Social Bookmarkにおけるユーザのタグによる分類傾向に基づいた情報推薦*

高野 博一[†]　古瀬 一隆[‡]　陳 漢雄[§]
[†,‡,§,¶] 筑波大学 システム情報工学研究科　〒305-8577 茨城県つくば市天王台 1-1-1

1　はじめに

本稿では, Social Bookmark（SBM）におけるブックマーク情報を他のユーザの情報から推薦する手法を提案する. このようなSBMに関する研究の多くはタグの文字列に着目したものである.

タグの文字列に着目した研究としては, タグがカテゴリ分けの役割を果たすと考え, 推薦対象となるユーザの登録タグと同じタグを用いたWebページを発見し, 推薦を行う手法がある. しかし, このような手法は, タグの文字列がWebページの情報を万人に向けて表している前提が必要がある. 現在, SBMサービスのタグの文字列はほとんど統制が行われておらず, 自己の利便性や表現の一種としてタグ付けを行うユーザが増加している. このため, タグの文字列に着目した研究では、推薦の精度が低くなる問題がある.

そこで, 本稿ではタグは分類におけるただの仕切りとし, タグによって分類されるWebページ群そのものに着目する. 各タグ同士のWebページ群の重なり具合によって重み付けを行い, 推薦すべきタグを選出する. このように選出されたタグの推薦対象ユーザが登録していないWebページを推薦することで, 精度の高い推薦が可能となる.

2　SBMの構造

ここで, SBMの構造について簡単に説明を行う.「ユーザ」はブックマークしたい「Webページ」に好きな「タグ」をつけブックマークする. 基本はこの,「ユーザ」,「タグ」,「Webページ」の三つ組みの構造となり, 本稿でも, この三つ組の構造を用いる.

3　タグによる分類傾向を用いた推薦手法

図1において, タグの文字列に着目した場合, ユーザ1は「memo」ユーザは「java」,「memo」と, それぞれのタグが付けられたページ集合がある. このようなケースでは, ユーザ1の「memo」でタグ付けされたページ集合と類似度が高いのは, 同じ「memo」でタグ付けされたページ集合ではなく,「java」でタグ付けされたページ集合であると言える. このように, タグは分類におけるただの仕切りとし, タグによって分類されるWebページ群そのものをタグによる分類傾向とし, この傾向を用いて情報推薦を行う.

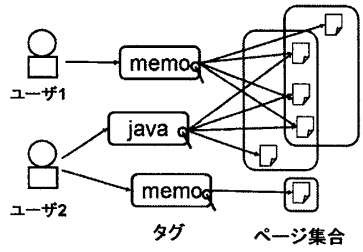

図1　タグの文字列に依存しない類似度の例

3.1　既存手法

ユーザ1のタグ i とユーザ2のタグ j におけるページの重なり方のベン図を図2に示す. この v, w, x, y, z を用いて, 類似度の算出を行う.

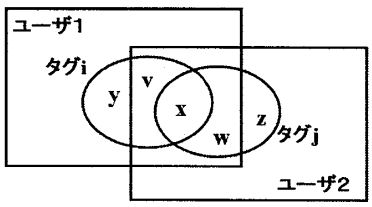

図2　2ユーザのタグによるページの重なり方

3.1.1　Jaccardを用いた類似度

単純な類似度計算手法として, Jaccard係数を用いた手法が挙げられる. 図2の場合は,

$$\frac{x}{y+v+x+w+z} \tag{1}$$

と表すことができる. しかし, この手法は, 単純な割合となるため, 母数の少ないタグが有利なる問題がある.

---

* Infomation Recommendation Based on Classification Tendency of Tagging in Social Bookmark System

† Takano Hirokazu, Graduate School of SIE, University of Tsukuba
‡ Kazutaka Furuse, Graduate School of SIE, University of Tsukuba
§ Hanxiong Chen, Graduate School of SIE, University of Tsukuba

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

3.1.2 尤度比検定を用いた類似度

また, 宮田らは尤度比検定を用いた手法を用いている.[1] 尤度比検定を用いた手法を図 2 を用いて式にすると,

$$log\left(\frac{{}_{v+x+w}C_x\, p_1{}^x(1-p_1)^{v+w}}{{}_{v+x+w}C_x\, p_0{}^x(1-p_0)^{v+w}}\right) \quad (2)$$

このようになる. p_1, p_0 については, 宮田らの値をそのまま用いることとし, $p_1 = 0.6, p_0 = 0.1$ とする. この式では, 図 2 の z の値を考慮していない. そのため, z が大きくなりがちな複数のトピックにまたがるようなタグの類似度も大きくなってしまう. また, x の割合が少ない場合, x の集合があるにもかかわらず, 類似度が 0 以下となるケースも存在する. そこで, 本稿ではどのようなケースにも対応できるような類似度計算手法を提案する.

3.2 提案手法

本稿で, 提案する式は以下のような式である。

$$\frac{x^2}{log(v+w)log(z)} \quad (3)$$

式 Jaccurd 係数のスコアは割合であったが, 式 3 は値となるので, 母数の少なさに影響されにくい. また, 尤度比検定では未考慮であった z の大きさも考慮している. y については, 式 3 でも考慮していない. それは, ユーザ 1 のタグ i が複数のトピックにまたがるようなタグだとしても, そのなかの類似した部分については, 十分に推薦対象となりうるためである.

4 実験

我々は, はてなブックマーク [2] からデータを収集した 1000 人のユーザとそれらのユーザによってブックマークされている約 190 万の Web ページの URL, URL につけられているタグ約 15 万件を対象に実験を行った.

ユーザとタグの組み合わせをランダムに 100 件を対象タグとして抽出する. これらの対象タグに登録されているブックマークのうちランダムに URL を 1 件削除する. この上で, 1000 人のユーザの各タグを比較タグとし, 上記の各式それぞれによる類似度の上位 n 件を抽出する. 上位 n 件のうち, 削除された URL を含む比較タグが含まれているかどうかを判断し, 対象タグ 100 件のその割合を再現率として, 実験を行った. この v, w, x, y, z を図 3 をみると, 提案式が他の式に比べ, 常に高い再現率であることが分かる. また, どの手法においても, 再現率の伸びが悪くなるのは, 削除された URL が他のユーザによってブックマークされていないケースが存在するためである.

次に行った実験は被験者を用いた適合率実験である. この実験は, まず, 7 人の被験者にそれぞれ, 2, 3 件のトピックに対するお気に入りのページを 20 件程度リストアップしてもらう. そして, そのリストアップされたページを用いて, DB から類似度の上位 20 件のユーザとタグ, そのタグ付けがされている URL の集合を抽出する. 各 URL のスコアは付けられたタグの類似度の合計とし, スコアの高い URL 上位 10 件を被験者に評価してもらう. 評価基準は, 同じトピックとして価値の高いものから順に, A, B , C の 3 段階である.

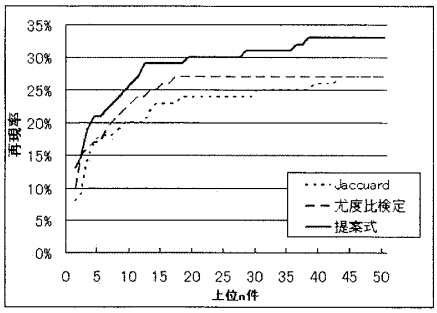

図 3 各式の再現率

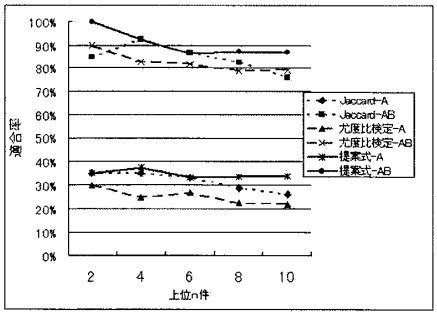

図 4 各式の適合率

図 4 をみると, 提案式が, 他の式に比べ比較的良い結果を出していることが分かる.

5 まとめ

本稿では, SBM におけるユーザのタグによる分類傾向に基づいた情報推薦の手法を提案した. この手法により, 従来のタグの文字列を見ない手法よりも精度の高い情報推薦が可能になった.

参考文献

[1] 宮田高道, 佐々木祥. SBM データを用いた web コンテンツ推薦. SBM 研究会, 発表資料, 2007.

[2] はてなブックマーク. http://b.hatena.ne.jp/.

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.