1H2-1

ソーシャルブックマークを用いた情報収集・分析ツールの開発

Information Gathering and Analysis Techniques Using Social Bookmark Data

宗片健太朗[*1] 福原 知宏[*2] 山田 剛一[*1] 絹川 博之[*1] 中川 裕志[*3]
Kentaro Munekata Tomohiro Fukuhara Koichi Yamada Hiroshi Kinukawa Hiroshi Nakagawa

[*1]東京電機大学 情報メディア学科 計算言語学研究室
Computational Linguistics Laboratory, Department of Information Systems and Multimedia Design, Tokyo Denki University

[*2]東京大学 人工物工学研究センター
RACE (Research into Artifacts, Center for Engineering), The University of Tokyo

[*3]東京大学 情報基盤センター 図書館電子化研究部門
Information Technology Center, The University of Tokyo

A system for gathering and analyzing social bookmark (SBM) contents, and analysis results are described. Because so many people use SBM services, various contents containing valuable and not-valuable (spam) contents are registered in SBM services. The aim of this research is twofold: (1) to filter out spam contents, and (2) to find valuable contents. We created a system for collecting and analyzing SBM contents. The system assists users to find (1) tags attached to a content, (2) users who bookmarked the content, and (3) contents which a user bookmarked. An overview of the system, and preliminary analysis results using valuable contents and spam contents are described.

1. はじめに

今日、Web 上で情報を共有できるソーシャルブックマーク (Social Bookmark;SBM) というサービスが存在する。SBM には人々の興味や関心を集める有用なコンテンツが含まれ、SBM を用いた様々な研究が行われている [1, 2]。しかし、SBM には、多くの人々にとって不要なスパムも含まれている。

しかし、計算機にはどのコンテンツが有用で、どのコンテンツが不要であるかを判断することはできない。SBM のデータを収集し、分析することによって、人々にとって有用なコンテンツ、あるいは不要なコンテンツ（スパム）の特徴が分かれば、計算機を用いて自動的に有用なコンテンツ、不要なスパムを発見することができるようになるだろう。

本研究では、SBM のうちの一つである「はてなブックマーク [3]」を用い、ブックマーク情報を収集するためのシステムを開発する。

本論文の構成は以下のとおりである。第 2 節では、SBM の概要と、スパムブックマークのモデル化を行う。第 3 節では、構築したシステムについて述べる。第 4 節では、収集したデータの分析結果について述べる。最後に、本論文のまとめを行い、今後の課題について述べる。

2. ソーシャルブックマーク

2.1 ソーシャルブックマークとは

SBM とは、ネットワーク上にブックマークを保存するサービスである。既存のブラウザ上のブックマークはそのコンピュータでしか見ることができないが、SBM はインターネットにつながっていればどのコンピュータからでも閲覧できる。

SBM では、単純にネットワーク上にブックマークを保存するだけではなく、分類、人気度、コメントなどの情報が付加され、複数のユーザでブックマークを共有できることに主眼が置かれている。

SBM の分類の方法にディレクトリのような階層的な分類法ではなく、ユーザが自由に設定できる「タグ」と呼ばれる単語やフレーズを利用しているサービスが多い。ユーザは登録時にひとつのアドレスに複数の任意のタグをつけることができ、複数のユーザがひとつのアドレスに対しタグをつけることができる。これにより、自分と同じ関心を持つユーザのブックマークを閲覧したり、自分の思いつかなかった切り口の共通点により新しいサイトを発見できたりする。

SBM の代表的なものに del.icio.us[4]、はてなブックマークなどがある。

2.2 スパムブックマーク

今日、SBM において、スパムブックマークが問題となっている。たとえば、はてなブックマークにおいて、「無料」というタグで新着ブックマークを調べたところ、102 件中 12 件（11.8 %）がスパムサイトへのブックマークであった。このように、SBM においても、スパムの影響が増えつつある。

本来、ブックマークとは、自分が興味を持った再度訪れる可能性のある URL を登録するものである。しかし、このような通常のブックマークとは異なる、商用目的などの悪意を持ったブックマークが存在する。このようなブックマークをスパムブックマークと呼ぶ。

スパムブックマークは、利益を得るためにアフィリエイトサイトや通販サイトなどへ誘導することを目的としたブックマークであると考えられる。

本研究では、ブックマーク先が通販サイトなどのスパムサイトとは断定できないものであっても、誘導を目的としているブックマークであればスパムブックマークであると考える。

また、スパムブックマークを行っているユーザをスパムユーザ、スパムブックマークがされているサイトをスパムコンテンツと呼ぶ。

連絡先: 宗片健太朗，東京電機大学 情報メディア学科 計算言語学研究室, 東京都千代田区神田錦町 2-2, 03-5280-3332, munekata@cll.im.dendai.ac.jp

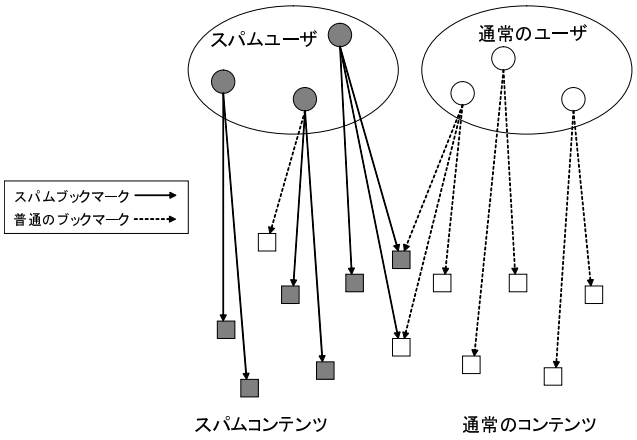

図 1: SBM におけるスパムブックマークの概念図

2.3 スパムブックマークのモデル

図 1 は、SBM におけるスパムブックマークの概念図である。この概念図では、(1) ユーザ、(2) ブックマーク（リンク）、(3) コンテンツの 3 つの要素を考える。スパムユーザは、スパムコンテンツに対して、スパムのブックマークを張る。これに対して、通常のユーザは、通常のコンテンツに対して通常のブックマークを張る。このとき、通常のユーザが、スパムコンテンツに対して通常のブックマークを張ることもある。

本研究の目的は、スパムブックマーク・スパムユーザを見つけ、スパムコンテンツをフィルタリングし、有用なコンテンツを取り出すことである。

以下、スパムブックマーク、スパムユーザ、スパムコンテンツの特徴について述べる。

2.3.1 スパムブックマークの特徴

すべてに共通しているものではないが、収集したスパムブックマークから、通常のブックマークとは異なる以下のような特徴が見られた。

1. はてなブックマークでは、ブックマークを公開するか非公開にするかを選択することができるが、スパムユーザは非公開にしていることが多い
2. 1 ユーザが、1 つの URL に対して 10 個程度のタグを付けてブックマークしている場合が多い
3. スパムブックマークがされている URL は、ブックマーク登録者がそのスパムユーザ 1 名のみである場合が多い
4. サイトの内容や商品を宣伝するコメントを付けている

2.3.2 スパムユーザの特徴

2.3.1 の 1 番で述べたとおり、スパムユーザのほとんどはブックマークを非公開にしていたが、公開しているユーザも存在する。その公開しているブックマークリストからは、以下のような特徴が見られた。

1. 同じドメインのサイトばかりが登録されている
2. 自動生成と思われるレイアウトがほぼ同じサイトばかりが登録されている
3. 短時間に多数のブックマーク登録を行っている
4. 「商品」「無料」など、特定のタグを常に使用している
5. ほかのユーザと比べて、使用しているタグの種類が多く、一貫性がない

2.3.3 スパムコンテンツの特徴

スパムユーザによってブックマーク登録されているコンテンツには、以下のような特徴が見られた。

1. 通販サイトの商品紹介ページ
2. 通販サイトへのリンクが並んでいるだけのサイト
3. 自動生成されたサイト
4. ほかの記事をコピーしたと見られるサイト

2.4 ソーシャルブックマーク空間を調査するシステムの要件

SBM 空間を調査するにあたって、以下の要件が必要である。

- URL、ユーザ、タグ、コメント、登録した時間のデータを収集し、これらの関係がわかるようにデータベースに格納する
- ある URL について、ブックマークしているユーザとそのユーザが付けているタグを表示する
- ある URL にブックマーク登録しているユーザの数を表示する
- ユーザが付けているコメントを表示する
- あるユーザがブックマークしている URL を表示する
- ブックマーク登録した時間を表示する
- あるユーザが使用しているタグの種類、使用頻度を表示する

3. ブックマークデータ収集システム

3.1 システム全体図

2.4 節であげた要件のうち、ブックマークデータの収集と解析に関する基本機能を実装した。本研究は、SBM のひとつである、はてなブックマークのデータを用いる。本研究で構築するデータ収集システムは、WEB 上から SBM の RSS（XML 形式のデータ）を取得し、そこからブックマーク情報を取り出してデータベースに格納するものであり、

(1)RSS 解析システム、(2) データ表示部

を開発する。（図 2 参照）

3.2 RSS 解析システム

はてなブックマークでは RSS が提供されている。この RSS には、ある URL をブックマークしているユーザの情報が含まれている。本システムは、RSS を解析し、そこに含まれている様々なデータをデータベースに格納する。

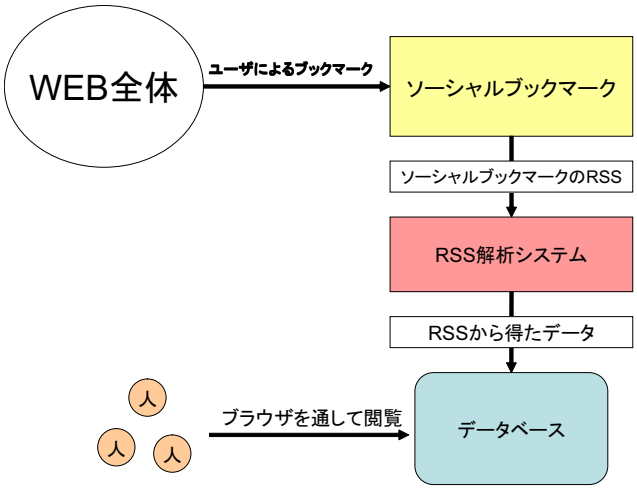

図 2: ブックマークデータ収集システムの全体図

http://b.hatena.ne.jp/entry/rss/http://blog.livedoor.jp/dankogai/archives/50983605.html

ユーザ名	タグ
toy2done	lifehacks 本
tomotiro-k	書評
hiroakiuno	lifehacks books
udy	lifehacks book
Alfa3733	
simonetta	書評
t_mori	lifehack Life 書評 考え方 仕事術
andoichi	book
no39929no	book 本 書評
h1romi	本
LukeSilvia	成長法 読み物
rararapocari	
muratamika	book 小飼弾
katsu-i	
julajp	
bash0C7	本 商品
memoclip	memoclip Life book review
ganseki	*あとで読む あとで読む 後で読む
picora	名言 dankogai
huh-gtr	dankogai

図 3: 収集したブックマークデータのブラウザへの表示例

3.3 データ表示

収集したデータを可視化するため、データベースの中身をブラウザ上に表示するプログラムを構築した。本システムでは、以下の内容をブラウザに表示することができる。

1. ある URL についてブックマークを行っているユーザのリスト
2. あるユーザがブックマークしている URL のリスト
3. あるタグが付けられている URL のリスト

図 3 は、上記 1 番の表示例である。図 3 において、ユーザ名をクリックすると上記 2 番、タグ名をクリックすると上記 3 番が表示される。

4. データと考察

4.1 収集したデータ

SBM から、URL とブックマークユーザ名だけでなく、タグ、コメントの情報も得られた。表 1 は本システムにて得たデータの総数である。また、ある URL に対する登録者数、登録タグ数、コメント数の平均はそれぞれ以下の通りである。

表 1: 収集したブックマークデータの量

URL 数	ユーザ数	タグ数	コメント数
88	5,254	1,711	1,665

- ユーザ数平均:132.0
- タグ数平均:42.7
- コメント数平均:21.6

4.2 有用なブックマーク登録データ

4.2.1 登録者数とタグ数の関係

登録者数と登録タグ数の関係を表したグラフを図 4 に示す。このグラフから、登録者数が増えるにしたがってタグ数も増加していることがわかる。

また、あるプログラマのインタビュー記事につけられたタグの一部を表 2 に示す。これをみると、関連性のない様々な言葉がひとつの URL につけられていることがわかる。また、同じ意味のタグでも、ユーザによって表現は異なっていることがある。

以上のことから、ユーザ一人一人が独自の観点で様々なタグをつけていることがわかる。

4.2.2 登録者数とコメント数の関係

登録者数とコメント数の関係を表したグラフを図 5 に示す。このグラフを見ると、タグ数とは違い、登録者が多ければコメントが多いとは限らないことがわかる。

また、あるプログラマのインタビュー記事につけられたコメントの一部を表 3 に示す。コメントの種類は、内容の要約や抜粋、見どころ、感想、自分の意見など様々である。意見を述べるにしても、インタビューの内容に関する意見であったり、インタビューそのものについての意見であったりと、人によってコメントの書き方に違いがあるのが分かる。

4.3 スパムブックマーク

スパムブックマークは、誘導することを目的としているため、一度にたくさんのタグを使用するなど、通常のブックマークとは異なる特徴が現れることがわかった。

表 4 は、はてなブックマークにおいて、タグ検索の結果の新着 25 件のデータである。(2008 年 4 月 14 日現在)

この表から、スパムコンテンツには、特定のタグが付けられている場合が多いことがわかる。

これらの特徴を利用して、スパムブックマークを自動検出し、処理することができると考えられる。

5. おわりに

本研究では、SBM の RSS を解析し、データを取り出すことで、SBM にはどのような情報が含まれ、またそのうちどの情報を取り出し利用することができるのかを知るために、RSS を解析するためのシステムを開発した。

SBM には、有用なコンテンツと不要なコンテンツが入り交ざっている状態である。本研究において、SBM のデータを分析し、有用なコンテンツと不要なコンテンツとに分類するための基礎となる部分を行った。

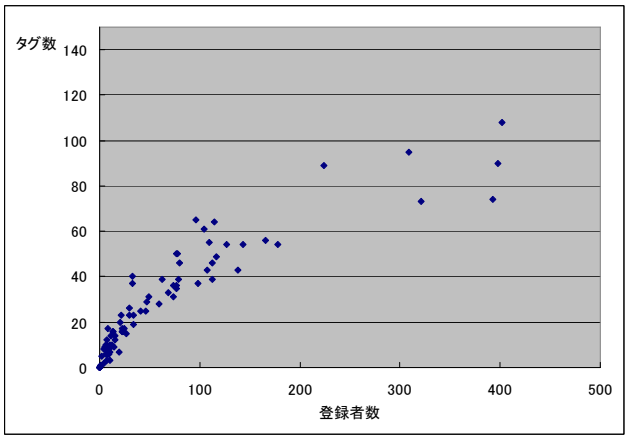

図 4: ブックマーク登録者数と登録タグ数の比較

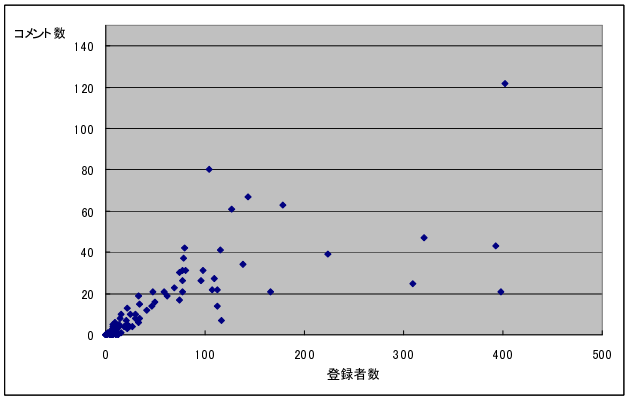

図 5: ブックマーク登録者数とコメント数の比較

表 2: プログラマのインタビュー記事につけられたタグの例

あとで読む	Life(人生, 生き方)
インタビュー (interview)	PHP
プログラミング (programming)	読み物
ひと (人物, 人,people)	PC
空気読まないおっさん	医
business	こけし
元気が出る言葉	達人

表 3: プログラマのインタビュー記事につけられたコメントの例

今ごろこの記事に気付いた orz 「べにぢょ」さん登場していましたね。
404 Blog Not Found でおなじみ dankogai 氏のインタビュー記事。
”Si vis pacem, para bellum——和を求めるなら戦に備えよ”
楽しく読み進められるインタビューだと思う。
楽しそうに生きている人間を見るとケチをつけたくなる
質問とか答えが面白かった。
「一生プログラマーでいれるかどうかは、言い換えれば年下から学べるか否か」
「オレはモテてる」　ナ、ナンダッテー！

表 4: タグごとの新着ブックマーク 25 件におけるスパム率

タグ名	スパム率
通販	95.8%
あとで読む	4.0%
アフィリエイト	84.0%
おすすめ	96.0%

スパムブックマークを行っているユーザは非公開ユーザが多いが、本収集システムでは、公開ユーザのデータしか取得することができない。今後の展望として、非公開のスパマーの情報をどのように取り扱うか、という問題が挙げられる。

参考文献

[1] 山家雄介, 中村聡史, アダム ヤトフト, 田中克己: ソーシャルブックマークの特性を利用した Web 検索のランキング精度の向上, 日本データベース学会 Letters Vol.6, No.1, pp.177-180

[2] 深見 嘉明, ソーシャルブックマークサービスにおけるアノテーション情報の機能分析, 第 21 回人工知能学会全国大会, 1G1-4 (2007).

[3] はてなブックマーク, http://b.hatena.ne.jp/

[4] del.icio.us, http://del.icio.us/