Temporal Analysis for Web Spam Detection: An Overview

Miklós Erdélyi\textsuperscript{1,2}, András A. Benczúr\textsuperscript{1}

\textsuperscript{1}Hungarian Academy of Sciences (MTA SZTAKI)
\textsuperscript{2}University of Pannonia
Motivation

• Try out more sophisticated machine learning methods.
• Compare recent methods in Web spam filtering exploiting temporal information.
• Extend existing methods to perform even better.
Overview of Temporal Features for Web Spam Detection

- Link-based features.
  - Change in linkage-related “static” properties.
    - Average number of out-links, number of neighbors at $n$ steps, etc.
  - Measuring linkage change in single-step neighborhood.

- Content-based features.
  - Change in standard content-based feature values over time.
  - Change in term-weight vectors.
  - Combination of term-weight vectors over time.
Measuring Linkage Change

• Considering only single-step neighborhood.

• Inlink death- and growth rate \([Shen \ et. \ al]\):

\[
IDR(a) = \frac{|I^{(t_0)}(a) - I^{(t_1)}(a)|}{|I^{(t_0)}(a)|}
\]

\[
IGR(a) = \frac{|I^{(t_1)}(a) - I^{(t_0)}(a)|}{|I^{(t_0)}(a)|}
\]

• Change rate of the clustering coefficient:

\[
CC(a, t) = \frac{|\{(b, c) \in G(t)| b, c \in \Gamma^{(t)}(a)\}|}{|\Gamma^{(t)}(a)|}
\]

\[
CRCC(a) = \frac{CC(a, t_1) - CC(a, t_0)}{CC(a, t_0)}
\]
Link-based Similarity for Temporal Features

- How to capture linkage change in multi-step neighborhood of a node?
- **Idea**: calculate similarity *across* graph instances having the same labeling.
- Consider node $u$ at time $t_0$ and $t_1$:

![Diagram showing two graph instances with node $u$](image-url)
Extended Jaccard Coefficient

- Captures linkage change in multi-step neighborhood.
- Calculation:
  - Take the $k$-step neighborhood of node $v(t_0)$, $v(t_1)$.
  - Calculate their similarity using Jaccard-coefficient.
  - Take the exponentially weighted sum in $k$:

$$ \text{XJac}_{t_0,t_1}(v) = \sum_{k=1}^{\ell} \frac{|\Gamma_k^{(t_0)}(v) \cap \Gamma_k^{(t_1)}(v)|}{|\Gamma_k^{(t_0)}(v) \cup \Gamma_k^{(t_1)}(v)|} \cdot c^k (1 - c) $$

- Generalizes link growth- and death rate to multi-step neighborhoods.
Why not SimRank?

• Explanation.
  • Recall:
    $$\text{Sim}_{\ell+1}(u, v) = \begin{cases} 1, & \text{if } u = v; \\ \sum_{v' \in I(v)} \sum_{u' \in I(u)} \text{Sim}_\ell(u', v') & \text{otherwise} \end{cases}$$

  • Pages $u$ and $v$ have $k$ witnesses for similarity, yet $\text{sim}(u, v) \approx 1/k$

• Use PSimRank instead…
PSimRank

• Allow coupling of random walks!
• Pair of random walks at vertices $u'$ and $v'$ meet with probability: \[
\frac{|I(u') \cap I(v')|}{|I(u') \cup I(v')|}
\]
• In case of no change self-similarity $= 1$.
• Computable in the same Monte Carlo framework as XJaccard.
• Performance (BAG-DT)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>No. of Features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth/death rates</td>
<td>29</td>
<td>0.605</td>
</tr>
<tr>
<td>XJaccard</td>
<td>42</td>
<td>0.626</td>
</tr>
<tr>
<td>PSimRank</td>
<td>21</td>
<td>0.593</td>
</tr>
<tr>
<td>XJaccard + PSimRank</td>
<td>63</td>
<td>0.610</td>
</tr>
<tr>
<td>Public link-based [5]</td>
<td>176</td>
<td>0.731</td>
</tr>
<tr>
<td>Public + growth/death rates</td>
<td>205</td>
<td>0.696</td>
</tr>
<tr>
<td>Public + XJaccard + PSimRank</td>
<td>239</td>
<td>0.710</td>
</tr>
<tr>
<td>All link-based</td>
<td>268</td>
<td>0.707</td>
</tr>
<tr>
<td>WSC 2008 Winner</td>
<td>-</td>
<td>0.852</td>
</tr>
</tbody>
</table>

• Considering multi-step neighborhood improves link-based temporal features.
Comparison of Link-based Temporal Features (2)

- Sensitivity (BAG-DT)
Content-based Temporal Features

- Simple bag-of-words representation.
- Term-weight vector based features [Dai et. al]:
  - Ave, AveDiff, Dev, DevDiff, Decay
- Selection of the dictionary based on:
  - term-frequencies
  - coverage
  - both.
- A well-selected dictionary is crucial for good classification performance!
Classification Framework

• Diverse set of classifier models combined by ensemble selection.
  • What is ensemble selection?
  • Why ensemble selection?

• Classifier models:
  • Bagged LogitBoost, Decision Trees, Bagged Cost-sensitive Decision Trees, Logistic Regression, Random Forest, Naïve Bayes.
Data Set

- WEBSPAM-UK2007 + 7 earlier snapshots.
- Web Spam Challenge 2008 training/test labels.

<table>
<thead>
<tr>
<th>Label Set</th>
<th>Instances</th>
<th>%Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>4000</td>
<td>5.95%</td>
</tr>
<tr>
<td>Testing</td>
<td>2053</td>
<td>4.68%</td>
</tr>
</tbody>
</table>

- Downloadable temporal feature sets from (soon):
  - [http://datamining.sztaki.hu/?q=en/downloads](http://datamining.sztaki.hu/?q=en/downloads)
  - Try them out!
Results (1)

- Ensembles on temporal link-based features:

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>No. of Features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth/death rates</td>
<td>29</td>
<td>0.617</td>
</tr>
<tr>
<td>XJaccard + PSimRank</td>
<td>63</td>
<td>0.625</td>
</tr>
<tr>
<td>Public link-based [5]</td>
<td>176</td>
<td>0.765</td>
</tr>
<tr>
<td>Public + growth/death rates</td>
<td>205</td>
<td>0.758</td>
</tr>
<tr>
<td>Public + XJaccard + PSimRank</td>
<td>239</td>
<td>0.769</td>
</tr>
<tr>
<td>All link-based</td>
<td>268</td>
<td>0.765</td>
</tr>
<tr>
<td>WSC 2008 Winner</td>
<td>-</td>
<td>0.852</td>
</tr>
</tbody>
</table>

- Temporal link-based features can slightly improve standard link-based features.
Results (2)

• Ensembles on content-based features ...

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>No. of Features</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public content [34]</td>
<td>96</td>
<td>0.879</td>
</tr>
<tr>
<td>Public content + BM25</td>
<td>10096</td>
<td>0.893</td>
</tr>
<tr>
<td>WSC 2008 Winner [25]</td>
<td>-</td>
<td>0.852</td>
</tr>
</tbody>
</table>

• ... and temporal content-based features:

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static BM25</td>
<td>0.736</td>
</tr>
<tr>
<td>Ave</td>
<td>0.749</td>
</tr>
<tr>
<td>AveDiff</td>
<td>0.737</td>
</tr>
<tr>
<td>Dev</td>
<td>0.767</td>
</tr>
<tr>
<td>DevDiff</td>
<td>0.752</td>
</tr>
<tr>
<td>Decay</td>
<td>0.709</td>
</tr>
<tr>
<td>Temporal combined</td>
<td>0.782</td>
</tr>
<tr>
<td>Temporal combined + BM25</td>
<td>0.789</td>
</tr>
<tr>
<td>Public content-based [34] + temporal</td>
<td>0.901</td>
</tr>
<tr>
<td>All combined</td>
<td>0.902</td>
</tr>
</tbody>
</table>
Conclusions

• About content-based features...
  • Term-weight based features are the best for both the temporal and the static setting.
  • The advantage of temporal link-based features diminish when used in combination with content-based features.

• Temporal link-based similarity features might be useful for domains where content is not available.

• Choice of machine learning method is as crucial as the choice of feature set.
Questions?

Miklós Erdélyi

http://datamining.sztaki.hu/?q=en/downloads/

{miklos, benczur}@ilab.sztaki.hu