Temporal Analysis for Web Spam Detection: An Overview

Miklós Erdélyi[1,2], András A. Benczúr[1]
[1]Hungarian Academy of Sciences (MTA SZTAKI)
[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]:

$$\mathrm{IDR}(a) = \frac{|I^{(t_0)}(a) - I^{(t_1)}(a)|}{|I^{(t_0)}(a)|}$$

$$\mathrm{IGR}(a) = \frac{|I^{(t_1)}(a) - I^{(t_0)}(a)|}{|I^{(t_0)}(a)|}$$

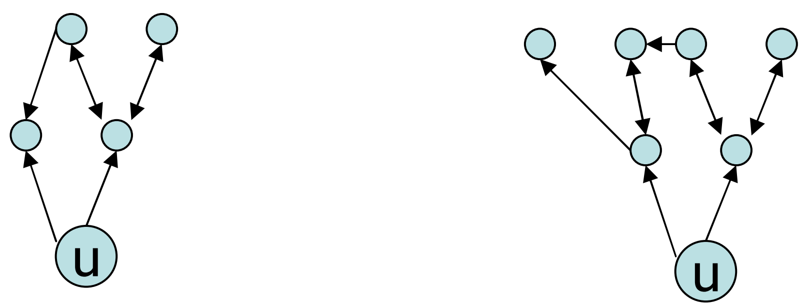
- Change rate of the clustering coefficient:

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

$$CRCC(a) = \frac{CC(a,t_1) - CC(a,t_0)}{CC(a,t_0)}$$

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:



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*:

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

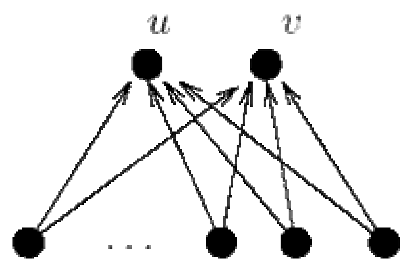
- Generalizes link growth- and death rate to multi-step neighborhoods.

Why not SimRank?

- Explanation.
  - Recall:

$$\begin{aligned}\mathrm{Sim}_{\ell+1}(u,v) &= 1, \text{ if } u = v;\\ \mathrm{Sim}_{\ell+1}(u,v) &= \sum_{\substack{v' \in I(v)\\ u' \in I(u)}} \mathrm{Sim}_{\ell}(u',v')\end{aligned}$$

  - Pages *u* and *v* have *k* witnesses for similarity, yet sim(*u*,*v*) ≈ 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.

Comparison of Link-based Temporal Features (1)

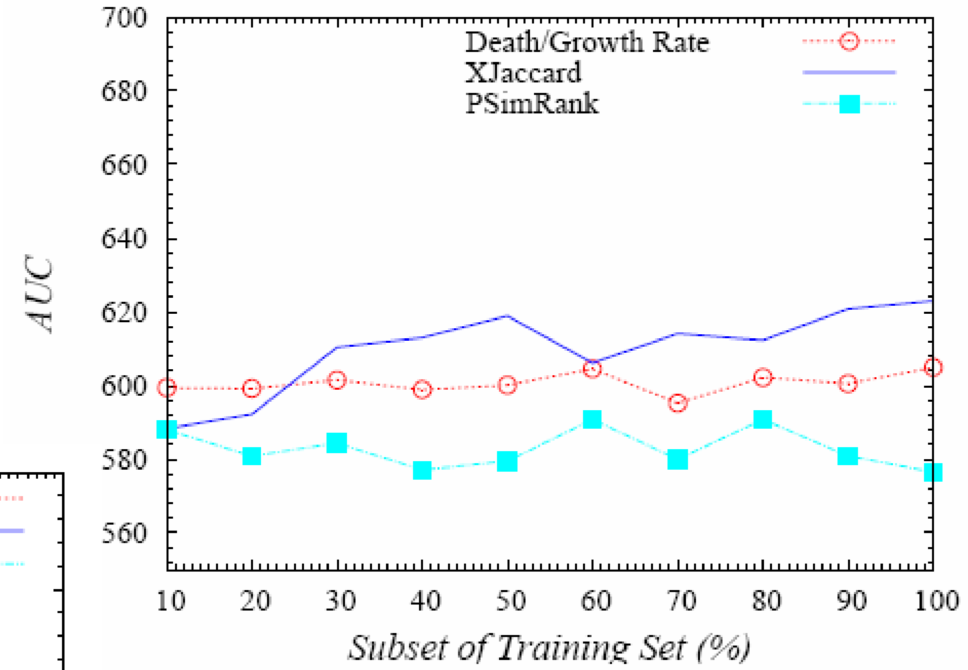
- Performance (BAG-DT)

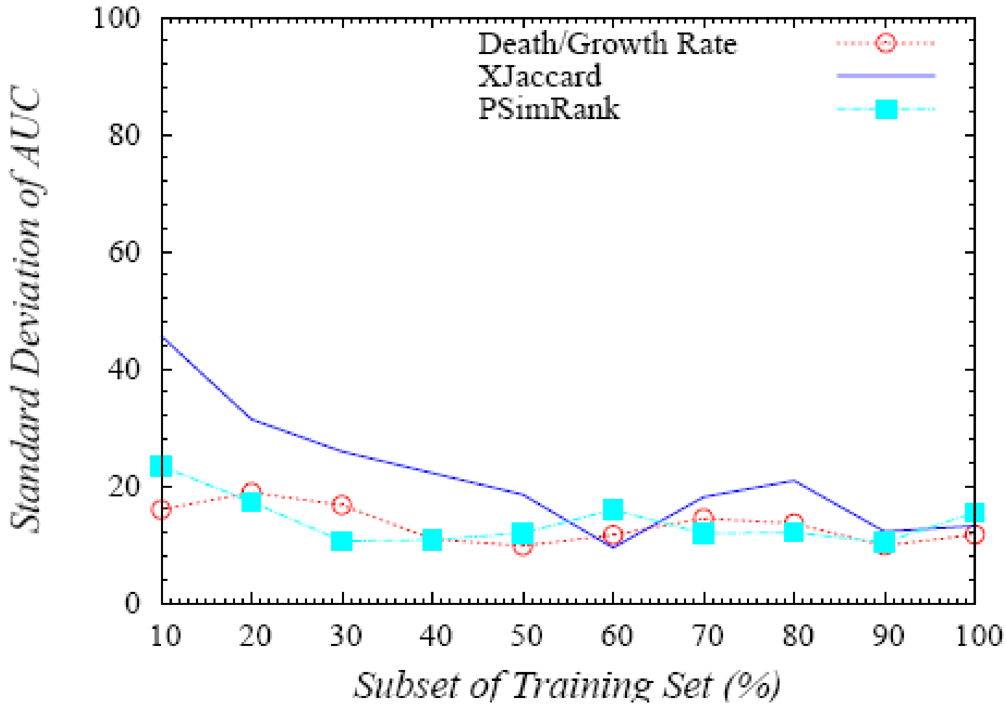
Feature Set	No. of Features	AUC
Growth/death rates	29	0.605
XJaccard	42	0.626
PSimRank	21	0.593
XJaccard + PSimRank	63	0.610
Public link-based [5]	176	**0.731**
Public + growth/death rates	205	0.696
Public + XJaccard + PSimRank	239	*0.710*
All link-based	268	0.707
WSC 2008 Winner	-	0.852

- 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.
  - uk-2007-May, uk-2007-Apr, …, uk-2006-Oct
- Web Spam Challenge 2008 training/test labels.

Label Set	Instances	%Positive
Training	4000	5.95%
Testing	2053	4.68%

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

Results (1)

- Ensembles on temporal link-based features:

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

- Temporal link-based features can slightly improve standard link-based features.

Results (2)

- Ensembles on content-based features …

Feature Set	No. of Features	AUC
Public content [34]	96	0.879
Public content + BM25	10096	**0.893**
WSC 2008 Winner [25]	-	0.852

- ... and temporal content-based features:

Feature Set	AUC
Static BM25	0.736
Ave	0.749
AveDiff	0.737
Dev	0.767
DevDiff	0.752
Decay	0.709
Temporal combined	0.782
Temporal combined + BM25	**0.789**
Public content-based [34] + temporal	0.901
All combined	**0.902**

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