Vandalism Detection on Wikipedia

The class imbalance problem & new approaches
Contents

Vandalism detection
The class imbalance problem
Content based classifiers
Wikipedia in Numbers

920 K

4.7 M

24

6 M
Vandalism

“Vandalism is any addition, removal, or change of content, in a *deliberate* attempt to compromise the integrity of Wikipedia.”

en.wikipedia.org/wiki/Wikipedia:Vandalism
# Demo

## Webis Wikipedia Vandalism Detection Bot

🔗 Tracking and classifying current edits received from en.wikimedia.org IRC channel #en.wikipedia.

<table>
<thead>
<tr>
<th>#</th>
<th>Article</th>
<th>Editor</th>
<th>Comment</th>
<th>Edited at</th>
<th>Confidence</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Kaun Banega Crorepati</td>
<td>Svpnikhil</td>
<td>/* Other versions */</td>
<td>16:41, 12 Oct 2014</td>
<td>0.5</td>
<td>Show Edit</td>
</tr>
<tr>
<td>12</td>
<td>Disney Channel Circle of Stars</td>
<td>80.31.102.189</td>
<td></td>
<td>16:42, 12 Oct 2014</td>
<td>0.7</td>
<td>Show Edit</td>
</tr>
<tr>
<td>15</td>
<td>Sean Bell shooting incident</td>
<td>159.92.9.130</td>
<td></td>
<td>16:42, 12 Oct 2014</td>
<td>1</td>
<td>Show Edit</td>
</tr>
<tr>
<td>65</td>
<td>Disney Channel Circle of Stars</td>
<td>80.31.102.189</td>
<td>[[WP:AES]]</td>
<td>Replaced content with 'gOOD mOORNING'</td>
<td>16:42, 12 Oct 2014</td>
<td>0.8</td>
</tr>
<tr>
<td>158</td>
<td>Streptococcus</td>
<td>99.11.160.173</td>
<td>/* Molecular taxonomy and phylogenetics */</td>
<td>16:43, 12 Oct 2014</td>
<td>1</td>
<td>Show Edit</td>
</tr>
<tr>
<td>195</td>
<td>Acropolis of Athens</td>
<td>92.75.118.59</td>
<td></td>
<td>16:44, 12 Oct 2014</td>
<td>0.6</td>
<td>Show Edit</td>
</tr>
</tbody>
</table>
Detecting Vandalism

Learning
Detecting Vandalism
The Detection System

Recall

Precision

PR-AUC

0.82

0.72

0.67

0.66

0.5

0.4

0.3

0.2

0.1

0

0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Recall

our system, PAN-WVC-10
Mola-Velasco 2010, PAN-WVC-10
our system, PAN-WVC-11
West et al. 2011, PAN-WVC-11
Class Imbalance

Training dataset

- 93% Regular
- 7% Vandalism
Class Imbalance Problem

Reasons:

1. minimizing the overall error
2. assuming balanced class distribution
3. assuming equal misclassification cost
Dataset Resampling

Random Undersampling

SMOTE = Synthetic Minority Oversampling Technique

Dataset Resampling

RealAdaBoost

Dataset Resampling

Training dataset resampling PAN-WVC-11

Random Forest

One-class Classification

training solely on vandalism samples
One-class Classification

One-class Classification

One-class SVM and resampling on PAN-WVC-11

Precision

Recall

One-class SVM

Content-based Classifiers

**article-based**: automatically compiled simple vandalism edits as training data

**category-based**: unique vandalism style in each article category
Content-based classifiers

Category: Geographical places
Conclusions

**Dataset Resampling**: no overall improvement using simple strategies

**One-class classification**: not suitable with the used settings

**Content based classifiers**: improved approaches may be promising
Code

webis-de/wikipedia-vandalism-detection
webis-de/wikipedia-vandalism-analyzer
webis-de/wikipedia-vandalism-bot
Precision & Recall

TP... true positive
FP... false positive
FN ... false negative

precision = TP / (TP + FP)
recall = TP / (TP + FN)
Detecting Vandalism

Johann Sebastian Bach: Difference between revisions

Revision as of 00:19, 8 September 2014 (view source)
Buxtehude (talk | contribs)
← Previous edit

Revision as of 14:54, 8 September 2014 (view source)
Gap24 (talk | contribs)
Next edit →

Line 9:

{{spaced ndash}}28 July 1750) was a **German** composer and musician of the [[Baroque]]

{{spaced ndash}}28 July 1750) was an **Austrian** composer and musician of the [[Baroque]]
References

Icons are taken from www.flaticon.com.


References (cont.)


Hempstalk, K.; Frank, E. & Witten, I. H.: One-Class Classification by Combining Density and Class Probability Estimation, *ECML/PKDD (1)*, 2008, 505,519