

## ソーシャルブックマークにおける ユーザのタグ付与行動分析に基づくコンテンツ分類

大石 剛司<sup>†</sup> 亀田 堯宙<sup>††</sup> 深見 嘉明<sup>‡</sup> 大向 一輝<sup>‡‡</sup> 武田 英明<sup>‡‡</sup>

<sup>†</sup> 東京大学工学部 <sup>††</sup> 東京大学大学院 <sup>‡</sup> 慶應義塾大学 SFC 研究所 <sup>‡‡</sup> 国立情報学研究所

### 1 はじめに

現在ブロードバンドの整備やPCの高速化によってインターネット上にあらゆるコンテンツが氾濫している。その中から自分にとって必要なコンテンツを探すために、ユーザが自分の興味のあるコンテンツを共有することができる del.icio.us<sup>†</sup> やはてなブックマーク<sup>††</sup> などのソーシャルブックマークサービス (以下 SBM と称す) が存在する。

SBM ではユーザがタグと呼ばれるキーワードを付与することによるコンテンツ分類を実現しており、従来の Yahoo!Japan<sup>‡</sup> のような、コンテンツの提供者によるディレクトリ型の分類とは違い、ポトムアップ型の分類でコンテンツ閲覧者から見れば好ましく思えるが、一方で、ユーザの多様性と同時に、付与されるタグが多様すぎてコンテンツが必ずしも適切に分類されていないという問題がある。また、SBM のユーザがある種のコミュニティを形成して、それぞれのコミュニティで主要となる興味・関心が異なるという性質がある。本研究では、以上を踏まえ、SBM のメタデータを用いて、コミュニティの主要な興味・関心を抽出し、これに基づくコンテンツの分類、評価、推薦を行うことを目指す。

### 2 研究目的

Golder ら [1] は、SBM の中でのタグの使用目的を調査し、タグを分類した。

その分類は基本的に客観的なタグ (コンテンツの主題、内容の種類、作成者を表す) と主観的なタグ (単独では意味がない分類のためのタグ、ユーザの感情を表す形容詞、コンテンツとユーザに関係するもの) に分けられる。この研究は SBM におけるタグ自体の内容に

注目した分析が行われている。

一方、本研究では個々のタグの意味内容に基づいた分類ではなく、コミュニティの興味・関心に基づく分類を目的としている。SBM でタグを分析する場合、同一タグを何人のユーザが何回使用したかを分析することで分類することが可能である。一方、個々のコンテンツをコミュニティの興味・関心に沿っているかを評価する場合、同一コンテンツに何人のユーザがタグを付与していることを分析することでコンテンツの「注目度」を抽出することができるが、それだけでは本研究の目標を達成するには不十分である。そこで、本研究では、ユーザが付与したタグの内容ではなく、ユーザのタグ付与行動自体に着目し、時系列に沿ってユーザの行動履歴を見ていくことによってコミュニティ全体の興味・関心を抽出することを目的とする。

### 3 研究手法

SBM では、ユーザがコンテンツにタグを付与する時、そのタグは自由な文字列を設定できる。ただ、分類・整理の観点から、以前使った文字列と同じタグを利用する傾向が強い。しかし、ユーザにとって、既存のカテゴリに納まらないコンテンツの場合には新しいタグ (「新規タグ」と定義する) が設定されると思われる。そこで、本研究では、ユーザがコンテンツに付与したタグが新規タグであれば、そのコンテンツにはなんらかの「新規性」がみられるのではないかと考える。

そして、付与されたタグの中で「新規タグ」の割合が低いコンテンツは、コミュニティ全体での主な興味・関心の対象となっているコンテンツであり、逆に、「新規タグ」の割合が高いコンテンツは、コミュニティ全体にとって「新規性」のあるコンテンツであると予測される。

詳細な研究手法については以下の通りである。

- 研究対象である SBM のデータを id-コンテンツの URL-タグ-ユーザ ID-登録された日時、という形式で用意する
- ユーザ毎にコンテンツの URL、タグ、登録された日時という履歴を作成する
- 上のユーザ行動履歴を参照して、コンテンツ毎に付与されたタグが新規タグであれば「新規ポイン

Contents Classification based on User Behavior on Social Bookmarking Service

<sup>†</sup> Takeshi OHISHI(starsky.508@gmail.com)

<sup>††</sup> Akihiro KAMEDA(kameda@race.u-tokyo.ac.jp)

<sup>‡</sup> Yoshiaki FUKAMI(yofukami@sfc.keio.ac.jp)

<sup>‡‡</sup> Ikki OHMUKAI(i2k@nii.ac.jp)

<sup>‡‡</sup> Hideaki TAKEDA(takeda@nii.ac.jp)

Faculty of Engineering, The University of Tokyo (<sup>†</sup>)

Graduate School of Frontier Sciences, The University of Tokyo (<sup>††</sup>)

Keio Research Institute at SFC (<sup>‡</sup>)

National Institute of Informatics (<sup>‡‡</sup>)

7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan

<sup>†</sup> del.icio.us <http://del.icio.us/>

<sup>††</sup> はてなブックマーク <http://b.hatena.ne.jp/>

<sup>‡</sup> Yahoo!Japan <http://www.yahoo.co.jp/>

表 1: 新規性指数の下位のコンテンツ

タイトル	総タグ数	新規性指数	主なタグ
ITmedia: Yahoo!と eBay の提携が Microsoft に及ぼす圧力	32	0.000	Microsoft, Yahoo!, Google
CNET: 映画ダウンロードの先の Apple と Google の提携	51	0.059	Apple, Google, 動画
ITmedia: ヤフーが CGM 化の号令, MySpace は連携も	31	0.065	Yahoo!, SNS, MySpace
CNET: 「EC ナビ人気ニュース」に新機能を追加	42	0.071	EC ナビ, CNET, News
MYCOM: グッズが集合するイベントの代表者に聞く	40	0.075	ガンダム, キャラ, ホビ

表 2: 新規性指数の上位のコンテンツ

タイトル	総タグ数	新規性指数	主なタグ
百式 - ネガティブ目標	31	0.871	酒, タバコ, 管理
山口もえ感涙! ほのぼの披露宴	40	0.850	芸能, 結婚式, おめでとう
EC ナビ人気ニュース 版 / ニュースに特化した SBM	34	0.794	EC ナビ, キーワード, 祝
asahi.com: アドウェイズが上場 社長上場企業最年少	32	0.781	上場, アフィリエイト, アドウェイズ
ITmedia: 人の心を読むコンピュータ, 間もなく登場?	41	0.780	テクノロジー, 科学, コンピュータ

ト」として 1, そうでなければ 0 を付加する

- コンテンツ毎に新規ポイントの合計を集計する
- 新規ポイントの合計をコンテンツ毎に付与されたタグ数の合計で割ることで「新規性指数」を算出する
- 付与されたタグ数の合計が 30 以上のコンテンツの中で新規性指数が高い集合と低い集合の上位を分析する

上記研究手法を用いた実験結果を次節で示す。

## 4 実験・考察

### 4.1 データ詳細

本研究において使用したデータは、日本の代表的なソーシャルブックマークサービスの一つである EC ナビ人気ニュース (現: Buzzurl[2]) の 2005 年 10 月 25 日から 2006 年 12 月 5 日の 406 日間のブックマークデータ 139597 件を採用した。基本的な統計情報としては、ユニーク URL 数が 88742, 付与されたユニークタグ数が 31076, 延べタグ数が 264058, 全ユーザ数が 7579 人である。

### 4.2 新規性指数が低いコンテンツ群

新規性指数の下位のコンテンツを表 1 に示す。

タイトルを見てみると Web サービスや Web に関わる企業に関するニュースで, CNET, ITMedia といった情報源からのものも多い。ここでは, SBM のコミュニティが定期的に関連しているサイトの情報が多い傾向があると言える。また, 趣味性の高いコンテンツに関しても同様のテーマの情報が比較的下位に存在し,

コミュニティの主要な興味・関心に含まれていると考えられる。コンテンツに付与されたタグも「MySpace」や「Yahoo」, 「Google」といった Web に関係するもので占められていた。新規性指数は 0.1 以下でほとんど全てのタグが以前に使用したタグである。

### 4.3 新規性指数が高いコンテンツ群

新規性指数の上位のコンテンツを表 2 に示す。

タイトルを見てみると生活や芸能・ニュースと言った内容のコンテンツが多いが, 統一的なテーマは見られず, 瞬間的な流行, 単発のイベントなどの情報となっている。また, 付与されたタグの中には, 「おめでとう」, 「祝」などの主観的なタグも見られた。新規性指数は 0.7 ~ 0.8 で付与されたタグのほとんどが新規タグである。

## 5 おわりに

SBM におけるユーザのタグ付与行動に注目し, 新規性指数を導入することによって, コンテンツを「新規性」という軸で分類し, コミュニティ全体の興味・関心を抽出することができた。今後の課題としては, 新たなコンテンツの分類基準を確立させることでコンテンツ推薦を行うことを目指す。

## 参考文献

- [1] S.A.Golder and B.A.Huberman, Usage Patterns of Collaborative Tagging Systems, Journal of Information Science, 32(2), pp.198-208, 2006
- [2] Buzzurl <http://buzzurl.jp>