Link-Based Similarity Search to Fight Web Spam

Károly Csalogány,
Computer and Automation Institute,
Hungarian Academy of Sciences

Joint work with András A. Benczúr and Tamás Sarlós

August 10, 2006

Contents

Brief recap of link-spam

- Honest links [Chakrabarti et al., 1999]:

  "hyperlink structure contains an enormous amount of latent human annotation that can be extremely valuable for automatically inferring notions of authority"

- High revenue for top search engine ratings
- Manipulations, "Search Engine Optimization"
  - content spam
  - link spam – focus of the talk

Personalized PageRank, TrustRank, BadRank

Definition: random surfer with *teleportation distr.* r

$$\mathrm{PPR}_r(u) = c \cdot r(u) + (1-c) \sum_{vu \in E} \mathrm{PPR}_r(v)/d^+(v)$$

- TrustRank [Gyöngyi et al., VLDB 2004]
  - Personalizes on trusted pages
  - Propagates trust forward
  - Needs very carefully selected trusted hub set
- BadRank [Google folklore]
  - Penalizes by personalization on known spam
  - Propagates distrust backwards

TrustRank, BadRank Cont'd

Different schemes for trust and distrust splitting and aggregation [Wu et al., MTW 2006]

- splitting: equal, constant, logarithm
- aggregation: simple, maximum, maximum parent
- combining trust and distrust

Contents

How to Detect Spam with Similarity Search?

- Algorithms produce top list of similar pages
- Extract features based on the known spam and honest hosts in the list
- Impose threshold on the features
  - different threshold - different quality
  - decreasing threshold - increasing recall
- Precision-recall curves

Spam Thresholds by Similarity Top Lists

- Similarity top list size: k
- Honest from evaluation sample in top list: h
  sum of their similarity value: h^*
- Spam from evaluation sample in top list: s
  sum of their similarity value: s^*
- In general $h + s < k$

SR	Spam Ratio: $s/(s+h)$
SoN	Spam over Non-spam: s/h
SVR	Spam Value Ratio: $s^*/(s^*+h^*)$
SVoNV	Spam Value over Non-spam Value: s^*/h^*

Algorithms

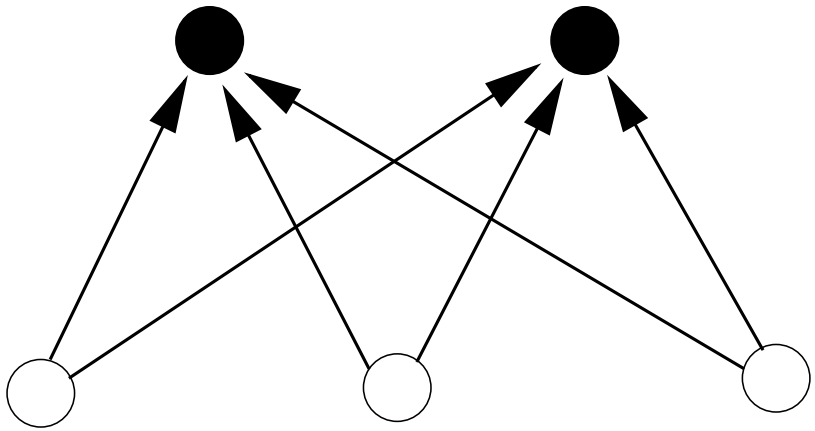
- Baseline algorithms
  - BadRank (distrust propagation)
  - Distrust propagation with different splitting and aggregation methods
  - Combined trust and distrust propagation
- Similarity algorithms
  - Cocitation
  - Companion
  - SimRank

  For each similarity algorithm compute the 4 features (SR, SoN, SVR, SVoNV)

Cocitation

Definition: Cocitation of a and b is the number of nodes that link to both a and b

- Easy to compute
- Easy to manipulate for spammers



SimRank

"Two pages are similar if referenced by similar pages" [Jeh–Widom KDD 2002]:

$$\mathrm{Sim}^{(0)}(u_1, u_2) = \begin{cases} 0 & \text{if } u_1 \neq u_2 \\ 1 & \text{if } u_1 = u_2 \end{cases}$$

$$\mathrm{Sim}^{(k)}(u_1, u_2) = \begin{cases} (1-c) \sum_{(v1,u1),(v2,u2)\in E} \dfrac{\mathrm{Sim}^{(k-1)}(v_1, v_2)}{d^-(u_1) \cdot d^-(u_2)} & \text{if } u_1 \neq u_2 \\ 1 & \text{if } u_1 = u_2 \end{cases}$$

- Similar to PageRank
- Generalization of cocitation
- Hard to manipulate
- Efficient algorithms [SBCsFR 2006]

HITS, Companion and the TKC effect

- Hypertext Induced Topic Search (HITS) [Kleinberg, 1999]
  - Finds good hub and authority pages
  - Hub and Authority scores in the vicinity of seed page(s)
  - Known to be vulnerable to Tightly Knit Communities (TKC)
- Companion [Dean–Henzinger, 1999]
  - Finds related pages
  - Performs HITS in the 2-step alternating neighborhood

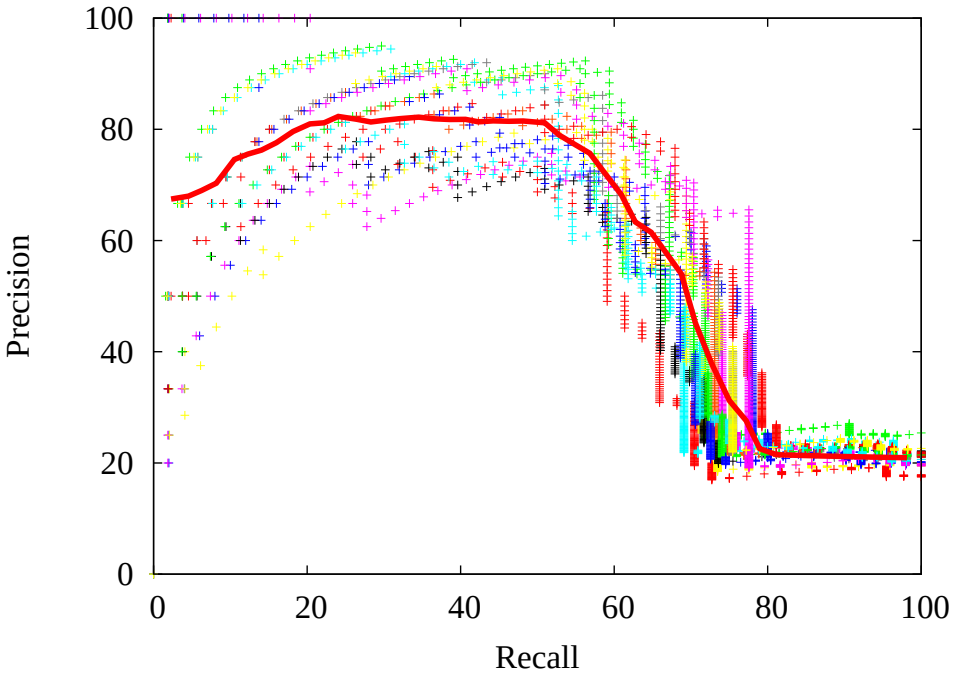
Outline

Evaluation Data Sets

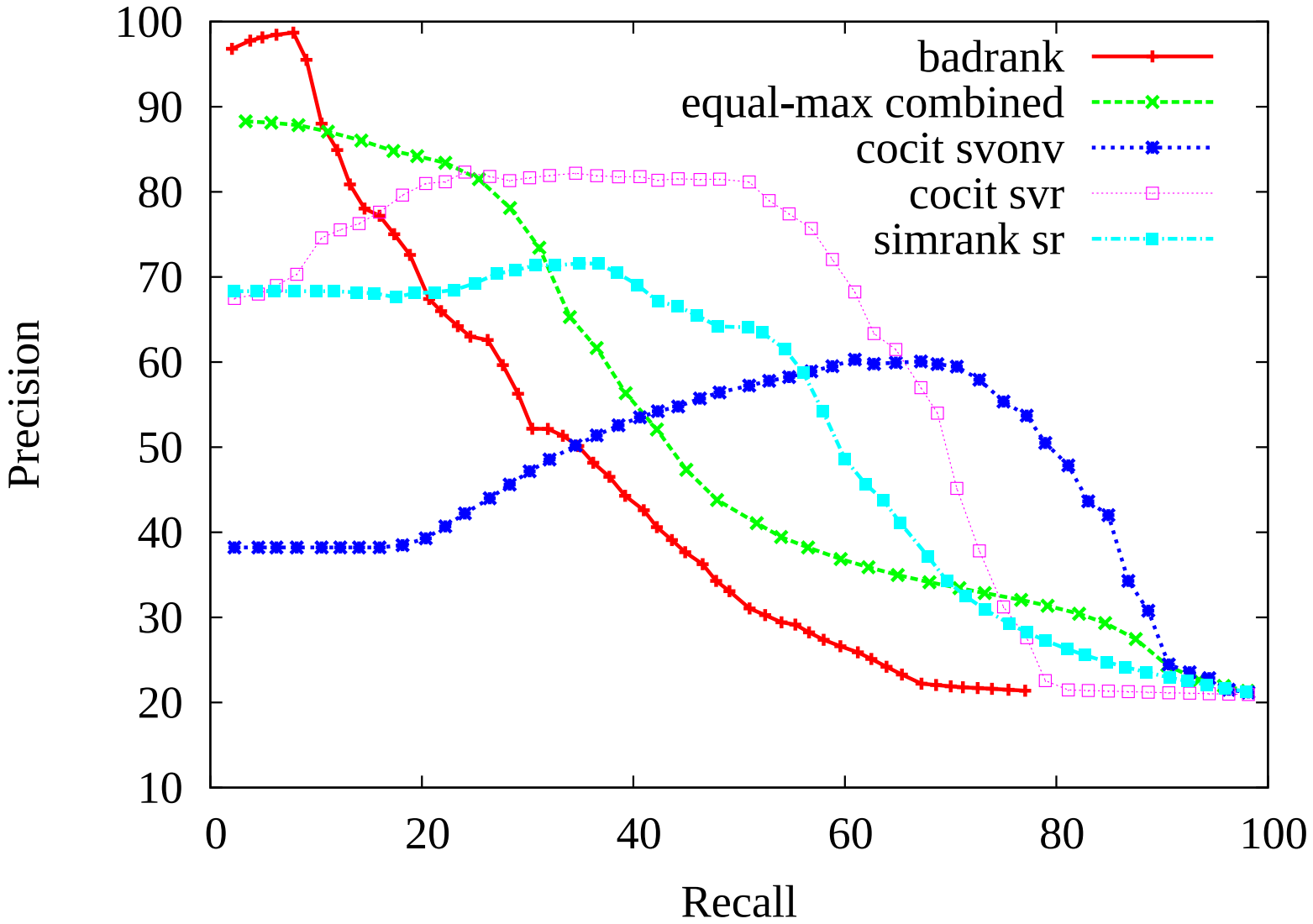
- `.de` domain
  - courtesy of T. Suel
  - 31M pages, 1B edges $\longrightarrow$ 800K sites, 25M edges
  - manually evaluated 1000-page sample with bias towards large PageRank
  - sample contains 20% spam
- `search.ch` data
  - courtesy of U. Müller
  - 20M pages, $\longrightarrow$ 300K domains, 24M edges
  - proprietary blacklist of `search.ch`
  - whitelist of B. Wu and B. Davison (Swiss ODP)
  - labeled set contains 4% spam

Evaluation Methodology

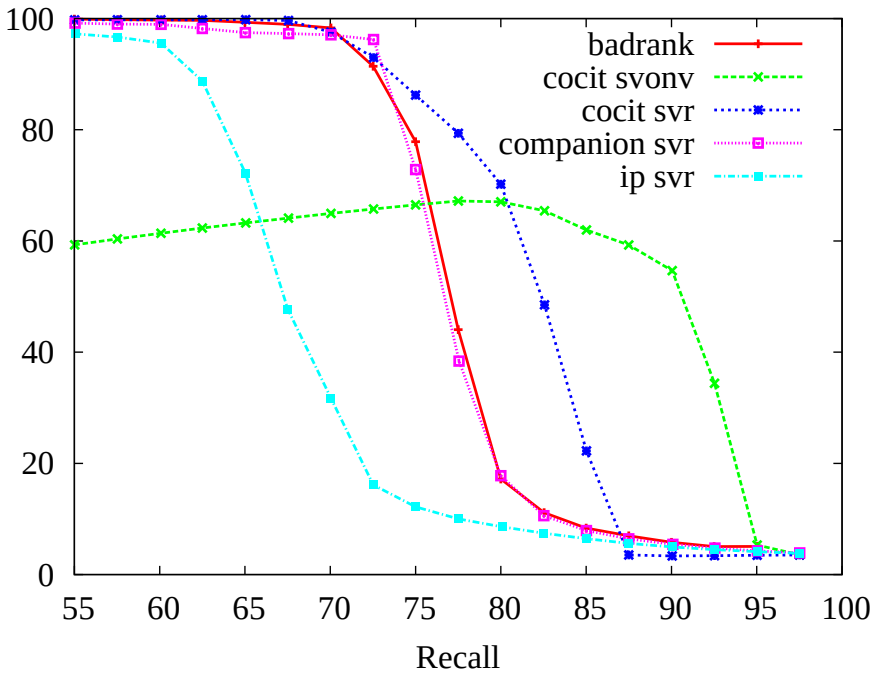
- 3-fold crossvalidation
- repeated 5 times with random splits
- Large variance in 15 results
- Averaging by interpolations of precision for all recall values



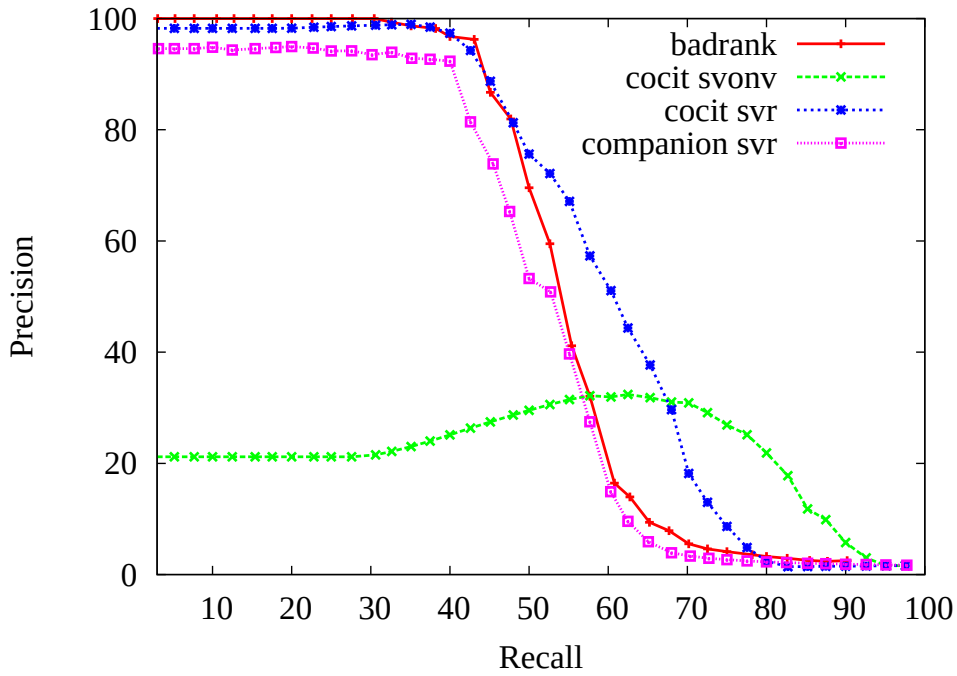
Results on the .de domain



Results on the search.ch data



Original sample



Reduced sample with unique IPs

Conclusion

- Link similarity based single feature classification
  - Capable of learning the difference between spam and nonspam
  - Better precision at higher recall than trust/distrust propagation
- Need for better data set
  - Uncorrelated spam pages
  - Good quality trusted set
- Future work
  - Further similarity algorithms
  - Use as input to classifiers
  - Use similarity algorithms to propagate output of other predictors along links

Thank you!

- Károly Csalogány, cskaresz@ilab.sztaki.hu
- http://www.ilab.sztaki.hu/websearch