On the Use of Reliable-Negatives Selection Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia

Edgardo Ferretti and Marcelo Errecalde
Universidad Nacional de San Luis
{ferretti,merreca}@unsl.edu.ar

Maik Anderka
University of Paderborn
maik.anderka@uni-paderborn.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

11th International Workshop on Text-based Information Retrieval, TIR’14
Munich, Germany

September 4th, 2014
Information Quality in Wikipedia

Situation

- extremely varying content quality
  - everyone can edit Wikipedia, even anonymously
  - heterogeneous community of Wikipedia authors
  - edits are not reviewed before publication

- comprehensive manual quality assurance is unfeasible
  - large data volumes, constantly evolving contents
Information Quality in Wikipedia

Situation

- extremely varying content quality
  - everyone can edit Wikipedia, even anonymously
  - heterogeneous community of Wikipedia authors
  - edits are not reviewed before publication

- comprehensive manual quality assurance is unfeasible
  - large data volumes, constantly evolving contents

Previous work

- research question: “Is an article featured or not?”
  - no practical support for Wikipedia’s quality assurance process
  - less than 0.1% of the English Wikipedia articles are featured
Quality Flaw Prediction in Wikipedia

Question

- How to improve the 99.9% non-featured Wikipedia articles?

Central idea

- automatic exploitation of human-defined cleanup tags  [Anderka et al., WWW’11]
Quality Flaw Prediction in Wikipedia

Question

- How to improve the 99.9% non-featured Wikipedia articles?

Central idea

- automatic exploitation of human-defined cleanup tags [Anderka et al., WWW’11]
  - each tag defines a specific quality flaw
  - tagged articles serve as human-labeled examples
  - machine learning is used to predict flaws in untagged articles

Existing flaw prediction approaches

- one-class classification [Anderka et al., WWW’11, SIGIR’12]
- binary classification [Ferschke et al., CLEF’12, ACL’13]
- **PU learning** [Ferretti et al., CLEF’12]
Outline

- Motivation
- Problem Statement
- Quality Flaw Prediction Using PU Learning
- Analysis and Empirical Evaluation
- Summary
Problem Statement
Quality flaw prediction in Wikipedia [Anderka et al., SIGIR’12]

- 3.8 M English Wikipedia articles $\rightarrow D$
- 445 quality flaws (cleanup tags) $\rightarrow F$

- Build a classifier $c : D \rightarrow \{1; 0\}$ for each flaw $f \in F$, given a sample of articles containing $f$.

flawed articles $\rightarrow$ article representation (document model) $\rightarrow$ one-class classifier
Problem Statement

Quality flaw prediction using PU learning [Ferretti et al., CLEF’12]

- exploit untagged articles to improve the effectiveness of a classifier \( c \)

PU learning: learning from \textit{Positive} and \textit{Unlabeled} examples [Liu et al., ICML'02]

- \textit{positive} examples = articles tagged with a flaw
- \textit{unlabeled} examples = untagged articles (either flawed or flawless)
Problem Statement

Background: PU learning  [Liu et al., ICML’02]

- set $P$ of positive examples
- set $U$ of unlabeled examples (containing both positive and negative examples)
- Build a classifier using $P$ and $U$ that can identify positive examples in $U$ or in a separate test set.

- two-stage approach:
  1. identifying **reliable negatives**
     - train a binary classifier using $P$ and $U$
     - apply this classifier to the examples in $U$
     - consider all examples not classified as “positive” as **reliable negatives**
  2. building the final classifier (non-iterative version)
     - train a binary classifier using $P$ and the set of **reliable negatives**
Problem Statement
Crucial aspects in the Wikipedia setting

1. unknown (flaw-specific) class imbalances
   - 1\(^{st}\) stage: ratio between \(P\) and \(U\)
   - 2\(^{nd}\) stage: ratio between \(P\) and the set of reliable negatives

2. effects of sampling (essential in practice due to the large number of existing Wikipedia articles)
   - 1\(^{st}\) stage: \(U\) is very large for most flaws
   - 2\(^{nd}\) stage: the set of reliable negatives can become considerably large

- have not—or only partially—addressed by Liu et al. and Ferretti et al.

→ we show where in the PU learning procedure sampling is useful
→ we analyze how different sampling strategies affect the flaw prediction effectiveness
Outline

- Motivation
- Problem Statement
  - Quality Flaw Prediction Using PU Learning
  - Analysis and Empirical Evaluation
- Summary
Quality flaw prediction using PU learning

1st stage: identifying *reliable negatives*

- $U_1$ is a sample from $U$
- training set is balanced, $|P| = |U_1|$

$\Rightarrow$ sampling strategy does not affect the flaw prediction performance

$\Rightarrow$ random sampling
Quality flaw prediction using PU learning

2nd stage: building the final classifier

- using $U_2 = U^n$ worsened the performance by up to 50% [Ferretti et al., CLEF’12]

- sampling strategies:
  
  $M_1$ selecting $|P|$ articles by random from $U^n$
  
  $M_2$ selecting the $|P|$ best articles from $U^n$
  
  (those assigned the highest confidence values by the first-stage classifier)
  
  $M_3$ selecting the $|P|$ worst articles from $U^n$
  
  (those assigned the lowest confidence values by the first-stage classifier)
Outline

- Motivation
- Problem Statement
- Quality Flaw Prediction Using PU Learning
- Analysis and Empirical Evaluation
- Summary
Analysis and Empirical Evaluation

Experimental design

- evaluation corpus of the “1st international competition on quality flaw prediction in Wikipedia”
  - 1,592,226 English Wikipedia articles
  - 208,228 tagged to contain one of ten important quality flaws

- 1st stage classifier: Naïve Bayes

- 2nd stage classifier: Support Vector Machine (SVM)

- balanced training sets: $|P| = |U_1|$ and $|P| = |U_2|$

- random sampling in the 1st stage

- $M_1$, $M_2$, and $M_3$ in the 2nd stage
Analysis and Empirical Evaluation

Selecting reliable negatives (2nd stage sampling)

- flaw Unreferenced: \(|U^n| = 29,635, \ |P| = |U_2| = 1,000\)
Analysis and Empirical Evaluation

Selecting *reliable negatives* (2nd stage sampling)

- flaw *Unreferenced*: $|U^n| = 29,635$, $|P| = |U_2| = 1,000$

→ strategy $M_3$ outperforms $M_2$

→ differences between $M_3$ and $M_1$ (random) are not statistically significant
Analysis and Empirical Evaluation

Flaw prediction effectiveness

effectiveness of PU learning in terms of F1 score for the ten quality flaws

<table>
<thead>
<tr>
<th>flaw name</th>
<th>baseline [Ferretti et al., CLEF'12]</th>
<th>proposed approach using strategy $M_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advert</td>
<td>0.8214</td>
<td>0.9440 (+14.93%)</td>
</tr>
<tr>
<td>Empty section</td>
<td>0.8216</td>
<td>0.9394 (+14.34%)</td>
</tr>
<tr>
<td>No footnotes</td>
<td>0.8264</td>
<td>0.9826 (+18.90%)</td>
</tr>
<tr>
<td>Notability</td>
<td>0.7944</td>
<td>0.9886 (+24.45%)</td>
</tr>
<tr>
<td>Orphan</td>
<td>0.8986</td>
<td>0.9960 (+10.84%)</td>
</tr>
<tr>
<td>Original research</td>
<td>0.7638</td>
<td>0.9338 (+22.26%)</td>
</tr>
<tr>
<td>Primary sources</td>
<td>0.8068</td>
<td>0.9891 (+22.60%)</td>
</tr>
<tr>
<td>Refimprove</td>
<td>0.8362</td>
<td>0.9382 (+12.20%)</td>
</tr>
<tr>
<td>Unreferenced</td>
<td>0.8365</td>
<td>0.9432 (+12.76%)</td>
</tr>
<tr>
<td>Wikify</td>
<td>0.7396</td>
<td>0.9818 (+32.75%)</td>
</tr>
<tr>
<td><strong>averaged over all flaws</strong></td>
<td><strong>0.8145</strong></td>
<td><strong>0.9637 (+18.31%)</strong></td>
</tr>
</tbody>
</table>
Outline

- Motivation
- Problem Statement
- Quality Flaw Prediction Using PU Learning
- Analysis and Empirical Evaluation
- Summary
Summary
What we have done

1. shed light on the effects of sampling in PU learning
   ➔ sampling is necessary (in both stages)
   ➔ in general, sampling strategy $M_3$ is favorable

2. improved PU learning approach for quality flaw prediction in Wikipedia
   ➔ average improvement of 18.31% compared to the baseline
Summary

What we have done

1. shed light on the effects of sampling in PU learning
   ➔ sampling is necessary (in both stages)
   ➔ in general, sampling strategy $M_3$ is favorable

2. improved PU learning approach for quality flaw prediction in Wikipedia
   ➔ average improvement of 18.31% compared to the baseline

Current work

- comparative study of the existing flaw prediction approaches
Thank you!

maik.anderka@uni-paderborn.de
Appendix
Article representation

- 65 state-of-the-art features, 30 new features

  - **content** characters, words, syllables, sentences, readability, parts of speech, closed-class word sets, . . .
  - **structure** sections, tables, images, references, categories, templates, lists, specific sections, . . .
  - **network** internal-, external-, interwiki-, broken links, PageRank, citation measures, . . .
  - **edit history** age, currency, connectivity, revisions, reverts, editors, cooperation, . . .