Automatic User Comment Detection in Flat Internet Fora

Mathias Bank
mathias.bank@uni-ulm.de

Faculty for Mathematics and Economics
University of Ulm
Community Analysis

Trust in Advertisement

- customer recommendation
- news Paper
- customer opinion
- product website
- TV
- magazines
- sponsoring
- online banner

Nielsen Consumer Report October 2007
User Content Detection
User Content Detection (1)

Davros
New Member

Adsense Integration

I recently bought vBulletin and downloaded the version with my Adsense integrated, but since doing this I have now been banned by Google for invalid click activity.

I have never and will never click ads that I publish on my web sites.

However I have been a part of the Adsense program since 2006 and have never had a problem. I run 5 websites only my latest site is vBulletin and I have made a reasonable amount of money from Adsense in the past so to say I am unhappy at losing the $300 march payment is an understatement. I have appealed against the decision but Google have said they have checked my account and the decision will stand.

Not only have they blocked my Adsense account but they have stopped Google Bots crawling my pages, or at least I haven't seen any since this has happened.

I'm just wondering if anybody else has had the same issues?

I've now signed up with AdBright, but can't help but think this is somehow related to vBulletin Adsense Integration.

Digicom
New Member

This also happened to me. I had been with Adsense since 2005 and in February this year I download Adsense integrated in vBulletin, then when I was due to be paid in march I was suddenly banned from adsense, I then appealed but was told

Quote

However, after thoroughly reviewing your account data and taking your feedback into consideration, we have re-confirmed we are unable to reinstate your account.

Since this happened Google Bots have also stopped crawling my site, it seems strange that my problem is identical to the problem Davros has and only happened after downloading Adsense integrated in vBulletin.

Nick
Senior Member
User Content Detection (2)
Main Content Detection - SST

Template based Systems

Style Trees
Main Content Detection – SST (1)

Node importance:
- number of presentation styles
- Number of content diversity
- Based on Shannon’s Entropy

Composite Importance
- Propagating child importance to its parent nodes

Noise detection by searching nodes with low Composite Importance
Main Content detection by searching nodes with high Composite Importance
Main Assumption: Node with Main Content Section will get the highest importance value.

/root/body/div[@class="content"][contains(@id,"thread")]/
Main Content Detection – Evaluation

Evaluation data:
- 51 real internet fora
- Based on 14 different fora systems
- Different template structures (not only color)
- Not used except for evaluation

Evaluation results:
- perfect: 82.4%
- Correct: 100%

For 17.6% (9 fora) the extracted XPath expression could be more specific.
User Content Detection (3)
Post Segmentation – Related Work

B. Liu et al:

- Data regions are presented in continuous regions with similar tags
- Data regions can be found in one subtree

⇒ Data list by searching „identical“ subtrees

- Post-Order traversing for nested structures

Detecting „identical“ subtrees in community systems very difficult!
# Post Segmentation - Observations

Analyzing
- 3,500 posts
- 13 different fora
- 8 different fora systems

<table>
<thead>
<tr>
<th>All Postings</th>
<th>No line breaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 320 characters</td>
<td>• 25% of all posts</td>
</tr>
<tr>
<td>• 4.3 line breaks</td>
<td>• 94 characters</td>
</tr>
</tbody>
</table>

Observations
1. User comments mainly consist of textual data
2. Complete comment in one subtree (Liu), splitted by further elements
3. No similarity between user comments except base structure
4. Nested structures are possible (e.g. Drupal)
Post Segmentation – Algorithm

Post Candidate Search

Wrapper Generation
Post Segmentation – Algorithm (1)

Post Candidate Search

Wrapper Generation

Merge inline elements

Ich habe die Suche bemüht und bin dabei auf zwei Threads gestoßen, die aber nicht genau mein Anliegen behandeln.

Mir geht's darum, die Elemente `<blockquote>`, `<q>` und `<cite>` richtig zu verwenden (mehr gibt's doch nicht für Zitate/Zitationen, oder?).

Merge inline elements & p, blockquote, pre

TEXT
Post Segmentation – Algorithm (2)

Merge inline elements
Text node identification

3 rules:
• > 150 characters
• > 50 characters & new lines
• Not beginning with ”___“

Warning:
• Not necessarily complete!
• Not necessarily correct!
Post Segmentation – Algorithm (3)

3 rules:
• > 150 characters
• > 50 characters & new lines
• Not beginning with “___“

Wrapper Generation

Ich bin sicher nicht auf dem Niveau, um das alles wirklich zu verstehen, auch, wenn ich schon einige Erfahrungen habe.

Dennoch erschließt sich mir der Sinn des Ganzen nicht so recht. Was soll dieses Standardmodul für Vorteile bedeuten?

```html
  <div class="mod">
    <div class="hd">Block Head</div>
    <div class="bd">Block Body</div>
    <div class="ft">Block Foot</div>
  </div>
```

Merge inline elements

Text node identification
Post Segmentation – Algorithm (4)

- Post Candidate Search
  - Merge inline elements
  - Text node identification
  - Text node cleaning
- Wrapper Generation

Ich bin sicher nicht auf dem Niveau, um das alles wirklich zu verstehen, auch, wenn ich schon einige Erfahrungen habe.

Dennoch erschließt sich mir der Sinn des Ganzen nicht so recht. Was soll dieses Standardmodul für Vorteile bieten?

```
<div class="mod">
  <div class="inner">
    <div class="hd">Block Head</div>
    <div class="bd">Block Body</div>
    <div class="ft">Block Foot</div>
  </div>
</div>
```

Complete comment in one subtree
Post Segmentation – Algorithm (5)

Post Candidate Search

- Merge inline elements
- Text node identification
- Text node cleaning
- Node generalization

Wrapper Generation
Post Segmentation – Algorithm (6)
Post Segmentation – Evaluation

Evalutation data:
Same data as for Main Content Detection
Main Content already selected (all fora used!)

Evaluation results:
- perfect: 64%
- Correct: 90%

<table>
<thead>
<tr>
<th>system</th>
<th>correct</th>
<th>perfect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burning Board</td>
<td>100,0%</td>
<td>66,7%</td>
</tr>
<tr>
<td>drupal</td>
<td>83,3%</td>
<td>0,0%</td>
</tr>
<tr>
<td>IPB</td>
<td>75,0%</td>
<td>50,0%</td>
</tr>
<tr>
<td>myBB</td>
<td>66,7%</td>
<td>66,7%</td>
</tr>
<tr>
<td>phpBB</td>
<td>100,0%</td>
<td>57,1%</td>
</tr>
<tr>
<td>SMF</td>
<td>100,0%</td>
<td>33,3%</td>
</tr>
<tr>
<td>Unclassified NewsBoard</td>
<td>100,0%</td>
<td>100,0%</td>
</tr>
<tr>
<td>Vanilla</td>
<td>100,0%</td>
<td>100,0%</td>
</tr>
<tr>
<td>vBulletin</td>
<td>92,3%</td>
<td>84,6%</td>
</tr>
<tr>
<td>miscellaneous</td>
<td>60,0%</td>
<td>60,0%</td>
</tr>
</tbody>
</table>