

# Web Spam Filtering in Internet Archives

Miklós Erdélyi<sup>1</sup>, András A. Benczúr<sup>1</sup>, Dávid Siklósi<sup>1</sup>  
Julien Masanès<sup>2</sup>

<sup>1</sup>Hungarian Academy of Sciences (MTA SZTAKI)  
Data Mining and Web Search Group

<sup>2</sup>European Archive Foundation, France

## Part I: Archival Institutions

- Web archives: 39 institutions under International Internet Preservation Consortium (IIPC) and more
- Loose collaboration, code and technology sharing
- Operated usually from moderate budget
- Effort sharing is crucial





# Web spam in Archives



- Slightly different policies
  - May want to archive spam to preserve whole picture
  - Might be worried more about false positives
  - Will perhaps not serve general search queries to users
- But increasingly affected by spam becoming more and more costly if not fought against:
  - 10+% of sites, near 20% of HTML pages
- We have conducted a survey ...

# Survey results (1)

- Participants: 20 archival institutions from all around the Globe
  - National and other libraries
    - Library of Congress
    - National Library of Denmark
    - ...
  - Internet Archive
  - Documentation Centre for Dutch Political Parties
  - Virtual Knowledge Studio
  - ...
- *“Is spam or fake Web content a problem in your crawling and capturing process?”*
  - Yes (39%)
  - No problem (4%)
    - Only 1 respondent expects no problem by spam even in the future
- The type of spam met by archives, counter measures ...

## Survey results (2)

- *“If you do meet spam during capturing, of what type is that spam?”*

Blog comment spam	20% (2)
Link farms	50% (5)
Copied content	60% (6)
Garbage content	70% (7)

- *“If spam has impact on your Web archiving process, what actions do you undertake?”*

→ We drop pages with spam or fake content.	18,20% (2)
We drop sites with spam or fake content.	45,50% (5)
→ We apply filters to avoid such noise.	54,50% (6)
After capturing we manually correct the crawl.	27,30% (3)
We see no options to avoid noise.	27,30% (3)

## Survey results (3)

- Low resources on spam filtering but...
- *“If you undertake actions to diminish the spam problem in the Web archive of your institute, can you estimate how much you invest in this?”*
  - “difficult to estimate”
  - “I would spend perhaps 3 or 4 days creating lists of seeds to filter out of the forthcoming crawl.”
  - “10 minutes - 1 hour per site”
  - “We use 2-5 minutes per website when going through the list of potential spam sites.”
  - “We do not do anything to edit captured content. We foresee that this would not scale, and that it would invite questions about the archive’s authenticity.”



# Archive specific needs



- Filtering
  - Analyze and train by a “bootstrap” crawl
  - Filter **newly appeared hosts** crawl time
  - Aid manual assessment (active learning)
- Collaboration
  - Aid information and label sharing
  - **Use a filter model trained possibly at another institution**
  - Catch spam farms that span more top level domains

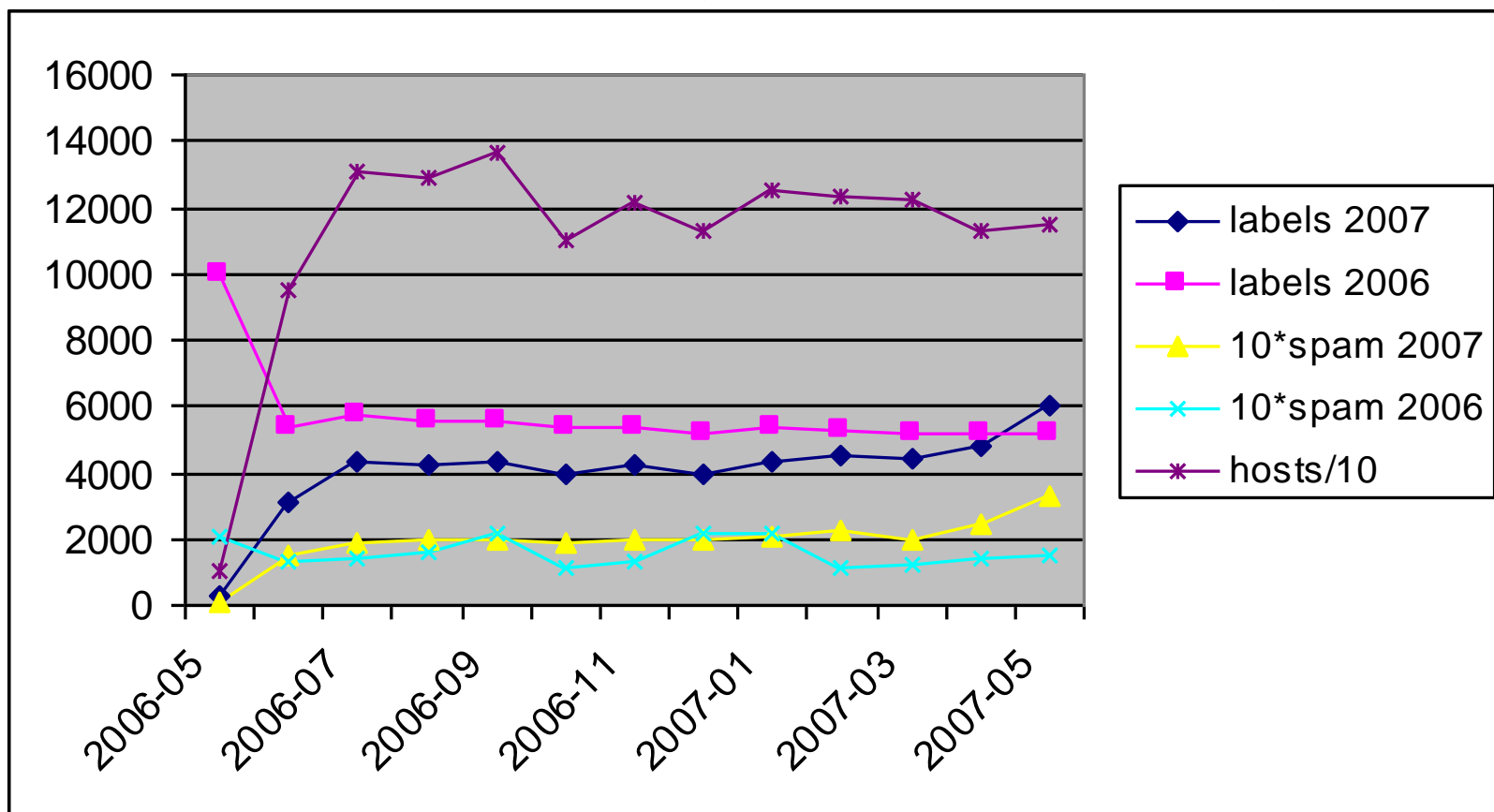
# Part II: Spam hunting in Archives

- Dataset  
(WEBSPAM-UK snapshots)
- Temporal features
- Results



- 13 UbiCrawler .uk snapshots

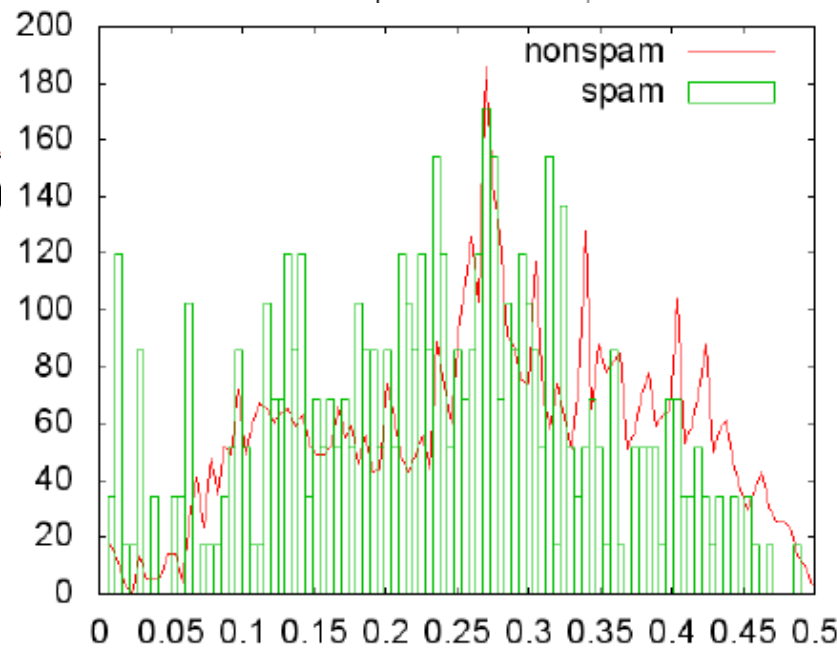
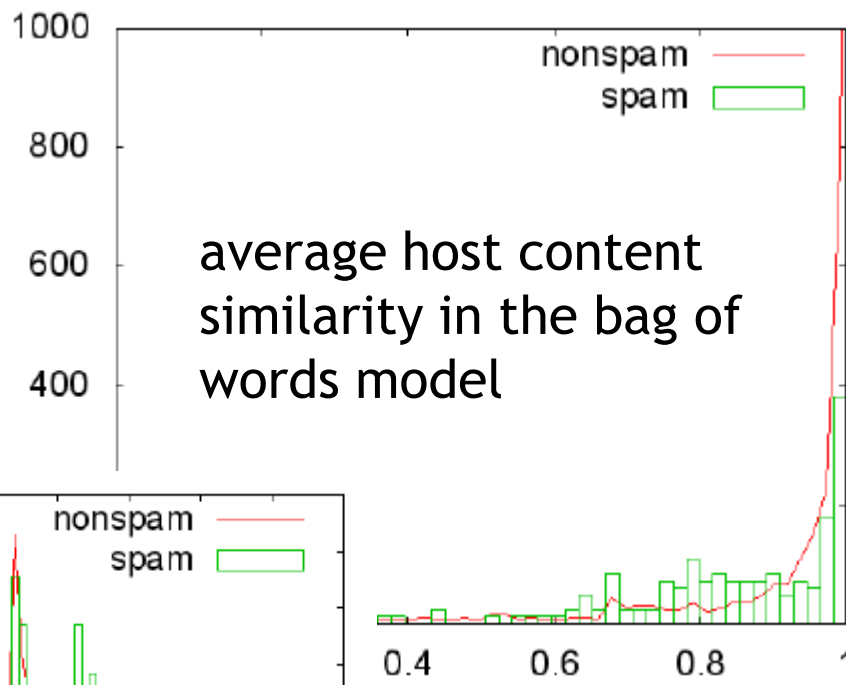
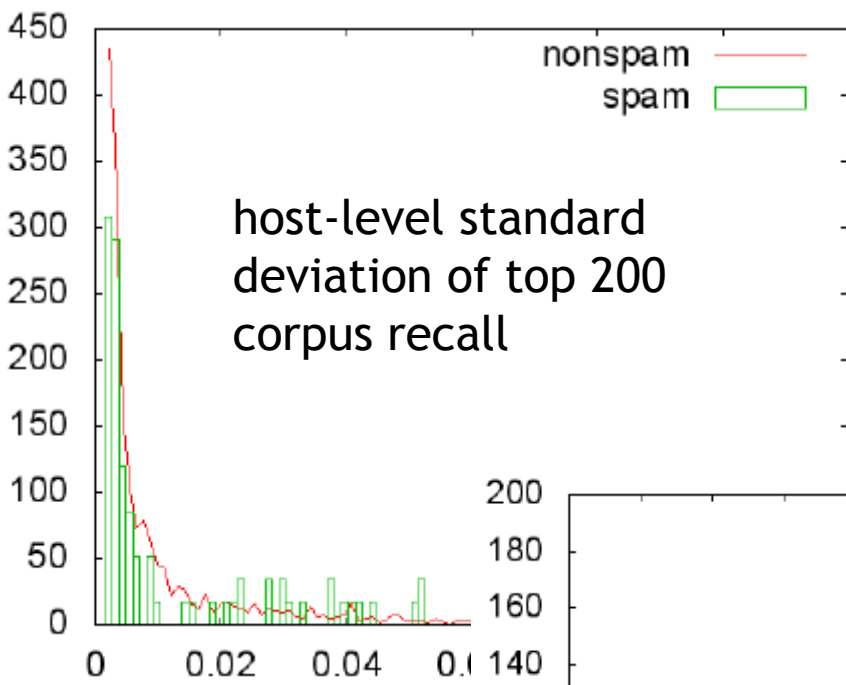
uk2006-05 (WEBSPAM-UK2006), uk2006-06, ..., uk2007-04, uk2007-05 (WEBSPAM-UK2007)



- Transformations - without generating new features
  - Normalization: centralized feature values
  - Variance of feature values across snapshots
- Content change
  - Simple bag-of-words model
  - Similarity between two snapshots
  - Aggregated by average, maximum and variance
- Classification stability
  - On average, how easy it is to classify the given host?



# Sample histograms



# Results (1)

- Using “public” content features, classification by C4.5
- Related feature sets:

Setup	Challenge	New host	2006 → 2007
Training set size	1,201	4,000	10,662
Public content	0.753	0.699	0.730
BOW	0.619	-	-
Stability	<b>0.776</b>	-	-
Variance	0.618	-	-

- Combination by log-odds averaging based random forest:

Combinations	Challenge
Content + BOW	0.729
Content + stability	0.766
Content + variance	0.726
Content + BOW + stability + variance	<b>0.777</b>

- *Conclusion*: temporal change based features seem to be useful by these preliminary experiments

# *Questions?*

**Miklós Erdélyi**

[datamining.sztaki.hu/](http://datamining.sztaki.hu/)

[miklos@ilab.sztaki.hu](mailto:miklos@ilab.sztaki.hu)