては，Folksonomy との親和性が高い点があげられる．入力記事に対して集められた k 個のタグつき記事においては，投稿したブロガーがその記事に対して適切な

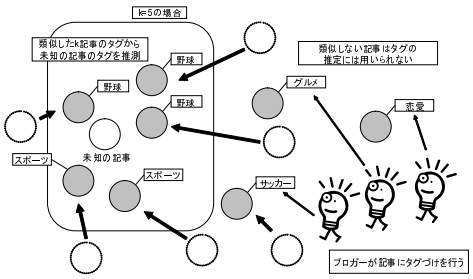

図 1: k-NN と Folksonomy

タグを付与していると考えられる．このタグが未知の記事へのタグづけに用いられることから，換言すると「類似記事」を通して，複数のブロガーが未知の記事に対してタグづけを行っていると捉えることができ，まさに擬似的な Folksonomy と考えることができる．上記について図で表したものが図 1 である．k-NN のこのような特徴から，Folksonomy のメリットである万人の尺度を基にした分類体系や複数の観点に基づくマルチタグ付与をブログの自動タグづけにも持ち込むことができるのではないかと考えている．

また， k-NN 法には既存の関連文書検索の機能を持った全文検索ソフトウェアとの親和性が高いというメリットもある．つまり，全文検索ソフトウェアのデータベースに備わる高度な文書管理機能を利用することによって，blogosphere の時間的な変化に対し対応が容易になる．換言すると，検索システムにおけるデータベースに新たな記事をインクリメンタルに挿入し，古くなった記事はデータベースからドロップするだけで時間的な変遷に対応が可能である．一方で，他の学習手法では窓関数を用いる等によって，一定期間ごとに判別器を再構築する必要がある．

さらに，ブログ記事のタグにおいては，適切なタグが設定されていない，もしくは設定されたタグが誤っているといった問題も考えられる．ここで，k-NN 法では k 件で集計を行うことによって，ある程度ノイズの影響を打ち消すことができるといった点もメリットと考えることができる．

ブログ記事に対するタグ付けにおいては，タグ毎の記事数に極端な偏りがある [3]．このような状況における k-NN 法の有力性については，Yang ら [4] の研究によって示されている．この研究では，分類先クラス毎の教師データ数に極端な偏りのある新聞記事のコーパス (Reuters-21578) を用いて，複数の分類手法を再評価している．結果として，最も性能が高い手法は SVM であったが，最適化された k-NN であれば SVM とほぼ遜色のない結果が得られることが示されている．したがって，ブログ記事においても k-NN は有力な選択肢であると考えている．

また，k-NN によりマルチなタグづけを行うためには閾値の設定が重要になる．そこで，今回は集計によって得られた各タグの記事数と，類似記事検索データベース内でのタグの分布を基にランダムに記事を取り出した場合に得られる記事数の推定値との比較により，有意差検定を行うことでそのタグを出力することとしている．つまり，k 個の類似記事の各記事においてタグ A が設定されているかどうかは近似的に独立であると考えられ，2 項分布で表すことが可能である．中心極限定理によって，データベース中でタグ A が設定された記事数を N_A ，データベース中の全文書数を N とすると，記事にタグ A が設定されている確率は $p_A(= N_A/N)$ であり，以下の不等式が成立する場合，k 記事中でのタグ A の出現は偶然と考えにくく，入力記事にタグ A を付与する．

$$Z < \frac{S_n - np_A}{\sqrt{np_A(1-p_A)}} \tag{1}$$

ここで各記号について，Z は t 分布に基づく検定統計量，n は k-NN における k，S_n は n 文書中でタグ A が設定された記事の実測値である．

2.2 クラスタリングによるタグの統合

表記ゆれや類義語の問題を解決するため，類似したタグをクラスタリングによって統合することを試みた．

ここで，タグ間の類似度として筆者らが文献 [3] で報告したものを用いている．筆者らは，タグの特徴ベクトルを作成する際の語の重み付け法として，そのタグと関連性の高い語が大きな重みを持つように，目的のタグが設定された文書群中での文書頻度の実測値とポアソン分布に基づく推定値との差を利用している．また，この特徴ベクトル同士のコサイン類似度によってタグ間の類似度を求めている．クラスタリングの手法としては階層的クラスタリングの最短距離法を用いた．最短距離法を用いた理由としては，類似度の再計算の必要がないこと，および閾値の設定が容易なことがあげられる．

自動タグ付けにおいて，類似タグを統合することのメリットは次の 2 点である．ひとつは，設定されている記事数が少ないため，付与することが困難なタグの中にも，類似したタグとまとめることによって，統計的に有力なタグとなりうることがあげられる．また，タグの乱立を防止することで，タグを用いたナビゲーションの際にユーザの利便性が向上する点である．

2.3 タグの有用性判定

「Weblog」や「日常」といったタグは，記事の内容を表しているタグとは考えにくい．したがって，タグを利用して話題を掴むといった用途においては，内容を表さないタグを取り除くことが望ましい．また，このようなタグを自動的に付与しないことも重要である．以降では，内容を表さないタグのことを「あいまいなタグ」と呼ぶこととする．

そこで，前節で述べたタグの特徴ベクトル自体に着目し，あいまいなタグを判定するために次の仮説をたてた．

> **仮説** 特徴的なタグは，特徴的な語すなわち特徴ベクトル内で大きな値を持つ語を複数持つ．したがって，ベクトル長が大きくなるため，タグの特徴ベクトル長によってあいまいなタグが判定できる．

上記の仮説に基づいて，タグの特徴ベクトルのベクトル長を算出した．ここで，実際には文献 [3] での特徴ベクトルに対して文書数での正規化を行った上でベクトル長を算出している．

3. 実験

本章では，まず今回の実験に用いたブログ記事について述べる．次に，タグの統合およびタグの有用性判定の結果について述べ，続いて自動マルチタグ付与について簡単な実験を行ったのでその結果について報告する．

3.1 実験データ

今回の実験に用いたブログ記事を表 1 に示す．ここで，計算量の問題からタグ付けの対象としては設定人数上位 5,000 タグとした．5,000 位のタグで，設定人数が約 10 人程度である．今回実験に用いたブログ記事においても，現状ブログのタグがいかにバラバラであるかが分かる．

表 1: 実験に用いたブログ記事

記事数	4,769,657
期間	2006/2/26〜4/1(2006 年 3 月分)
対象	日本語ブログ
ブロガー数	399,414
タグの種類	263,071

```
<cluster no="33">
    <node no="1">run</node>
    <node no="2">ジョギング</node>
    <node no="3">マラソン</node>
    <node no="4">ランニング</node>
</cluster>
<cluster no="26">
    <node no="1">ソーイング</node>
    <node no="2">手作り</node>
    <node no="3">手芸</node>
    <node no="4">パッチワーク</node>
    <node no="5">HANDMADE</node>
    <node no="6">Handmade</node>
    <node no="7">handmade</node>
    <node no="8">作る</node>
    <node no="9">ハンドメイド</node>
    <node no="10">てづくり</node>
    <node no="11">hand made</node>
</cluster>
<cluster no="100">
    <node no="1">一口馬主</node>
    <node no="2">愛馬</node>
</cluster>
<cluster no="145">
    <node no="1">武寺</node>
    <node no="2">音ゲー</node>
    <node no="3">IIDX</node>
</cluster>
```

図 2: クラスタの例

3.2 タグの統合・有用性判定

タグの統合におけるクラスタリングの際には，ヒューリスティックに類似度 0.65 を閾値に設定した．この値は，目視で誤った統合が起こりにくい類似度の下限を確認し設定している．このとき, 閾値を超えた類似度を持っているタグの組み合わせは 2,399 組であった．ここで，タグの組み合わせは最大 ${}_{5000}C_2$ 組考えられる．また，クラスタリングの結果，実際に得られたクラスタ数は 203 であった．幾つかのクラスタの例を，図 2 に示す．図から，例えば「一口馬主」と「愛馬」のように，表記上は全く異なっているにもかかわらず同じような話題を持っていると考えられるタグが，一つのクラスターを形成していることが分かる．

次に，タグの有用性判定について，ベクトル長の上位・下位のタグについて示したのが図 3 である．図から，ベクトル長が短いものは確かにあいまいなタグから構成されており，また，ベクトル長が長いものは固有の話題を有していると考えられるタグが多い．

ベクトル長によるあいまいなタグの判定を評価するために，設定人数上位 500 タグについて人手であいまいなタグかどうかを判定し正解データを作成した．ベクトル長について閾値を設定し，その閾値以上のタグは全てあいまいではないタグ，閾値未満のものはすべてあいまいなタグとする非常に単純な判定方法で閾値を変えた場合の precision，recall，F 値は図 4 のようになった．

図より，閾値を 0.35 に設定した場合に F 値が最大となり，その際の F 値は 0.87 と非常に良好な結果が得られた．

3.3 自動マルチタグ付与の評価実験

本稿では，タグ統合およびタグの有用性判定という事前処理を行ったうえで，自動マルチタギングシステムを作成した．有用性判定においては，F 値最大となったベクトル長 0.35 より

ベクトル長上位 ()内 はベクトル長

F1(1.78), 競馬 予想(1.76), 高校野球(1.70), ワイン(1.68),
阪神タイガース(1.66), 予想(1.64), ラーメン(1.52), リードメール(1.50),
三国志大戦(1.49), 温泉(1.47), baton(1.45), 本・雑誌(1.43),

ベクトル長下位 ()内 はベクトル長

携帯から(0.025), 日記・エッセイ・コラム(0.031), モブログ(0.042),
moblog(0.051), 携帯より(0.054), 日々(0.057), いろいろ(0.065),
にっき(0.073), ひとりごと(0.079), DIARY(0.101), 独り言(0.108),

図 3: ベクトル長の上位・下位のタグ

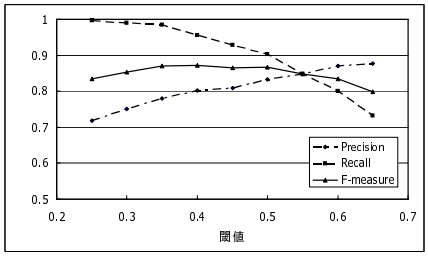

図 4: P-R，および F 値のグラフ

も短いものはタグ推定結果として出力を行わないこととした．クラスタリングについては，クラスタ中の設定人数が少ないタグを設定人数が多いタグに統合することで，いわゆるマイナーな話題においてもシステムによってタグが付与されるように考慮している．ここで，関連文書検索には Hyper Estraier[5] を実装が簡単なため利用している．

関連文書検索用データベースには本稿での実験用データの 2/5 である 2/26〜3/11 までの 2 週間分の記事を利用している．データ量を減らしているのは，計算量の問題のためである．ただし，タグ未設定記事や設定人数 3 人以下のタグがつけられた記事，および「日記」,「Weblog」,「(スペース)」のタグがつけられた記事はタグづけ精度の悪化を招くものと考え，データベースへの登録を行わなかった．結果として，約 48 万件の記事をデータベースに登録している．また，k-NN における k については $k = 100$ としている．

本稿での評価実験としては，タグづけ結果を人手により判定することで簡単な実験を行う．実験データとして，3/12〜3/18 までのタグ付けに利用していないブログ記事のうち，タグ “野球”，タグ “ボクシング”，ランダムに収集したものをそれぞれ 100 記事ずつ集め，データセットとしている．ここで，野球はメジャーな話題，ボクシングはややマイナーな話題を想定している．

推定されたタグのスコア上位 1 件，3 件における精度，およびシステムによって推定された平均タグ数について表 2 に示す．ここで，スコアは式 (1) を用いて以下のように表される．

$$Score(tag) = S_n - np_A - Z\sqrt{np_A(1-p_A)}(> 0) \qquad (2)$$

ただし，タグ推定の結果，記事が短い等の理由により，システム推定タグが一つもない場合は評価から除いている．また，3 件における精度とは，上位 3 つの推定タグ全て正解の場合は 100% ，2 つ正解の場合は 66% ，と各記事ごとに精度を求めた上での全体の平均値であり，システムにとってはやや厳しい評価基準である．[*5]

*5 ただし，推定タグが 2 つ以下の場合には，2 つ正解していれば 100% のように精度を求める．

表 2: タグ推定の精度および平均推定タグ数

タグ	p@1	p@3	avg num
“野球”	0.83	0.51	4.53
“ボクシング”	0.58	0.52	6.72
random	0.70	0.47	3.78

表から，確かにマルチタグを出力できていることが分かる．また，複数の軸を持ったブログ記事はテストデータ中にそれほど多くなかったため，結果としてスコア第 3 タグまでの精度にそれほどの差は見られなかったが，第 1 タグの精度においては，話題がメジャーなほど精度が高い傾向が得られた．

3.4 考察

本手法で出力されるタグに着目すると，特に第 1 タグでは「スポーツ」や「グルメ」といった，非常に汎用的なタグが多く出力されていた．これは，k-NN という手法の特性上，文書データベース中のタグの分布に非常に影響を受けやすく，元々多く含まれているタグが出力されやすいためである．本手法においては，汎化されたタグも重要であるが，幅広い話題に対応することを重要な要素として考えている．したがって，閾値の設定法で有意差検定を行うことで特化されたタグの出力を試みているが，現状では十分に機能しているとは言い難い．一方で，SVM 等のように 2 値判別器を複数組み合わせる手法においては，各判別器は独立であるため，汎化されたタグが出力されやすいといった影響はない．

しかし，2 値判別器の組み合わせでは，各判別器が独立であるために記事中のわずかなニュアンスから余分なタグを多く出力してしまうことが考えられる．この点に関しては，k-NN では出力される最大のタグ数がある程度限定されるため，ひとつの記事中であまりにも多くのタグが付与されることで，タグによる記事分類が機能しなくなるといった問題は起こりにくいと考えられる．

両手法は互いに一長一短であると考えられるが，今後は本手法を改良し，特化したタグも容易に出力が可能となるよう検討していきたい．

4. 関連研究

平野ら [6] は，ポータルサイトのディレクトリ型検索を基にして作成した分類先 (91 クラス) を採用し，Yahoo!掲示板の書き込みを訓練データとしてベイジアンフィルタによってブログの記事分類を行った．しかし，ブロゴスフィアの話題は果たして 91 種類で良いのかといった問題がある．既存の分類体系の利用だけでは，ブロゴスフィアに適した分類先クラスの設定は難しいと考えられる．本稿では，より粒度の細かい概念もタグによって取り込まれている．

Ohkura ら [7] は，分類先クラスとしては不適切なタグを排除するために，SVM で分類器を作成し，その分類の精度が低いものは分類先クラスとして不適切であると判断する手法を提案している．また，SVM を複数組み合わせることでマルチタグを付与する手法を提案している．本稿では，設定人数が多いが分類先として不適切なタグを除くだけでなく，表記ゆれや類義語をまとめるといったタグの統合を考慮している．また，k-NN には既存の関連文書検索機能を持った全文検索ソフトウェアとの親和性が高いといったメリットもある．

Brooks ら [8] は，各記事中で tf·idf 値が大きい語をタグとすることで，実際にブロガーがつけたタグよりも各記事間の類似度の平均値が大きくなることを示した．しかし，この手法では記事中にない語はタグとして付与されることがなく，また，タグ自体も非常にバラバラになってしまうといった事が懸念される．

5. おわりに

本稿では，k-NN を基にしてブログ記事に自動的にマルチタグを付与する手法を提案した．また，表記ゆれや類義語による類似タグを統合する手法，および「Weblog」等，あいまいなタグを識別する手法についての検討も行った．実際に自動マルチタグ付けを行うことで，簡単な評価実験を行った結果について報告した．

今後の課題としては，精度を向上させるために関連文書検索を行う際のデータベース中の記事全てに対し，再びタグ付けを行うことで元のタグと類似したタグが付与されるかにより，データベース中の記事を精選することを考えている．また，今回はブロガーがタグ付けが上手いか下手かという点を考慮していなかったので，ブロガーの属性という点も考慮していきたい．

参考文献

[1] Mathes Adam. Folksonomies - Cooperative Classification and Communication Through Shared Metadata, 2005.
http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[2] James Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business,Economies, Societies and Nations. Bantam Dell Pub Group, 2004.

[3] 藤村滋, 藤村考, 片岡良治, 奥雅博. Blog のタグ間類似度のスコアリング. 日本データベース学会 Letters,Vol.5,No.4, pp.33-36, 2007.3.

[4] Yiming Yang, Xin Liu. A Re-examination of Text Categorization Methods. 22nd Annual International ACM SIGIR Conference(SIGIR 99), pp.42-49,1999.

[5] 平林幹雄．全文検索システム Hyper Estraier の設計と実装．ACM SIGMOD 日本支部第 35 回 大 会,http://qdbm.sourceforge.net/mikio/hesigmodj.pdf,2006.

[6] 平野耕一，古林紀哉，高橋淳一．日本語圏ブログの自動分類．情報処理学会研究報告 (2005-NL-170), pp.21-26, 2005.11.

[7] Tsutomu Ohkura, Youji Kiyota, Hiroshi Nakagawa. Browsing System for Weblog Articles based on Automated Folksonomy. Workshop on the Weblogging Ecosystem(WWW2006), 2006.5.

[8] Chirstopher H.Brooks, Nancy Montanez. Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering. 15th International World Wide Web Conference(WWW2006), 2006.5