Mining User Comment Activity for Detecting Forum Spammers in YouTube

Ashish Sureka
[ashish@iiitd.ac.in]
Indraprastha Institute of Information Technology, Delhi (IIIT-D), India

RESEARCH MOTIVATION

- *Comment spam* in *online discussion forums* is prevalent
- Several *content-based* methods have been proposed to automatically identify spam comments
- Content-based methods
  - Analyze the text of the post or message
  - Such as checking the presence of pre-defined terms or links
  - Shown encouraging results: not perfect: need to augment or complement their capabilities
- *Research Gap*
  - Relatively unexplored: analyzing *commenting behavior/activity* of a user to identify spammers
- *Hypothesis*/Belief
  - Examining the commenting activity (*usage analysis and characterization*) of a user can play a role in identifying *spammers*
RESEARCH AIM

➤ Broad objective

➤ To investigate the application of usage-based features
➤ Derived from a user's comment activity (log of recent comments)
➤ To identify comment spammers

➤ Specific objective

➤ Mine usage-based discriminatory patterns and markers
➤ To identify comment spammers in YouTube forums
➤ YouTube: a very popular and largest video sharing website on Internet
RESEARCH CONTRIBUTIONS

Literature Survey

- Heymann et al. present a survey of approaches for fighting spam on social websites [9]
- Hayati et al. present an evaluation and analysis of Web 2.0 anti-spam methods [7]
- Benevenuto et al. provide an overview of pollution in video sharing systems such as YouTube [1]
- Yo-Sub Han et al. present an algorithm to evaluate the reputation of a user in YouTube [6]
- Benevenuto et al. introduce a technique to detect video spammers in YouTube [2]

Unique contributions

- First study (on YouTube) of mining the recent activity log of a user
- To extract usage-based features
  - Prevalence of high comment repeatability
  - Presence of exactly same comment across videos
  - Presence of ultra low time difference between comments
  - Presence of a large number of spam tags by the community or moderators
- To identify forum/comment spammers
- An empirical study on dataset crawled from YouTube
- Demonstrates that the proposed usage-based and behavioral features are reliable/effective
- Fresh perspective /insights on the characteristics and properties of comment spammers on YouTube
HIGH-LEVEL SOLUTION FRAMEWORK/METHODOLOGY

ATDC: Average Time Difference between Comments
PCHF: Percentage of comments having hasSpamHint flag
CRAV: Comment repeatability across videos
CRR: Comment repetition and redundancy
The tagging of hasSpamHint is performed at the comment-level and not at the user-level
Many (significant percentage) spam comments are not tagged as hasSpamHint
Retrieve comments marked with hasSpamHint for a given video
Extract userids behind the spam comments
Retrieve recent commenting activity (log of comments and the associated metadata) of a given user

FILTER USERS BASED ON hasSpamHint
FEATURE EXTRACTION

ATDC: Average Time Difference between Comments

CRAV: Comment Repeatability Across Videos

CRR: Comment Repetition and Redundancy

PCHF: Percentage of Comments with hasSpamHint Flag
EMPIRICAL ANALYSIS, FINDINGS

- Users A, B, C and D have posted more than 30 comments (A 35, B 118, C 36 and D 30)
- Have a CRR value of more than 0.7 (which means posting same comment multiple times)
- Have 80% of the comments marked as spam by the moderator
- Users on top right corner are potential spammers
- Manual inspection of such users and confirms the hypothesis to be true

Plot of users in the evaluation dataset across two dimensions: spam percentage and comment repetition and redundancy (CRR)
Users having *high comment overlap* and *low video overlap*
Means several similar comment posting but in a single or small set of videos

Users having *high comment overlap* and *high video overlap*
A phenomenon wherein a user posts exactly same comments across multiple videos
spam percentage and number of comments

log of average time difference between comments (ATDC) and number of comments
An illustrative list of comments of some of the users identified as spammers in the experimental dataset

<table>
<thead>
<tr>
<th>Watch?v=JNVeTR9MhAo?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check out my channel</td>
</tr>
<tr>
<td>Please watch my vids at airborne2048</td>
</tr>
<tr>
<td>CAT almost BIT ME A FINGER OFF)))))) view on my channel )))) password please !!!!!!!!!!!!!!!!!!!!!!!!!!!!!1</td>
</tr>
<tr>
<td>Watch?v=DGHC-AgB8Us?</td>
</tr>
<tr>
<td>TV4500Channels.blogspot.com</td>
</tr>
<tr>
<td>CHECK OUT MY VIDS AND COMMENT</td>
</tr>
</tbody>
</table>

**SPAMMER HEURISTIC/RULE**

\[
\text{SPAMMER} = (\text{PCHF} > 70) \text{ OR } (\text{ATDC} < 150) \text{ OR } (\text{COMOVP} > 0.60) \text{ OR } (\text{VDOVP} > 0.60)
\]
Thank You