Detecting Spammers on Twitter

Fabrício Benevenuto
(joint work with Gabriel Magno, Tiago Rodrigues, and Virgílio Almeida)

Federal University of Minas Gerais (UFMG) - Brazil
New forms of Spam

- Online social networks
  - Very popular
  - Easy to create content

- New forms of spam
  - Unsolicited invitations on Facebook
  - Vandalism on Wikipedia
  - Video spam on YouTube [SIGIR’09]
  - Spam on Twitter
Spam on Twitter
Spam on Twitter

notorrious: i wish #worldcup games came on at night...not at 7am.
less than 20 seconds ago via Twitter for iPhone · Reply · View Tweet

aplusk: Man, I didn't expect Germany to look this good #worldcup
about 3 hours ago via Brizzly · Reply · View Tweet

about 3 hours ago via Twitter4J · Reply · View Tweet

SPAM

Users post URLs unrelated to content
Negative impact of spam

- Pollute real time search
- Interfere on mining tools and statistics about trendings and events on Twitter
- Consume user and system resources
- Waste human attention
Goal and methodology

- **Goal:** Detect spammers on Twitter

- **4-step approach**
  1. Collect data from Twitter
  2. Manually create a collection of users labeled as spammers or non-spammers
  3. Identify attributes able to distinguish spammers from non-spammers
  4. Classification approach to detect spammers
Part 1. Motivation & Problem

Part 2. 4-step approach

Part 3. Experimental results
Step 1. Collecting Twitter

- Crawls subject to rate-limiting
  - Twitter provide us a white list for 58 machines at MPI-SWS
  - Rate limit = 20,000 requests/hour.

- Inspected all user IDs collecting
  - User profile information
  - List of followers and followees
  - All tweets they posted

- In total we collected 54,981,152 users, 1,963,263,821 unique links, and 1,755,925,520 tweets
Step 2. Labeled collection

- Desired properties

1) Have a significant number of spammers and non-spammers

2) Include spammers who are aggressive in their strategies

3) Choose users randomly and not based on their characteristics
Step 2. Labeled Collection

Focus on popular events of 2009

#musicmonday

Michael Jackson

Susan Boyle
Step2. Labeled Collection

• Volunteers analyze tweets of randomly selected users that post to the three trending topics analyzed
  – Development of a Web system to ease the process
  – Each user is analyzed by at least two volunteers
  – Agreement in 99% of cases

• 8,207 users were analyzed out of which 355 are spammers

Labeled collection:
355 spammers + 710 non-spammers = 1,065 users
Step 3. Attributes

- **User behavior** (total = 23)
  - number of followers
  - number of followees
  - number of tweets
  - age of the user account
  - etc.

- **Tweet content** (total = 39)
  - Fraction of tweets with spam words
  - Fraction of tweets with URL
  - etc.
Importance of the attributes

<table>
<thead>
<tr>
<th>Position</th>
<th>$\chi^2$ ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fraction of tweets with URLs</td>
</tr>
<tr>
<td>2</td>
<td>age of the user account</td>
</tr>
<tr>
<td>3</td>
<td>average number of URLs per tweet</td>
</tr>
<tr>
<td>4</td>
<td>fraction of followers per followees</td>
</tr>
<tr>
<td>5</td>
<td>fraction of tweets the user had replied</td>
</tr>
<tr>
<td>6</td>
<td>number of tweets the user replied</td>
</tr>
<tr>
<td>7</td>
<td>number of tweets the user receive a reply</td>
</tr>
<tr>
<td>8</td>
<td>number of followees</td>
</tr>
<tr>
<td>9</td>
<td>number of followers</td>
</tr>
<tr>
<td>10</td>
<td>average number of hashtags per tweet</td>
</tr>
</tbody>
</table>

Attributes from content and user behavior are both important
Distinguishing classes of users (1)

Spammers post most of the tweets containing URLs.
Distinguishing classes of users (2)

Spammers have less followers than followees
Step 4. Classification approach

- SVM (Support vector machine) as classifier
- Use all attributes
- 5-fold cross validation
Part 1. Motivation & Problem

Part 2. 4-step approach

Part 3. Experimental results
Classification results

88% of accuracy

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spammer</td>
<td>Non-spammers</td>
</tr>
<tr>
<td>True Spammer</td>
<td>70.1%</td>
<td>29.9%</td>
</tr>
<tr>
<td>True Non-spammer</td>
<td>3.6%</td>
<td>96.4%</td>
</tr>
</tbody>
</table>

- **J = 0.1**: correctly classify 44% spammers, misclassifying <0.3% non-spammers
- **J = 3**: correctly classify 81% spammers, paying the cost of misclassifying 18% non-spammers
Reducing the attribute set

Different subsets of features can obtain competitive results
Detecting tweets instead of users

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th>84.5% of accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spam</td>
<td>Non-spam</td>
</tr>
<tr>
<td>True</td>
<td>78.5%</td>
<td>21.5%</td>
</tr>
<tr>
<td>Non-spam</td>
<td>92.5%</td>
<td>7.5%</td>
</tr>
</tbody>
</table>

Good results, but tweet content is easy to be faked

We can still obtain good results classifying users even if we disregard content attributes

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spammer</td>
</tr>
<tr>
<td>True</td>
<td>69.7%</td>
</tr>
<tr>
<td>Non-spam</td>
<td>4.3%</td>
</tr>
</tbody>
</table>
Conclusions

• We propose a mechanism to detect spammers on Twitter
  – Twitter dataset and labeled collection
    • Publicly available (soon) at www.dcc.ufmg.br/~fabricio
  – Attribute identification
  – Classification approach
    • Correctly identified majority of spammers
    • Different subsets of features can obtain competitive results
    • Detection of spam also works, but attributes are easier to be faked
Questions?

fabricio@dcc.ufmg.br

http://www.dcc.ufmg.br/~fabricio