User Comment Analysis for Android Apps and CSPI Detection with Comment Expansion

Lei Cen, Luo Si, Ninghui Li
Department of Computer Science
Purdue University

Hongxia Jin
Samsung Information Systems America
Outline

1. Background

2. Data Collection & Annotation

3. CSPI Detection

4. Experiment

5. Conclusion
Background

• Security/Privacy issues of mobile apps

• Values of user comments in revealing the issues
Background

Task: Detect CSPI
Comments with Security/Privacy Issues (CSPI)

Reviews

User reviews

Alex Pantoja  July 3, 2014

Wow! So cool sad when lee died but still I had a fun time! The swipes are very hard to master cause it’s dark red and I can’t see it sometimes. But still good game :D

Laura Duda  July 3, 2014

Absolutely wonderful! It takes a long time to load and it stutters occasionally otherwise I would rate five stars. But it is completely fun and enjoyable to play. It’s probably the best game (including season 1 game) that I’ve played yet on my tablet.

Greg Perez  July 3, 2014

The walking dead season 2 I am going to have so much fun choosing the bad choices.

Kadir Yeles  July 2, 2014

When will episode 4&5 come out in AUS? Really want to know when it’ll come out! Can’t wait for it.

Patrick Benjamin  July 2, 2014

Good game. BTW good job and I know we’ve been bothering u tell tale for so long about making season two for android Google play store but now there’s two more things I ask of you plz make the wolf among us come out for android and make the episodes come out faster on twds2 game.

Nicole Bayardo  July 3, 2014

Awesome! I can’t finish the first season of the walking dead cause it costs money u should make the other episodes on the second part of the walking dead free these games are awesome and I want to play more but I can’t cause the money! I recommend this game to mature audiences after that have fun and kill zombies and decide what to do and when to do it I plan on being on the actual show to sign up to be a zombie and maybe depending on the episodes I want to possibly be on the walking dead.

Stacy Sager  July 3, 2014

Maybe the best way to play TWD Performance on my older Galaxy S3 causes me to fail many action sequences I would never fail on PC or with a controller. But on a higher end device the experience rivals that of the PC version’s. The time for inputting commands on a touch screen should maybe be longer than what it takes to press a button. When performance isn’t a problem I still seem to fudge things up more than I do on other versions off the game. Can’t wait for episode 4 and 5.

Scott Pollan  July 3, 2014

Awesome! Love this franchise! Graphics & game play couldnt be any better! Keep up the great work! Must download!

Wanda Castejano  July 3, 2014

Game is wonderful! When lee died is kinda sad but I have a good time Now do episode 3.
Data Collection

- From Google Play
- 6,938 apps
- 5,108,538 comments
Data Annotation

General

Scenario

When?

Nature

What?

Before Installation

Executing

After Uninstall

Foreground

Background

System

Privacy

Spam

Finance

Others
Data Annotation: More Detail

• Done by two people.
• Labeled 36,464 comments.

<table>
<thead>
<tr>
<th>Nature</th>
<th>Label</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>System</td>
<td>Issues causing negative effect to the system</td>
</tr>
<tr>
<td></td>
<td>Privacy</td>
<td>Issues about getting unauthorized access to user info.</td>
</tr>
<tr>
<td></td>
<td>Spam</td>
<td>Issues about unpleasant ads and related.</td>
</tr>
<tr>
<td></td>
<td>Finance</td>
<td>Issues about suspect money stealing.</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>Security / privacy issues not included above.</td>
</tr>
<tr>
<td>Scenario</td>
<td>Before</td>
<td>Issues occur before installation of the app.</td>
</tr>
<tr>
<td></td>
<td>Execution</td>
<td>Issues occur when the app is executing on the phone.</td>
</tr>
<tr>
<td></td>
<td>Foreground</td>
<td>Issues occur when the app takes the foreground screen.</td>
</tr>
<tr>
<td></td>
<td>Background</td>
<td>Issues occur when the app is running background.</td>
</tr>
<tr>
<td></td>
<td>After</td>
<td>Issues occur after uninstall of the app.</td>
</tr>
</tbody>
</table>
Data Annotation: Observation

• Short in length
  • “spam”

• Miss-spelling
  • “I bought $2.00 in *** and I never got them. I’m really mad about it.”

• Made-up words
  • “spamspamspam”
CSPI Detection: The Big Picture

- Comments
- CSPI w. Labels
- Post-Process
- Base Detection
- Filtering
- Feature Extraction
- Comment Expansion
CSPI Detection: Filtering

• Key word based
• Key words by semi-auto selection
  • security,
  • privacy,
  • permission,
  • money,
  • spam,
  • steal,
  • phish,
  • etc.
• Only select comments with low ratings (<4)
CSPI Detection: Features

• Pre-process
  • Stemming
  • Remove stop words

• Bag of word feature
  • Remove most popular (>1,000,000) and least popular words (<100)
  • 13,135d
CSPI Detection: Comment Expansion

• Borrowed from Pseudo relevant feedback for query expansion

\[ f_{new} = (1 - \alpha) f_{old} + \alpha \cdot \frac{1}{|R|} \sum_{f \in R} f \]

• Motivation
  • Utilize similar comments to get less noisy comment feature.
CSPI Detection: Comment Expansion (Cont’d)

- Considerations
  - Define “relevant” comments
    - TF-IDF + cosine similarity
    - Indexed by Lemur engine
  - Scope:
    - Same app?
    - Same category?
    - Whole set?
  - Time Constraint
  - Size of “relevant” comments

Size
“Relevant” comments

All
Category
App

Time line
CSPI Detection: 
Base method

- Independent Logistic Regression (ILR)

\[
\min_{w_j, b_j} NLL(X, w_j, b_j) + \lambda(|w_j|^2 + |b_j|^2)
\]

\[
NLL(X, w_j, b_j) = \sum_{i=1}^{N} \ln(1 + \exp^{-y_i^j(w_j^T x_i + b_j)}), j = 1, \ldots, 11
\]

- Train one LR classifier for each label

- Known as Binary Relevant (BR) in multi-label learning
CSPI Detection: Post-Process

• Motivation
  • Make use of label correlation

• Process
  • Another round of ILR
  • User the output of the first round as input
CSPI Detection: Post Process (Cont’d)
Experiments: Settings

• 50/50 train/test setting

• Independent Logistic Regression as baseline

• Model selection using 5-fold cross validation (scope, expansion size, regularization parameters)
Experiments:

General Comparison

$CDCE^- : \text{ILR + comment expansion}$

$CDCE^+ : \text{ILR + comment expansion + post-process}$

$CDCE^* : \text{ILR + comment expansion + selective post-process}$

<table>
<thead>
<tr>
<th>method</th>
<th>General</th>
<th>Scenario</th>
<th>Nature</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILR</td>
<td>0.7962</td>
<td>0.6713</td>
<td>0.7032</td>
<td>0.7153</td>
</tr>
<tr>
<td>CDCE^-</td>
<td>0.8037†</td>
<td>0.6740†</td>
<td>0.7159†</td>
<td>0.7223†</td>
</tr>
<tr>
<td>CDCE^+</td>
<td>0.8004†</td>
<td>0.6814†</td>
<td>0.7096†</td>
<td>0.7225†</td>
</tr>
<tr>
<td>CDCE^*</td>
<td>0.8037†</td>
<td>0.6836†</td>
<td>0.7159†</td>
<td>0.7263†</td>
</tr>
</tbody>
</table>
Experiments:
Label level comparison

<table>
<thead>
<tr>
<th>Label</th>
<th>F1</th>
<th>P.P.</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ILR</td>
<td>CDCE*</td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>0.7962</td>
<td>0.8037†</td>
<td>No</td>
</tr>
<tr>
<td>Before</td>
<td>0.6965</td>
<td>0.7012</td>
<td>Yes</td>
</tr>
<tr>
<td>Execution</td>
<td>0.7277</td>
<td>0.7358†</td>
<td>Yes</td>
</tr>
<tr>
<td>Foreground</td>
<td>0.4347</td>
<td>0.4689†</td>
<td>Yes</td>
</tr>
<tr>
<td>Background</td>
<td>0.6674</td>
<td>0.6750†</td>
<td>No</td>
</tr>
<tr>
<td>After</td>
<td>0.1327</td>
<td>0.0882</td>
<td>No</td>
</tr>
<tr>
<td>System</td>
<td>0.3991</td>
<td>0.4012</td>
<td>No</td>
</tr>
<tr>
<td>Privacy</td>
<td>0.7264</td>
<td>0.7350†</td>
<td>No</td>
</tr>
<tr>
<td>Spam</td>
<td>0.8181</td>
<td>0.8304†</td>
<td>No</td>
</tr>
<tr>
<td>Finance</td>
<td>0.5238</td>
<td>0.5320†</td>
<td>No</td>
</tr>
<tr>
<td>Others</td>
<td>0.0235</td>
<td>0.0384</td>
<td>No</td>
</tr>
</tbody>
</table>
Conclusions

• An analysis on CSPI on data from Google Play, with a two-dimensional label system describing “what” and “when” of the CSPI.
• Propose to use comment expansion for improving the comment feature for CSPI detection.
• Provide a foundation for further exploration for evaluation of mobile app security/privacy risk.
Q & A