2H1-4

Folksonomyタグを用いた個人の視点に基づくコンテンツ検索手法

Content Retrieval Based on Personal Viewpoints and Folksonomies

数原 良彦*1 篠沢 佳久*1 櫻井 彰人*1
Yoshihiko Suhara Yoshihisa Shinozawa Akito Sakurai

*1慶應義塾大学大学院理工学研究科
Graduate School of Science and Technology, Keio University

There have emerged content management service, which is called folksonomy, focused on web content sharing with tags put by users to the contents. The tags might be considered as of common meaning but in fact are of different meaning depending on concepts of users who put the them. This is one of the reasons why conventional tag search engines do not give appropriate replies. We thought the problem might be solved by building a classifier that classifies contents based on the tags that the target user had put. The experiments gave positive results.

1. はじめに

近年，ウェブ上でコンテンツを管理，共有するサービスが普及しており，ユーザがタグと呼ばれる自由記述のメタデータを付与することで管理を行う folksonomy と呼ばれる分類体系を用いられている．Folksonomy で通常用いられるタグ検索では，クエリタグが付与されたコンテンツを取得するが，コンテンツに付与されたタグは，そのタグを付与したユーザの視点に基づいているため，例えタグ名称が同じ場合でも検索の目的とする意図とは異なる概念によって付与された可能性がある．そのためタグ検索によって異なる視点で付与されたコンテンツが取得されてしまい，目的のコンテンツを見つけ出すのが困難であるという問題が指摘されている [Voss 07]．

そこで本研究では folksonomy において，ユーザの視点に基づいたコンテンツ検索の実現を目指す．我々は，ユーザが同一タグを付与したコンテンツ群は，あるひとつの概念を反映していると考え，未知コンテンツがその概念に含まれるかを判別する分類器を学習することで未知コンテンツ群をあたかもユーザの視点に従ってタグ付けしたものとし，ユーザの視点に基づいたコンテンツ検索を行う手法を提案する．

具体的には，コンテンツを記述する特徴量として，当該コンテンツに付与されたタグを用いる．各タグに対する視点はユーザ毎に異なるかもしれないが，各ユーザ毎に一貫し，したがってあるユーザの分類器学習時には，当該ユーザの視点に沿ったタグが他に優先して用いられるであろうと考えたからである．

Folksonomy を用いたコンテンツ検索・推薦の研究は近年盛んに行われている．丹羽らはユーザの嗜好をユーザとタグの親和度によって表現し，ユーザ間の嗜好類似度に基づくコンテンツ推薦手法を提案している [丹羽 06]．この手法ではタグ名称が同じものを同一の意味づけをしていると解釈するため，個人の視点に基づいたコンテンツ検索の実現は難しいと考えている．また，佐々木らは，各ユーザが同じタグを付与したコンテンツ群におけるコンテンツの共起性に基づく推薦システムを提案している [佐々木 07]．この手法では，選択されたタグが付与されたコンテンツ群へ未知コンテンツを推薦するためには，他ユーザが同じ名称のタグを付与したコンテンツ群との類似度を計算する必要があり，計算量が多くなるという問題がある．

連絡先: 数原良彦, 慶應義塾大学大学院理工学研究科, 〒 223-8522 神奈川県横浜市港北区日吉 3-14-1, suhara@ae.keio.ac.jp

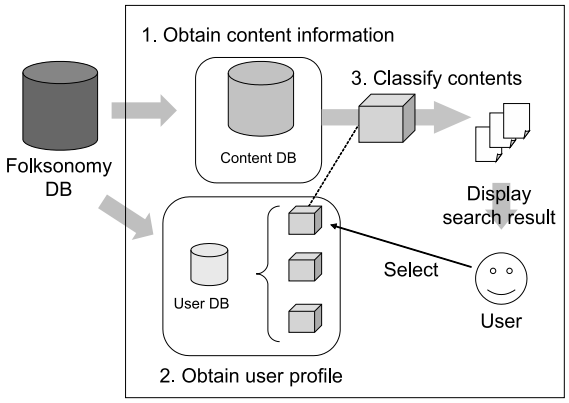

図 1: 提案手法の概要

2. 提案手法

本研究では folksonomy ユーザが使用しているひとつのタグが付与されているコンテンツ群がユーザのひとつの視点を反映すると考え，ユーザが使用しているタグ毎に分類器を学習することで，未知コンテンツに対して当該ユーザの視点でタグを付与する手法を提案する．

我々は，全てのユーザがそれぞれの視点に基づいてコンテンツの内容を特徴づけるようにタグ付与を行っているとし，多数のユーザによって付与されたタグがコンテンツの特徴をよく表現するのではないかと考えた．そこで，コンテンツの特徴として複数ユーザが付与したタグの頻度集合を特徴として分類器の学習を試みる．なお，コンテンツに付与されたタグを，あたかもテキストを bag-of-words として扱うのと同様に使用するため，本稿ではこれを **bag-of-tags** と呼ぶことにする．

具体的には，各ユーザ毎に，コンテンツを入力としユーザが付与した当該タグをクラスとする分類器をコンテンツの表現を bag-of-tags を用いて作成することを提案する．図 1 に提案手法の概要を示す．

提案手法では，ユーザが当該タグを付与したコンテンツを正例とし，ユーザがタグ付与を行ったコンテンツ集合の中から当該タグが付与されていないものを負例の訓練事例として分類器の学習を行う．

表 1: 提案手法とベースライン手法の比較

	マクロ平均			マイクロ平均		
	再現率	適合率	F 値	再現率	適合率	F 値
MNB (Tag)	0.518	0.135	0.208	0.595	0.161	0.245
MNB (Text)	0.081	0.087	0.082	0.185	0.277	0.207
CNB (Tag)	0.572	0.123	0.195	0.639	0.139	0.216
CNB (Text)	0.083	0.089	0.084	0.188	0.265	0.207
SVM (Tag)	0.221	0.303	0.254	0.249	0.337	0.284
SVM (Text)	0.093	0.148	0.107	0.158	0.301	0.194

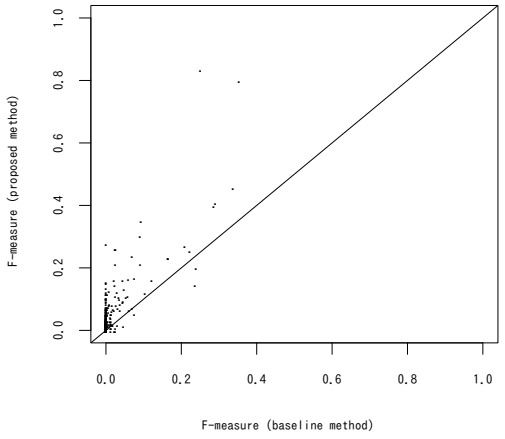

図 2: 提案手法とベースライン手法の F 値の比較

3. 評価

3.1 評価実験

2007 年 5 月から 6 月にかけて取得した del.icio.us のデータ約 1.6 万ユーザについてのブックマークデータとそのタグ情報を取得した．その中から 17 人のユーザが使用したタグ合計 759 個について，評価実験を行った．評価は 2-fold cross validation (2-CV) で行った．実験には，データマイニングツール Weka[*1]を使用し，各分類器の実装として MNB には NaiveBayesMultinomial を，CNB には ComplementNaiveBayes を，SVM には SMO を用いて，正例の適合率，再現率，F 値によって評価を行った．

提案手法の有効性を評価するための比較手法として，コンテンツの URL に存在するウェブページから HTML タグの除去，不要語の除去，接辞処理を行った単語の頻度ベクトルで表現される bag-of-words をベースライン手法として用いた．なお，不要語の除去には Perl モジュールの Lingua::EN::StopWords を利用し，接辞処理には Porter アルゴリズムを用いた．

3.2 結果

結果を表 1 に示す．また，全てのタグの分類結果について，提案手法とベースライン手法の F 値をプロットしたものを図 2 に示す．図 2 の x 軸はベースライン手法の F 値，y 軸は提案手法の F 値を表す．したがって，$y = x$ の直線よりも上側に存在するプロットは，当該タグの分類実験において提案手法がベースライン手法よりも高い精度を示したことを表している．

NB 手法におけるマイクロ平均適合率を除き，全ての指標において提案手法がベースライン手法より高い値を示した．ベースライン手法では再現率が著しく低い値を示しており，このことから誤って負例と分類してしまう FN (false negative) の数が多いことが推測できる．これよりテキストの内容を特徴とした手法では正例，負例を分類する特徴をうまく選択することができず，ほとんどを負例と判別していると考えられる．

提案手法ではベースライン手法に比べ，再現率，F 値ともに高い値を示しており，bag-of-tags によって分類器をより適切に学習することがいえる．したがって，bag-of-tags はコンテンツの特徴をよく表現していると考えられ，複数のユーザによって付与されたタグは，全体としてコンテンツの特徴を反映すると考えられる．

4. おわりに

本稿では，folksonomy ユーザが使用しているタグが，当該ユーザのひとつの視点を表していると考え，folksonomy におけるタグ情報を用いることで，当該タグが付与されたコンテンツ群についての分類器を学習し，未知コンテンツを判別することでユーザの視点に基づく検索手法を提案した．

評価実験によって，テキストから抽出された単語の頻度ベクトルである bag-of-words を特徴とするベースライン手法に比べ，本稿で提案したコンテンツに付与されたタグの頻度ベクトルで表現される bag-of-tags によって高い精度で分類できることを示した．

本研究では，ユーザが付与したひとつのタグがひとつの視点に基づくとみなして提案を行った．しかし，実際にはひとつのタグを異なる視点で付与したり，異なるタグを同じ視点で付与することが起こると考えられる．このようなタグ付与に対応し，より的確な意味で個人の視点に基づく手法を今後の課題として考えている．

参考文献

[Voss 07] Voss, J.: Tagging, Folksonomy & Co-Renaissance of Manual Indexing?, Proc. the International Symposium of Information Science, pp.234-254 (2007)

[佐々木 07] 佐々木祥, 富田高道, 稲積泰宏, 小林亜樹, 酒井義則: Social Bookmark におけるコンテンツクラスタ間の類似度を用いた web コンテンツ推薦システム, 情報処理学会論文誌: データベース, Vol.48, No.SIG20 (TOD36), pp.14-27 (2007)

[丹羽 06] 丹羽智史, 土肥拓夫, 本位田真一: Folksonomy マイニングに基づく Web ページ推薦システム, 情報処理学会論文誌, Vol.48, No.5, pp.1382-1392 (2006)

∗1 http://www.cs.waikato.ac.nz/ml/weka/