Blog のタグ間類似度のスコアリング

Similarity Scoring between Blog Tags

藤村 滋♣ 藤村 考♣
片岡 良治♣ 奥 雅博♣
Shigeru FUJIMURA Ko FUJIMURA
Ryoji KATAOKA Masahiro OKU

blog のタグに基づいて記事を集約することで，Folksonomy のメリットを生かした blog の分類システムの構築を目指している．blog のタグの現状分析の結果，タグが設定されている記事の割合が少ないといった問題に加えて，表記ゆれや類義語，多義語の問題によって，タグに基づく集約のみでは多くの課題があることが分かってきた．本論文では上記の問題のうち，特に表記ゆれや類義語の問題の解決を目指し，blog のタグ間の類似度の測定法を提案する．本手法では，語句の重みとして文書頻度とポアソン分布による推定文書頻度の差を採用し，タグの特徴ベクトルとして利用する．また，この類似度を用いて，タグの簡単なクラスタリングを行った結果についても報告する．

Our goal is to implement a blog classification system that utilizes the advantage of Folksonomy, by aggregating blog tags. A preliminary analysis of blog tags identified problems related to orthographic, synonymous and polysemic word variation, in addition to the problem of blog entries that had few tagged. Our solution is a new similarity measurement method for blog tags. It weights related terms based on differences between document frequency and value estimate obtained by assuming a Poisson distribution. These differences are used as the feature vector of tag. We use this similarity to conduct a simple tag clustering. Experiments show that the similarity measured by our proposed method matches the human view.

1. はじめに

消費者が発信する情報を集約することでコンテンツを形成するCGM(Consumer Generated Media)に注目が集まっている．CGMの多くのサービスは，個々のユーザーが設定した「タグ」というコンテンツに関するメタデータを集約し，タグを通して皆でコンテンツを分類するFolksonomy[1]を利用している．Folksonomyは，Folks(人々)とTaxonomy(分類学)を組み合わせた造語であり，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，タグにはコンテンツの話題ともいえる端的な語句が設定されることも多く，話題に沿った情報収集の有力な手段となること等が利点といわれている．

ここで，CGMの一形態であるblogの多くはタグの機能を有しており，bloggerが自身の記事を整理するために用いられている．一方で，他のCGMサービスと比較して，現状ではblogのタグは有効活用されているとは言い難い．Technorati™のタグ検索[1]をはじめとした，タグによる検索サービスに用いられる程度である．

blogのタグの現状分析を行ったところ，タグの種類はスケールフリー性に基づき膨大であり，表記ゆれや類義語の問題も含め，タグに基づく記事の集約のみでは，同一の話題を持つ記事の集約が難しいことが分かった．また，そもそもタグが設定されない記事が多いことや，多義性を持ったタグにより状況によっては不必要な記事までも集約するといった問題もある．

そこで，本論文では上記の問題の中でも，同一の話題を持つ記事の集約に向け，タグ間の類似度について議論する．タグの特徴ベクトル作成時の語の重み付け法として，実際の文書頻度とポアソン分布に基づく推定文書頻度の差を用いる手法を提案する．タグの特徴ベクトルを用いて，タグ間のコサイン類似度を測定する．タグ間の類似度によって，タグ検索における関連タグの推薦や，表記ゆれや類義語に対してタグの自動統合に役立てることが可能となる．

以下，本論文の構成を示す．2章では，blogのタグに関する現状の分析について述べる．次に，3章では，タグ間の類似度を求めるための特徴ベクトル作成手法について具体的に述べる．4章では，タグ間の類似度計測に関する簡単な実験について示す．5章では，関連研究について示す．最後に，6章では，まとめについておわりにとして記す．

2. blog のタグの現状分析

本章では，現状でのblogのタグの設定状況の調査・分析について報告する．まず，今回の調査対象としたblog記事の諸情報について表1に示す．

表1 調査対象のblog記事に関する情報
Table 1 Statistical Information of Blog Entries for Our Analysis

記事数	4,769,657
期間	2006/2/26～4/1(2006年3月分)
対象	日本語blog
blogger数	399,414
タグの種類	263,071

表1の対象における日本語blogとは，主にgooブログやLivedoorブログ等のホスティングサービスが中心である．また，次章以降での実験は，このデータを利用している．

上記の対象を基に，x軸を設定人数，y軸をタグの種類として両対数軸を用いてグラフ化したものが，図1である．両対数軸上ではタグの設定人数とタグの種類はほぼ直線を形成しており，Power law(べき乗則)が成り立っていると推測される．換言すると，タグの種類はタグ設定人数の-2乗に比例しており，次の式が成り立つ．

$$Num(tags) \propto Num(people)^{-2}$$

♣ 正会員 日本電信電話株式会社 NTTサイバーソリューション研究所 {fujimura.shigeru, fujimura.ko, kataoka.ryoji,oku.masahiro}@lab.ntt.co.jp

[1] http://www.technorati.jp/tags/ 2005年12月20日に日本でも開始

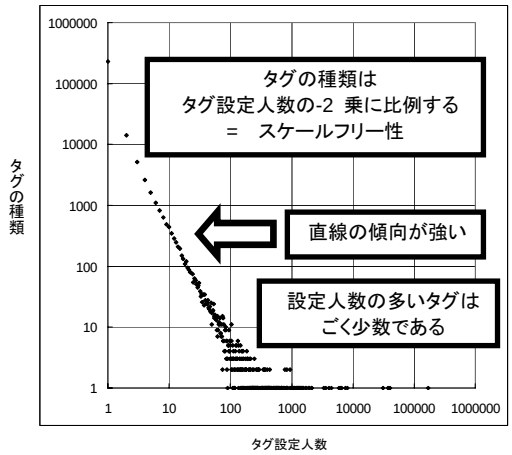


図1 タグ設定人数とタグの種類の関係
Fig.1 Number of Bloggers and Tags

したがって，設定人数の多いタグはごく少数であり，スケールフリー性を有していることが分かる．

また，タグ未設定の記事数は1,893,197件と約40%となっていた．ホスティングサービスによるタグ未設定時のデフォルトタグを考慮すると実際の割合はより大きいと考えられる．

さらに，設定人数が最も多いタグは「日記」であった．続いて，「未分類」，「Weblog」，「日記・コラム・つぶやき」であり，実際の記事内容を推測できないタグが設定人数上位にも含まれていることが分かった．

ここで，設定人数上位200位までの中で，食べ物に関する概念のタグを抽出したものが図2である．

()内は設定人数

グルメ・クッキング(1703), グルメ(938), 料理(854) , 食べ物(832), 食(782), お菓子(347), ごはん(292), パン(250), ラーメン(245), 食事(245), food(229), おいしいもの(207), おやつ(203)

図2 食べ物に関するタグの例
Fig.2 Tags Related to Foods

この例からも，同じような概念を持つと考えられるタグが乱立していることが分かる．また，「おいしいもの」が設定された記事の中には，ラーメンやお菓子について書かれた記事もあると考えられ，タグの意味的な粒度の違いによって，多義性が生じるといったことも考えられる．また，より直接的な多義語の問題としては，例えば「NEWS」はジャニーズのアイドルグループなのか，事件報道のことであるのかが分からないといった問題がある．

ただし，本論文においてはタグの表記ゆれや類義語の問題も含め，同一の話題を持つタグを特定することが特に重要と考え，タグ間の類似度測定によってこの問題の解決を目指す．

3. タグの特徴ベクトル作成手法

3.1 残差 df 値による語の重み付け

タグの特徴ベクトルを作成する場合においては，処理の対象は記事ではなくタグが設定された記事群である．従来から用いられてきた tf・idf 法などは，ひとつの記事内での語の重みを求める手法であるため，文書群での語の重みを求めるために最適化されている訳ではない．

そこで，本論文では**「特定のタグがつけられた記事群に偏って出現する語がそのタグの特徴語である」**という仮定を基に，語の出現の偏りを表す指標として，タグづけされた記事群での文書頻度とポアソン分布によって推定される文書頻度の差（以降，残差 df と呼ぶ）を用いる．残差 df は，文書群における推定文書頻度よりも実際の文書頻度が大きい語が大きなスコアを持つ．

ここで，文書においては，ポアソン分布は語がランダムに生起する場合の生起回数を確率的に表現するモデルである[3]．ポアソン分布は次の式で表される．ここで k は生起回数であり，λは期待値である．

$$P(k;\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$$

ここでF_iを語iの大域的頻度，nを全文書数とすると，ある文書中で語iが一回以上出現する確率pは以下の式で表される．

$$p = 1 - P(0;\frac{F_i}{n}) = 1 - e^{-\frac{F_i}{n}}$$

さらに，df_{ij}をある語iのタグj中での文書頻度とし，n_jをタグjに属する全文書数とすると，残差df値rdf_{ij}は次の式で求められる．

$$rdf_{ij} = df_{ij} - n_j(1 - e^{-\frac{F_i}{n}})$$

本論文では，rdf 値の閾値を 1.0 とし，閾値以上の語を特徴ベクトルの要素として採用している．

ここで，ポアソン分布に基づく推定文書頻度を利用した指標としては，文献[3]に述べられているように，一般的に残差 idf が知られている．従来，残差 idf は大域的頻度の大きな語は結果的に df 値も大きくなり，idf 値が小さくなるため，語の重み付け法である tf・idf 法がうまく機能しないといった問題を解決するために用いられてきた．重要な語は同一の文書内で複数回繰り返されて使われるため，ポアソン分布によって推定される df 値よりも実際の df 値が小さくなるという仮定に基づいている．

一方で，同一のタグが付与された記事群は内容的にも，用いられている語彙的にも似ている可能性が高い．残差 df は，同一タグが付与された記事群内で，内容を表すような語句はその記事群中に偏って出現している可能性が高いため，その df 値はポアソン分布によって推定される df 値よりも大きくなるという仮定に基づいている点で異なっている．

換言すると，残差 idf は語が内容語であるかを考慮した上で，文書特定性が高い語を調べるための手法であり，残差 df は特定の文書群の中での内容語を調べるための手法である．

3.2 コサイン類似度によるタグ間類似度計算

前節で述べたタグの特徴ベクトルを利用してタグ間の類似度を求める．類似度の計算法としては，一般的に用いられているコサイン類似度を利用する．C_i, C_jをカテゴリの特徴ベクトルとすると，類似度は以下の式によって求められる．

$$similarity = \frac{C_i \cdot C_j}{|C_i||C_j|}$$

4. 実験・考察

4.1 タグ間類似度のスコアリング

前章で述べた手法を基に，タグの特徴ベクトルを作成した．本論文では，計算量の問題から設定人数上位 5,000 タグまで特徴ベクトルを求めた．ただし，「日記」，「未分類」，「Weblog」や「(空白)」のタグは，設定人数，記事数共に膨大なため対象から除外している．さらに設定人数上位のタグは記事数が多いので，特徴ベクトルを作成する際にはランダムに選んだ 5,000 記事を用いている．

実際に，得られた特徴ベクトルの例を表 2 に示す．

表 2 タグの特徴ベクトルの例
Table 2 Examples of Word Vector of Tag

タグ「映画」		タグ「スポーツ」	
Term	rdf	term	rdf
映画	3314	日本	1441
観	1337	WBC	1108
作品	1018	試合	1108
シーン	777.8	選手	945.0
監督	772.8	スポーツ	866.4
面白	733.8	チーム	796.5
物語り	689.7	勝	700.8
ストーリー	685.3	Korea	699.5
ＤＶＤ	589.1	負け	689.5
見	558.1	America	672.8
:	:	:	:
(要素数：約 9,500)		(要素数：約 7,500)	

表から，rdf 値上位の語は直感的にもタグの内容を表していると考えられる．また，「スポーツ」については blog を収集した期間中に野球の WBC(World Baseball Classic)が開催されており，大きな話題となった．そのため多数の野球関連の語が大きな重みを持つこととなっているが，これはその時々の blog の状況をよく反映しているといえる．

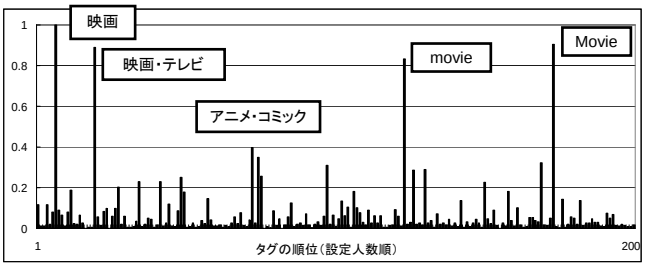
次に，タグ間の類似度の例について図に示す．「映画」とタグの設定人数順上位 200 位までのタグ間での類似度を表したものが図 3(a)である．同様に，「グルメ」の類似度を表したものが図 3(b)である．

この結果から，表記ゆれや類義語により同一の概念を指すと考えられるタグ間の類似度は特に大きいことが分かる．また，図 3(b)に注目すると，「グルメ」は比較的大きな類似度を持つタグが多いことが分かる．したがって，食べ物に関する概念を持つタグは表記ゆれや類義語の影響が大きいと言える．

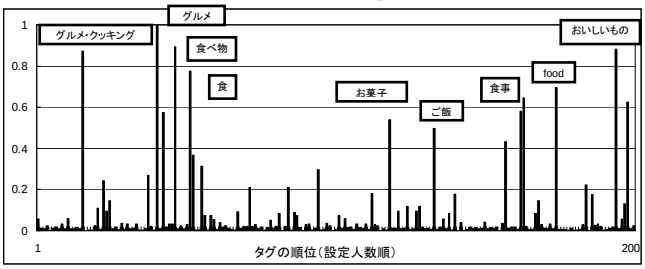
また，比較対象として，各記事ごとに tf・idf 重み付けを行った単語ベクトルを求め，タグの特徴ベクトルをそのタグが設定された記事の単語ベクトルの平均とする（以降，tf・idf 平均手法と呼ぶ）ことでタグ間類似度を求める実験を行った．この手法で得られた「グルメ」の類似度グラフは図 3(c)である．

両グラフを比較すると，残差 df による特徴ベクトルの方が類似度のピークがはっきりと現れる手法であるといえる．tf・idf 平均による手法では，tf の影響によって特徴ベクトル内に一般語が含まれているため，類似度が平均的に大きな値になっていると考えられる．換言すると，類似度が特に大きい部分では両手法に顕著な差は見られないが，残差 df による手法では，例えば，「お菓子」や「ご飯」といった「グルメ」とある程度の関連性を持っているタグを識別可能である．

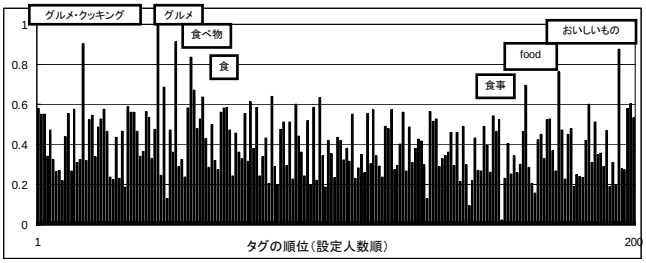
上記の点から，残差 df による特徴ベクトルを用いた手法がタグ間の類似度計算にはより適していると考えられる．



(a)タグ「映画」



(b)タグ「グルメ」



(c)tf・idf 平均によるタグ「グルメ」

図 3 類似度グラフ
Fig.3 Similarity Graph

4.2 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自動的に解決するため，タグ間の類似度を利用したクラスタリングを行うことにより同一概念のタグの統合を試みた．

ここで，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．その理由としては，類似度の再計算の必要がないため高速であることが期待されるためである[4]．

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．一方で，タグの組み合わせは最大 $_{5000}C_2$=1,250 万組考えられる．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 4 に示す．図から，例えば「弐寺[2]」と「音ゲー」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

[2] コナミ株式会社の DJ シミュレーションゲームの略称

グが，一つのクラスタを形成しており，ユーザーが関連するタグを思いつくことが出来ないような状況において，有効な支援が可能であることを示唆している．

```
<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">弐寺</node>
        <node no="2">音ゲー</node>
        <node no="3">IIDX</node>
</cluster>
```

図 4 クラスタの例
Fig.4 The Example of Cluster

5. 関連研究

平野ら[5]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)を用いて，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 blog の記事分類を行っている．本研究では，blogger が設定したタグを用いることで，より粒度の細かな話題にも対応が可能になると考えている．

Brooks[6]らは，タグの関連性を文書間の平均コサイン類似度として設定し，2 分木構造を基にした階層型クラスタリングにより，タグ間の階層性を設定する手法を提案している．しかし，この手法では文書間の類似度を基にしているため，類似度はそれほど大きな値とならない．タグ間の類似度を考慮した本手法の方がコサイン類似度のピークの差がより大きく表れるため，結果として，タグの統合の際の精度が高くなると考えられる．

6. おわりに

本論文では，まず blog のタグについての現状分析を行った結果，タグの種類と設定人数についてスケールフリー性が成り立っていること，および blog のタグには表記ゆれや類義語の問題，多義語の問題，タグ未設定記事が多いといった問題があることを示した．次に，タグの表記ゆれや類義語の問題の解決に向け，タグ間の類似度を計算する際にタグの特徴ベクトル中の語の重み付けの方法として，実際の文書頻度とポアソン分布によって推定された文書頻度との差を利用する方法を提案した．また，この手法で得られたタグの特徴ベクトルを用いて，実際にタグ間の類似度の計測を行った結果，関連性があると考えられるタグ同士の類似度は確かに大きくなること示した．また，簡単なクラスタリングを行った結果，良好な結果が得られた．

今後の課題としては，以下の点が挙げられる．本論文では，類似度の測定に関して他手法との詳細な比較までは行っていない．従って，今後は他手法との定量的な比較検討が必要になると考えられる．また，類似度が高いタグ同士が人間の直感に従っているかどうかも調査する必要がある．

また，今後はタグ未設定の記事や複数の話題を持つのにシングルタグしか設定されていない記事に対して，適切なタグをマルチタグとして自動付与することも検討していきたい．

[文献]

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

[2] James, S.: “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] 北研二, 津田和彦, 獅々堀正幹 :”情報検索アルゴリズム”, 共立出版(2002).

[4] 宮本貞明 :“クラスター分析入門 – ファジィクラスタリングの理論と応用”, 森北出版(1999).

[5] 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)

[6] Chirstopher, H, B. and Nancy, M.: “Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering”, 15th International World Wide Web Conference (WWW2006), pp.625-632, (2006).

藤村 滋 Shigeru FUJIMURA
NTT サイバーソリューション研究所 所属．2005 年 東京大学大学院情報理工学系研究科修士課程修了．同年，日本電信電話株式会社入社．Web マイニングの研究に従事．日本データベース学会会員．

藤村 考 Ko FUJIMURA
NTT サイバーソリューション研究所 主任研究員．電気通信大学大学院情報システム学研究科客員教授．1989 年 北海道大学大学院工学研究科博士課程修了．同年，日本電信電話株式会社入社．トランザクション処理記述言語，汎用電子チケットシステム，電子決済システム，blog マイニングの研究開発に従事．工学博士．情報処理学会，電子情報通信学会，日本社会情報学会会員．

片岡 良治 Ryoji KATAOKA
NTT サイバーソリューション研究所 主幹研究員．1987 年 千葉大学大学院電子工学専攻修士課程修了．同年，日本電信電話株式会社入社．トランザクションの並行処理制御方式の研究，マルチメディア情報システムの研究，ポータルサービスシステムの研究開発に従事．情報処理学会会員．

奥 雅博 Masahiro OKU
NTT サイバーソリューション研究所 主幹研究員．1984 年 大阪府立大学大学院工学研究科博士前期課程修了．同年，日本電信電話公社(現 NTT)入社．機械翻訳，日本文推敲支援技術等の自然言語処理，検索をはじめとするポータルサービスの研究開発に従事．博士(工学)．情報処理学会，電子情報通信学会，言語処理学会会員．