ソーシャルブックマークにおけるスパムの検出

宗片健太朗 † 福原知宏 †† 山田剛一 † 絹川博之 † 中川裕志 ††
† 東京電機大学 †† 東京大学

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登録されている。

しかし、SBM にはスパムコンテンツも含まれている。このスパムコンテンツをフィルタリングしなければ、SBM の有用性が損なわれる。本研究では、SBM サービスの一つである「はてなブックマーク [3]」を用い、スパムコンテンツのフィルタリングに向けた分析を行った。

本論文の構成は以下の通りである。2. ではソーシャルブックマークの概要について、3. ではブックマークデータの収集のためにシステムについて、4. ではスパムブックマークの分析及び考察、5. では本論文のまとめを述べる。

2 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、タグ、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 の代表的なものに delicious[4]、はてなブックマークなどがある。

3 ブックマークデータ収集システム

3.1 全体図

本研究では、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る。（図 1 参照）

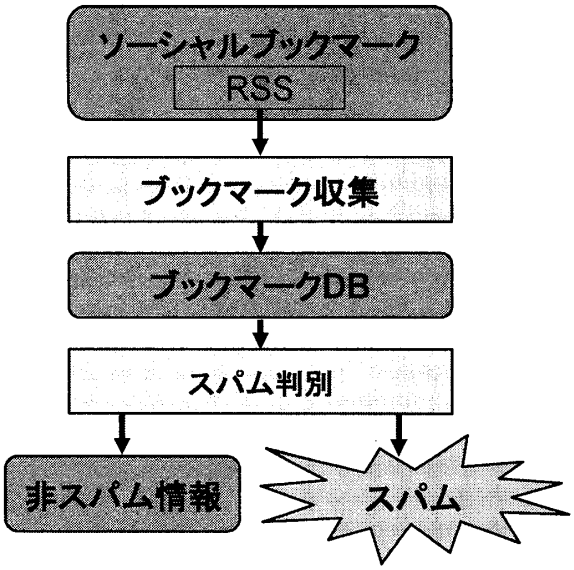

図 1: 本研究の全体図

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、あるひとつのページに対するブックマークの情報を表すもの、あるユーザのブックマーク情報を表すもの、あるタグの付けられたブックマークの情報を表すものなど様々である。

本システムは、あるページに対するブックマーク情報が得られる 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。RSS から取得できるデータは以下の通りである。

1. URL
2. ブックマークしているユーザ
3. 各ユーザが付けているタグ
4. 各ユーザが付けているコメント
5. 各ユーザがブックマークした日付、時刻

4 スパムブックマークの分析

4.1 スパムブックマーク

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。

Analysis of spam bookmarks in a social bookmark service
† Kentaro Munekata
†† Tomohiro Fukuhara
† Koichi Yamada
† Hiroshi Kinukawa
†† Hiroshi Nakagawa
Tokyo Denki University (†)
The University of Tokyo (††)

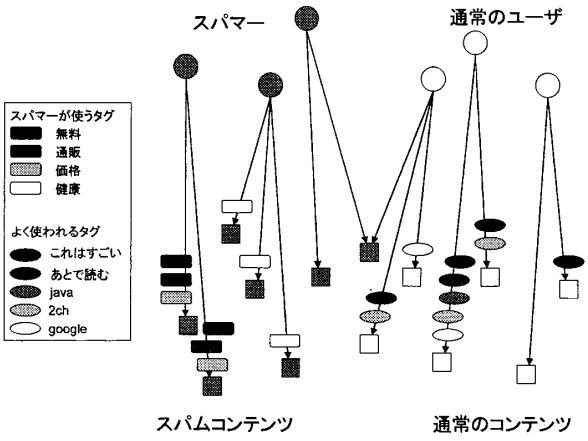

図 2: スパムブックマークの概念図

表 1: 収集したブックマークデータの量

URL 数	6,856
ユーザ数	39,707
タグ数	31,722
ブックマーク数	617,116
コメント数	149,212

このようなブックマークをスパムブックマークと呼ぶ。また、スパムブックマークを行っているユーザをスパマー、ユーザにとって価値の低いコンテンツをスパムコンテンツと呼び、スパムブックマークをされるほど、コンテンツのスパム度は高くなる。図 2 はスパムブックマークの概念図である。図 2 において，白抜きは非スパムユーザおよび非スパムコンテンツを，黒のものはスパマーおよびスパムコンテンツを表す。また、一般的によく使われているタグとスパマーがよく使うタグの一例を図 2 に示している。

4.2 収集したデータ

2008 年 10 月〜12 月中に収集したデータの量を表 1 に示す。また、このうち約 2000 ユーザを調査者 1 名により目視で調査し、64 人のスパマーを取り出した。

4.3 考察

表 2 は、スパマーの特徴として考えられるものをいくつか挙げ、それに当てはまったスパマーの数を表している。このうち、「ブックマーク者数が一人のブックマークの割合が多い」という特徴を持つユーザが 64 人中 60 人と非常に多かった。

次いで、「ブックマークしているページのドメインが同じ」という特徴を持つユーザが 32 人だった。1〜3 種類程度のドメインに対してのみブックマークしており、ページの内容やレイアウトなどが類似しているものが多かった。

表 2: 特徴に当てはまったスパマーの数

特徴	ユーザ数 (%)
ブックマーク者数が一人の ブックマークの割合が多い	60(93.8%)
ブックマークしているページの ドメインが同じ	32(50.0%)
それぞれのブックマークに付けている タグの数がほぼ一定	19(29.7%)
複数のページのタイトルの 先頭または語尾が一致	13(20.3%)
ブックマーク数に対して、 使用しているタグ数が多い	11(17.2%)

また、表 2 にはないが、コメントがそのページ内からのコピーである、コメントとページタイトルが一致している、といったコメントについての特徴も少数ではあるが見られた。

以上のような特徴を用いることで、スパマーと非スパマーを分類し、スパムコンテンツにフィルターをかけることができると考えられる。

5 おわりに

SBM の Web コンテンツの共有機能という利点を保持するために、SBM のデータを収集し、スパマーやスパムコンテンツをフィルタリングするための分析を行った。

4.3 で述べたとおり、ほとんどのスパマーのブックマークのブックマーク者数は一人であったが、現在の収集システムではユーザのブックマークをたどって集めているため、スパマーがあまり集まっていない。収集方法を見直し、スパマーのデータが増えれば、よりよい分析結果が得られると考えられる。

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(2007)

[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] delicious, http://delicious.com/