D-015

ソーシャルブックマークにおけるスパム検出のための特徴とその評価

Analysis and Evaluation of Features for Detection of Spam in Social Bookmark

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[*2]
Kentaro Munekata, Tomohiro Fukuhara, Koichi Yamada, Hiroshi Kinukawa, Hiroshi Nakagawa

1. はじめに

今日、Web 上でブックマーク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。また、[3], [4]などの SBM のデータを用いた研究も行われている。

しかし、SBM にはスパムが存在し、これに対処しなければコンテンツの有用性が損なわれてしまう。

本研究では、SBM のうちの一つである「はてなブックマーク[1]」のデータ用い、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.では研究の概要について、3.ではスパムブックマークの定義や特徴について、4.ではスパムブックマークの判別とその考察、5.では本論文のまとめを述べる。

2. 研究概要

2.1 全体図

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る。収集システムにより収集したデータに対して機械学習を用いて、スパマー、非スパマーの判別を行う（図 1）。

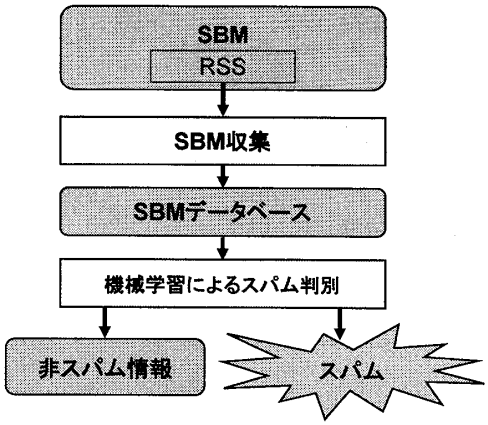

図 1.　研究の全体図

```
<item
rdf:about="http://b.hatena.ne.jp/ddaysogre/20090701
#bookmark-14308423">
   <title>ddaysogre</title>

<link>http://b.hatena.ne.jp/ddaysogre/20090701#book
mark-14308423</link>
   <description>集客実験。よい記事。</description>
   <dc:date>2009-07-01T20:29:45+09:00</dc:date>
   <dc:subject>マーケティング</dc:subject>
   <dc:subject>はてな</dc:subject>
</item>
```

図 2.　はてなブックマークの RSS の例

2.2 収集データ

はてなブックマークの RSS から、以下の情報を収集している。

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

図 2 は RSS の一部の具体例である。ここには<title>要素にユーザ名、<description>要素にコメント、<dc:date>要素にブックマーク日時、<dc:subject>要素に付与されたタグの情報が含まれている。

表 1. 実験に用いたデータ

	訓練データ	テストデータ
ユーザ総数	1000	66
スパマー数	87	17

3. スパムブックマーク

3.1 スパムブックマークとは

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。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、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。

*1 東京電機大学 大学院　Tokyo Denki University
*2 東京大学　The University of Tokyo

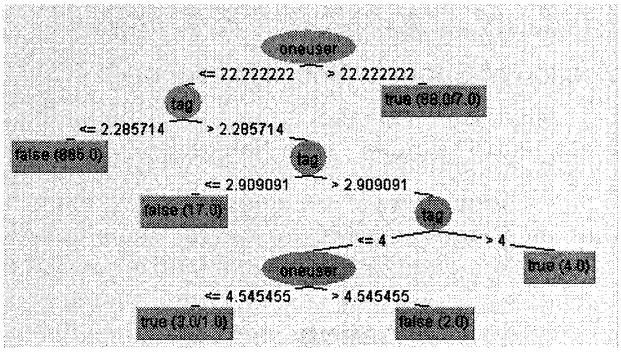

図 3. スパム判別のための決定木

3.2 スパマーの特徴

スパマーには、ブックマーク行動に通常のユーザとは異なる点が見られることがある。スパマーの特徴として考えられるものを以下に挙げる。

特徴 1: スパマーは単独で行っていることが多く、非スパマーのユーザはスパムコンテンツをブックマークしないため、スパマーのブックマークはブックマーク者数が 1 人であるブックマークの割合が多くなる。

特徴 2: スパマーは、検索にかかりやすくするために、ひとつのブックマークにたくさんのタグを付けている場合が多く、ブックマーク数に対する使用タグ数の割合が非スパマーに比べて多い。

特徴 3: スパマーは自分の作成したページをブックマークするため、ブックマークしているページのドメインが一種類、または数種類程度しかない。

特徴 4: それぞれのブックマークに付けられているタグの数が一定である。プログラムにより自動的にブックマークさせていると考えられる。

特徴 5: 多数のブックマークのタイトルに「○○ブログ」など、一致する部分がある。

特徴 6: コメントが、ブックマークしているページのタイトルや本文からの引用である。

4. スパムブックマークの判別

4.1 スパムブックマークの機械学習

本研究では、3.2 で挙げた特徴を用い、ユーザがスパマーであるか否かを判別することにより、スパムを検出することを目的とする。

3.2 で挙げた特徴のうち、特徴 1 と特徴 2 の二つについての分析を行った。それぞれの特徴について以下のような式を用いた。

特徴 1:

$$oneuser = \frac{\text{このユーザ固有のブックマーク数}}{\text{ブックマーク数}}$$

表 2. テストデータの適用結果

	精度	再現率	F 値
スパマー	0.607	1	0.756
非スパマー	1	0.776	0.874

特徴 2:

$$tag = \frac{\text{タグの異なり数}}{\text{ブックマーク数}}$$

調査した 1000 ユーザを基に、機械学習を行った。学習ツールには Weka を用い、学習アルゴリズムには決定木を用いた。図 3 は、Weka により生成された決定木である。図 3 において、true がスパム、false が非スパムであり、oneuser が特徴 1、tag が特徴 2 をあらわしている。

4.2 学習に基づく判別

訓練データには、2008 年 10 月~12 月中に収集したデータのうちの 1000 ユーザ分、テストデータには 2009 年 6 月に収集したユーザのデータを用いた(表 1)。

訓練データにより生成した決定木をテストデータに適用した結果を表 2 に示す。精度、再現率ともにおよそ 7 割となった。

4.3 考察

今回の実験では、スパマー検出の精度が特に低かった。理由として考えられるのは、ブックマーク数の少ない新規ユーザの存在である。

ある程度の数のブックマークを行っているユーザであれば、そのユーザの特徴を把握することは容易であるが、ブックマーク数が少ないユーザの場合、特徴を正しく把握することが困難である。テストデータにおけるユーザは新着ブックマークから集めており、まだブックマーク数の少ないユーザも含まれていたため、そういったユーザが誤って判定されたと考えられる。

5. おわりに

SBM の有用性を保持するため、スパマーの特徴を分析し、スパマー判別のための学習を行った。

今回の実験では、ある程度の効果は出たが、ブックマーク数の少ないユーザでは正しい結果が得られないということが分かった。

今後の展望として、上記の問題への対処と、3.2 で挙げた特徴のうち、今回用いなかった特徴 3~6 の適用が挙げられる。

参考文献

[1]はてなブックマーク, http://b.hatena.ne.jp
[2]Robert Wetzker, Carsten Zimmermann, Christian Bauckhage, "Analyzing Social Bookmarking Systems: A del.icio.us Cookbook", ECAI 2008 Mining Social Data Workshop , pp. 26-30 (2008).
[3]Pierpaolo Basile, Domenico Gendarmi, Filippo Lanubile, Giovanni Semeraro."Recommending Smart Tags in a Social Bookmarking System".Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007), pp.22-29(2007).
[4]Miranda Grahl, Andreas Hotho, Gerd Stumme."Conceptual Clustering of Social Bookmarking Sites".Machine Learning, Vol. 56, No. 1 – 3, pp.115-151 (2004).