Web Spam Hunting

Dávid Siklósi
sdavid@ilab.sztaki.hu

joint work with
András A. Benczúr István Bíró Zsolt Fekete
Miklós Kurucz Attila Pereszlényi Simon Rácz
Adrienn Szabó Jácint Szabó
{benczur, ibiro, zsfekete, mkurucz, peresz, sracz,
aszabo, jacint}@ilab.sztaki.hu

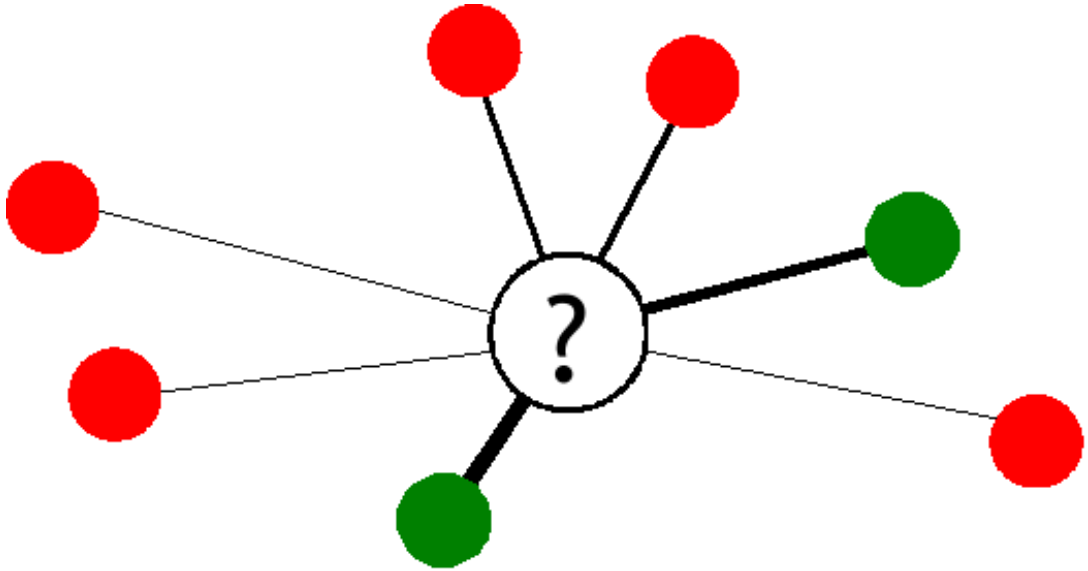
Data Mining and Web Search Group
Computer and Automation Research Institute
Hungarian Academy of Sciences

Content classification

- Bagging cost sensitive SVM (linear kernel $\gamma = 1$) over tf.idf
- Content Classification by Latent Dirichlet Allocation (AIRWeb 2008)
  - Blei, Ng, Jordan, 2003 (LDA)
  - topic model: distribution over words
  - document model: distribution over topics
  - spamicity measured as similarity to spam vs. honest topic models
- Dynamic Markov Compression
  - Bratko, Filipič, Cormack, Lynam, Zupan (2006)
  - Ratio of compression rates when added to spam vs. normal

Stacked Graphical Learning: Overview

- Predict spamicity $p(v)$ of node v
- For target node u, aggregate $p(v)$ for neighbors to form new feature $f(u)$
- Rerun classification by adding feature $f(.)$
- Iterate



Stacked Graphical Learning: features

- Choice of neighbors (node similarities), direction, aggregation (Spam Challenge 2007 part II)
  - Similarity: edge weight, neighborhood (Jaccard, cosine, Adamic/Adar, ...), path ensemble (PageRank, SimRank, Katz, ...)
- Site Structure Analysis and Stacking
  - Apply the "Connectivity Sonar" features of Amitay et al. (Hypertext 2003)
    - Average, most populated level
    - In and outlinks distributed across pages and levels
    - Leaf and root level linkage
  - Extend in a graph stacking framework: weight by predicted spamicity
    - Honest directories may contain some spam at bottom
    - Virtual hosting may contain spam below the root

Additional features

- Commercial Intent Features (AIRWeb 2007)
  - Microsoft OCI (commercial intention) scores
    - publicly available at http://adlab.msn.com/OCI
  - Penalties for high rank:
    - own search engine (Okapi-based)
    - competitive queries measured by Google AdWords
- New features
  - Number of document formats (doc, pdf etc)
  - existence and value of `robots.txt` and robots meta
  - existence and average of server last modified dates
  - distance and personalized PageRank from DMOZ top categories

Combination and Results

- Random forest over classifiers
  (outperforms logistic regression proposed last year by Gordon Cormack)
  - 1–3 SVM, LDA, Compression
  - 4–6 C4.5's over public link, public content and additional features
- Compute graphical features, add C4.5 classification above

Results

	F-measure	ROC
link	0.198	0.689
sonar	0.204	0.684
additional	0.305	0.71
content	0.349	0.73
LDA	0.448	0.77
SVM	0.496	0.926
DMC	0.667	0.95
Combined	0.744	0.98

Future challenges?

- The LiWA: Living Web Archives EU FP7 Project
  - User partners: European Internet Archive, Sound and Vision (NL), ...
  - Research partners: L3S Hannover, MPI Saarbrucken
  - We lead spam filtering efforts
- We plan to provide time history crawl for spam filtering experiments
  - Recrawl by archive crawler – reuse existing assessment labels
  - Needs careful definition as an Archive .uk crawl is huge with 2M sites
- New areas
  - Active learning to optimize manual assessment efforts
  - Spam time evolution analysis in archives
  - New forms of Spam: social media, multimedia, ...