Methods for Web Spam Filtering

Ph.D. Thesis Summary

Károly Csalogány

Supervisor: András A. Benczúr Ph.D.

Eötvös Loránd University
Faculty of Informatics
Department of Information Sciences

Informatics Ph.D. School
János Demetrovics D.Sc.

Foundations and Methods of Informatics Ph.D. Program
János Demetrovics D.Sc.

Budapest, 2009

1 Introduction

Identifying and preventing spam was cited as one of the top challenges in web search engines in a 2002 paper [12]. Amit Singhal, principal scientist of Google Inc. estimated that the search engine spam industry had a revenue potential of $4.5 billion in year 2004 if they had been able to completely fool all search engines on all commercially viable queries [20]. Due to the large and ever increasing financial gains resulting from high search engine ratings, it is no wonder that a significant amount of human and machine resources are devoted to artificially inflating the rankings of certain web pages. The birth of the highly successful PageRank algorithm [17] was indeed partially motivated by the easy spammability of the simple in-degree count; its variants [14, 9, 11, 3, 23, 2, and many others] proved successful in fighting search engine spam.

Web spam is an expensive operation requiring the registration and operation of diverse domain names and IP ranges. Unlike for email spam where the target is to reach the user mailbox, Web spammers also compete against each other in obtaining high search engine rankings for their target pages. Web spamming is hence a professional business with the purpose of financial gains.

There is no sound definition of Web spam in the literature. Our techniques should instead primarily be justified by improved user satisfaction with search engine hit lists. The fact that there is no clear boundary between honest pages and spam over the Web imposes a bias in the evaluation that we have to live together with.

Although web spam appears in increasingly sophisticated forms, basic spamming techniques fall into two categories, *content spam* and *link spam*. Content spam is manipulating the textual content of the page, link spam is the creation of artificial link structure that misleads the link-based ranking algorithms of search engines.

2 New Results

Spam and various means of search engine optimization seriously deteriorate search engine ranking results; as a response, building black and white lists belongs to the daily routine of search engine operation. Our goal is to extend this invaluable source of human annotation either to automatically demote pages similar to certain known spam pages or to suggest additional pages for the operator to be included in the blacklist.

In my thesis I give five methods for spam detection. *SpamRank* and *Link Based Similarity* are based on the linkage neighborhood and structure of the pages; *Commercial Intent Detection* and *Language Model Disagreement* on the textual content; finally *Graph Stacking* combines both sources.

Algorithm 2.1 Overall Structure of SpamRank Algorithm

for all Web pages i **do**
  Support$_{i,\cdot}$ $\leftarrow$ empty sparse vector of reals
Phase 1: Supporter Generation
generate nodes into vector Support$_{i,\cdot}$ that have high contribution in the PageRank of i
Phase 2: Penalization
for all Web pages i **do**
  give Penalty$_j$ based on irregular behavior of PageRank over Support$_{i,j}$
Phase 3: SpamRank as Personalized PageRank (PPR) over Penalties
SpamRank $\leftarrow$ PPR(Penalty)

Claim 1 SpamRank: a method based on supporter irregularity estimation

Spammers intend to increase the PageRank of certain spam pages by creating a large number of links pointing to them. We proposed a novel method based on the concept of personalized PageRank that detects pages with an undeserved high PageRank value without the need of white or blacklists or other means of human intervention of any kind. We assumed that spammed pages have a biased distribution of pages that contribute to the undeserved high PageRank value.

The overall structure of the SpamRank method is shown in Algorithm 2.1. In Phase 1 (Algorithm 2.2) we compute the approximate personalized PageRank vector of all pages j. We use the Monte Carlo approximation of [J1]. This algorithm under practically useful parameter settings computes a set of roughly 1,000 nodes i together with a weight Support$_{i,j}$. This weight can be interpreted as the probability that a random PageRank walk starting at j will end in i.

Algorithm 2.2 Phase 1: Finding supporters by Monte Carlo simulation

for all Web pages j **do**
  for $\ell = 1, \ldots, N = 1000$ **do**
    $t \leftarrow$ random value from geometric distribution with parameter ϵ
    $i \leftarrow$ endvertex of a random walk of length t starting at j
    Support$_{i,j} \leftarrow$ Support$_{i,j} i + 1/N$

For a fixed page i, penalties are defined by considering the PageRank histogram of all j with Support$_{i,j} > 0$ for pages that receive enough supporters. Pages with less than n_0 supporters (in our experiment $n_0 = 1000$) are ignored.

In Phase 2 of our algorithm we find the method of identifying irregularities in the PageRank distribution of a page's supporters We build on models of evolution of Web [1, 15] that predict a power law degree distribution. The distribution of PageRank behaves very similar to the in-degree and experiments [18, 8] showed that PageRank should have power law distribution in the entire web and as well as in

Algorithm 2.3 Phase 2:SpamRank Penalty Calculation for Web pages

Initialize vector Penalty by all 0
for all Web pages i with at least n_0 supporters j with nonzero $\text{Support}_{i,j}$ **do**
  $\rho \leftarrow$ regularity of the supporters of i
  if $\rho < \rho_0$ **then**
    for all Web pages j with $\text{Support}_{i,j} > 0$ **do**

$$\text{Penalty}_j \leftarrow \text{Penalty}_j + \begin{cases} (\rho_0 - \rho) & \{\text{Variant I}\} \\ (\rho_0 - \rho) \cdot \text{Support}_{i,j} & \{\text{Variant II}\} \end{cases}$$

    {we use $\rho_0 = 0.85$}
for all Web pages j **do**
  if $\text{Penalty}_j > 1$ **then**
    $\text{Penalty}_j \leftarrow 1$

smaller neighborhood. The key element of our method is that the neighborhood of a spam page will look different from an honest one since the neighborhood of a link spam will consist of a large number of artificially generated links that came from similar objects.

As a measure of regularity of supporters in Phase 2 we use a simple approach to test its fit to a power law distribution. We split pages into buckets by PageRank; we let bucket boundary values grow exponentially. If the PageRank values follow a power law distribution, then the logarithm of the theoretical size of bucket k is linear in k. In our algorithm we use the Pearson correlation coefficient between the index and the logarithm of the count within the bucket as the measure of regularity. Given such a measure $\rho \leq 1$ where $\rho = 1$ means perfect regularity, we proceed by penalizing all the supporter pages proportional to $(\rho_0 - \rho)$ if the measure is below a threshold ρ_0.

Finally in Phase 3 we define *SpamRank* as PageRank personalized on the vector of penalties.

SpamRank is one of the earliest results on web spam filtering. It appeared in [C1], accepted for publication in Information Retrieval [J2]. The results represent a joint work with András Benczúr, Tamás Sarlós, and Máté Uher. The conference version is cited by 54 publications. My contribution is the evaluation part, the algorithmic details and ideas on feature generation.

Claim 1.1 The SpamRank method [J2, C1]

Our experiments demonstrated that SpamRank is indeed capable of differentiating among spam and non-spam pages. The method was tested on a 31-million page crawl of the `.de` domain with a manually classified 1000-page stratified random sample with bias towards large PageRank values.

Claim 2 Link-based similarity

We investigated the usability of similarity search in fighting Web spam based on the assumption that an unknown spam page is more similar to certain known spam pages than to honest pages. In order to be successful, search engine spam never appears in isolation: we observe link farms and alliances for the sole purpose of search engine ranking manipulation. The artificial nature and strong inside connectedness however gave rise to successful algorithms to identify search engine spam. One example is trust and distrust propagation, an idea originating in recommender systems and P2P networks, that yields spam classificators by spreading information along hyperlinks from white and blacklists. While most previous results use PageRank variants for propagation, we form classifiers by investigating similarity top lists using the following similarity measures:

- *Co-citation* $\mathrm{coc}(u, v)$ is defined as the number of common in-neighbors of u and v.
- *Companion*: The Companion algorithm [7] builds the 2-step alternating neighborhood of the given vertex; then performs the HITS authority computation [6] and returns the top authorities.
- *SimRank* was suggested by Jeh and Widom [13] as a measure of link-based similarity of two nodes in the graph. The basic idea is that two nodes are similar if they are pointed by similar nodes. SimRank is defined by the following equation similarly to the PageRank power iteration:

$$\mathrm{Sim}^{(0)}(u_1, u_2) = 1 \text{ if } u_1 = u_2 \text{ and } 0 \text{ otherwise} \tag{1}$$

$$\mathrm{Sim}^{(i)}(u_1, u_2) = \begin{cases} (1-c) \cdot \frac{\sum \mathrm{Sim}^{(i-1)}(v_1, v_2)}{d^-(u_1) \cdot d^-(u_2)} & \text{if } u_1 \neq u_2 \\ 1 & \text{if } u_1 = u_2, \end{cases} \tag{2}$$

  where the summation is for all pairs $(v_1, u_1) \in E, (v_2, u_2) \in E$.
- We use *Singular Value Decomposition* for nearest neighbor search after a low dimensional projection of the adjacency matrix of the host graph. We represent host u by row u of $V_t \Sigma_t$ and measure similarity as the Eucledian distance in this t dimensional space.

For every unknown pages, we define our features as follows:

- Spam Ratio (SR): fraction of the number of spam within labeled spam and honest pages, $s/(s+h)$.
- Spam over Non-spam (SON): number of spam divided by number of honest pages in the top list, s/h.

- Spam Value Ratio (SVR): sum of the similarity values of spam pages divided by the total similarity value of labeled spam and honest pages under the appropriate similarity function, $s^*/(s^* + h^*)$.
- Spam Value over Non-spam Value (SVONV): similarity value sum for spam divided by same for honest, s^*/h^*.

These results appeared in [C3] and represent a joint work with András Benczúr and Tamás Sarlós. My contribution is the evaluation, design and efficient implementation of the similarity algorithms.

Claim 2.1 Spam filtering based on linkage similarity [C3]

As our main result, we claim that similarity based single features perform better than trust or distrust propagation based single features at higher recall values. Cocitation performs best among similarity measures.

The method is tested over two datasets previously used to measure spam filtering algorithms: The 31-million page crawl of the `.de` domain with a manually classified 1000-page stratified random sample with bias towards large PageRank values and the `.ch` domain data in courtesy of the `search.ch` engine [19].

Claim 3 Graph Stacking

We compared a wide range of semi-supervised learning techniques for Web spam filtering. Semi-supervised learning based methods has the assumption that the label of a node in a graph is similar to those of its neighbors. Our experiments demonstrated that stacked graphical learning in combination with graph node similarity methods improve classification accuracy. The result show a very good performance of co-citation and little actual use of the neighborhood beyond two steps in the graph.

Besides spam filtering, we evaluated these methods on a different problem. We applied the methods on a telephone call graph where the task is to predict telephone churn i.e. users who are going to cancel their telephone line. This problem is quite similar to spam detection, since we can use the telephone graph instead of the web graph and demographic and traffic data instead of web content. Our assumption was that the label (spam and churned, respectively) of a node in a graph is similar to those of its neighbors. The same method may work because churn occurs similarly in bursts in groups of a social network.

For a given unknown node u and edge weight function w (that may be in or out-degree, co-citation, PageRank etc.), our algorithm selects the k largest weight neighbors of u to generate a new feature based on the known spam and honest hosts in this set. As in previous result we extract four different features (Spam Ratio, Spam over Non-spam, Spam Value Ratio, Spam Value over Non-spam Value) from this set of size k or possibly less if u has less than k neighbors. Each element v is either classified as spam with

weight $p(v)$ or else labeled spam or non-spam; in these cases we let $p(v)$ be 0 and 1, respectively. Let s and h be the sum of $p(v)$ and $1 - p(v)$ in the set; remember $s + h < k$ is possible. We define a weighted version s^* and h^* as the sum of $w(uv) \cdot p(v)$ and $w(uv) \cdot (1 - p(v))$. We add the new feature defined by either of the above to the existing ones and repeat the classification process with the extended feature set. k

The results appeared in [C4] and represent a joint work with András Benczúr, Dávid Siklósi and László Lukács. My main contribution is the definition and implementation of the stacked features and graph similarity measures.

Claim 3.1 Using graph stacking features for spam detection [C4]

Our experiments demonstrated that stacked graphical learning in combination with graph node similarity methods improve classification accuracy of spam detection.

The method is tested on the WEBSPAM-UK-2006 [4] dataset.

Claim 3.2 Using graph stacking features for telco churn detection [C4]

Our experiments demonstrated that stacked graphical learning in combination with graph node similarity methods improve classification accuracy of telco churn detection.

The method os tested on a dataset from a small Hungarian landline telephone service provider.

Claim 4 Language Model Disagreement

Mishne et al. [16] demonstrated that the distribution of words (a unigram language model) is a strong feature for telling legitimate and spam blog comments apart. We applied this language model disagreement technique for spam filtering in the scale of the entire Web, both in terms of algorithmic efficiency and quality.

By using Interpolated Aggregate Smoothing as in [16], the language model for document D has the form

$$p(w|D) = \lambda \frac{tf(w, D)}{\sum_{v \in D} tf(v, D)} + (1 - \lambda) \frac{tf(w, C)}{\sum_{v \in C} tf(v, C)} \tag{3}$$

where C is the text of the entire corpus and w is a word.

As in [16], our key ingredient is the Kullback-Leibler divergence (KL) between the unigram language model of the target and source pages of hyperlinks.

$$KL(A \,||D) = \sum_{w} p\,(w|A) \log \frac{p\,(w|A)}{p\,(w|D)}, \tag{4}$$

Unfortunately, it is infeasible to compute KL for all pairs of documents connected by hyperlinks. Two computationally easier tasks are to compare each *anchor text* to (i) the document containing it (as in [16]) and to (ii) the document pointed by it. We observe best performance when we extend the anchor text by a few neighboring words to properly handle very short anchor such as `"here"`; we obey segment boundaries defined by HTML and punctuation.

As suggested in [16] the distribution of KL values is a mixture of Gaussians. KL will have normal distribution over the documents if all anchor text behave the same since we sum random variables that correspond to words and the words themselves have sufficient independence to yield a normally distributed sum. If however we have fair and malicious hyperlinks, the two categories will be biased towards the smaller and the larger values, respectively.

In our algorithm we form the set of suspicious hyperlinks with KL value (4) above a threshold. We obtain NRank by feeding suspicious edges into PageRank by keeping edge above 7. Results are useful in the range 4-7; increased values of the threshold give NRank results farther from original PageRank and improve recall.

These results appeared in [P1] and represent a joint work with András Benczúr and István Bíró. My contribution to the results is the evaluation and the solutions for the large scale implementation.

Claim 4.1 Applicability of hyperlink downweighting by means of language model disagreement [P1]

We demonstrated the applicability of hyperlink downweighting by means of language model disagreement. The method filters out hyperlinks with no relevance to the target page without the need of white and blacklists or human interaction. Our method is capable of fighting against various forms of nepotism such as common maintainers, ads, link exchanges or misused affiliate programs.

The method was tested on a 31-million page crawl of the `.de` domain with a manually classified 1000-page stratified random sample with bias towards large PageRank values.

Claim 5 Commercial Intent

As noticed by Gyöngyi and Garcia-Molina [10], most spammers just want financial gain from their activities. In contrast to previous content-based spam features such as distribution, entropy, compressibility targeting the templatic nature of machine generated pages, our features hence try to capture the semantics of spam content. By utilizing external classifiers we also enrich the available training and test data.

We investigate the following features for Web spam detection:

- Online Commercial Intention (OCI) value assigned to an URL in a Microsoft adCenter Labs Demonstration.

- The Yahoo! Mindset classification of Web pages as either commercial or non-commercial.
- Google AdWords advertisement keyword suggestions for the sites as well as keyword scores.
- The distribution of Google AdSense ads over pages of a site.
- A measure for queries based on spammer success in obtaining high rank for the particular query, measured on our own search engine.

We merge our features with the publicly available ones of [5] and then classify by the C4.5 implementation of the machine learning toolkit Weka [22].

These results appeared in [C5] and represent a joint work with András Benczúr István Bíró and Tamás Sarlós. The methods described here achieved the highest F-measure in the Web Spam Challenge competition in 2007 [21]. My contribution to the results is the evaluation part, feature combination, search engine success and Google AdWords features.

Claim 5.1 Commercial intent features improve spam classification accuracy [C5]

Inclusion of our new features improves the spam classification accuracy. We run our tests on the WEBSPAM-UK2006 dataset as a standard means of measuring the performance of Web spam detection algorithms. Our features improve the classification accuracy of the publicly available WEBSPAM-UK2006 features by 3%.

3 Other Results

Besides spam filtering, I have results in other areas related to web search:

- Develepment and experiments over a Hungarian search engine [P2, H1, H2, C9].
- Cross-modal information retrieval using text and image features [J5, J6, C8].
- Spectral clustering of large graphs and its applications in recommendation systems and telecommunication networks [J4, C6, C7, H3].

References

[1] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the word-wide web. *Physica A*, 281:69–77, 2000.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of web spam. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2006.

[3] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. SpamRank – Fully automatic link spam detection. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with WWW2005*, 2005. To appear in *Information Retrieval*.

[4] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[5] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, 2007.

[6] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[7] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. In *Proceedings of the 8th World Wide Web Conference (WWW)*, pages 1467–1479, 1999.

[8] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, 2002.

[9] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International World Wide Web Conference (WWW)*, pages 403–412, 2004.

[10] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[11] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 576–587, Toronto, Canada, 2004.

[12] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[13] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 538–543, 2002.

[14] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th International World Wide Web Conference (WWW)*, pages 640–651, New York, NY, USA, 2003. ACM Press.

[15] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1–10, 2000.

[16] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005.

[17] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford University, 1998.

[18] G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*, pages 330–339, Singapore, 2002.

[19] Räber Information Management GmbH. The Swiss search engine, http://www.search.ch/, 2006.

[20] A. Singhal. Challenges in running a commercial search engine. In *IBM Search and Collaboration Seminar 2004*. IBM Haifa Labs, 2004.

[21] 2007. `http://airweb.cse.lehigh.edu/2007/ presentations/challenge_results.p`

[22] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[23] B. Wu, V. Goel, and B. D. Davison. Topical TrustRank: Using topicality to combat web spam. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, Edinburgh, Scotland, 2006.

Publications

Journal papers

[J1] Dániel Fogaras, Balázs Rácz, Károly Csalogány, and Tamás Sarlós. Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments. *Internet Mathematics* 2(3):333-358, 2005.

[J2] András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máte Uher. SpamRank – Fully Automatic Link Spam Detection. To appear in *Information Retrieval*.

[J3] M. Kurucz, A. Benczúr, K. Csalogány, and L. Lukács. Spectral Clustering in Social Networks. *Lecture Notes In Artificial Intelligence*, pages 1–20, 2009.

[J4] M. Kurucz, D. Siklósi, L. Lukács, A. A. Benczúr, K. Csalogány, and A. Lukács. Telephone call network data mining: A survey with experiments. In Bolyai Society Mathematical Studies, Vol. 18., B. Bollobás, R. Kozma, D. Miklós, eds., *Handbook of Large-Scale Random Networks, published by Springer Verlag in conjunction with the Bolyai Mathematical Society of Budapest*, 2008.

[J5] A. Benczúr, I. Bíró, M. Brendel, K. Csalogány, B. Daróczy, and D. Siklósi. Multimodal Retrieval by Text–Segment Biclustering. In *ADVANCES IN MULTILINGUAL AND MULTIMODAL INFORMATION RETRIEVAL*. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers. Lecture Notes in Computer Science Vol. 5152 (2008).

[J6] P. Schönhofen, A. A. Benczúr, I. Bíró, and K. Csalogány. Cross-language retrieval with wikipedia. In *ADVANCES IN MULTILINGUAL AND MULTIMODAL INFORMATION RETRIEVAL*. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers. Lecture Notes in Computer Science Vol. 5152 (2008).

Conference papers

[C1] András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máte Uher. SpamRank – Fully Automatic Link Spam Detection. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with WWW2005*, 2005.

[C2] Tamás Sarlós, András A. Benczúr, Károly Csalogány, Dániel Fogaras, and Balázs Rácz. To Randomize or Not To Randomize: Space Optimal Summaries for Hyperlink Analysis In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 297-306, 2006.
Full version available at http://datamining.sztaki.hu/www/index.pl/publications-en.

[C3] A. A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight web spam. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with SIGIR2006*, 2006.

[C4] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn. In *Graph Labeling Workshop in conjunction with ECML/PKDD 2007*, 2007.

[C5] A. A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. Web spam detection via commercial intent analysis. In *Proceedings of the 3th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with WWW2007*,2007.

[C6] M. Kurucz, A. A. Benczúr, and K. Csalogány. Methods for large scale svd with missing values. In *KDD Cup and Workshop in conjunction with KDD 2007*, 2007.

[C7] M. Kurucz, A. A. Benczúr, K. Csalogány, and L. Lukács. Spectral clustering in telephone call graphs. In *WebKDD/SNAKDD Workshop 2007 in conjunction with KDD 2007*, 2007.

[C8] A. Benczúr, I. Bíró, M. Brendel, C. Károly, B. Daróczy, and D. Siklósi. Cross-modal retrieval by text and image feature biclustering. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary,2007.

[C9] P. Schönhofen, A. A. Benczúr, I. Bíró, and K. Csalogány. Performing cross-language retrieval with Wikipedia. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, Sept. 2007.

Conference posters

[P1] A. A. Benczúr, I. Bíró, and K. Csalogány. Detecting nepotistic links by language model disagreement. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, 2006.

[P2] András A. Benczúr, Károly Csalogány, Dániel Fogaras, Eszter Friedman, Tamás Sarlós, Máte Uher, and Eszter Windhager. Searching a Small National Domain – a Preliminary Report. In *Poster Proceedings of the 12th International World Wide Web Conference (WWW)*, 2003.

[P3] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. On the Feasibility of Low-rank Approximation for Personalized PageRank. In *Poster Proceedings of the 14th International World Wide Web Conference (WWW)*, pages 972–973, 2005.

Publications in Hungarian

[H1] András A. Benczúr, Károly Csalogány, Dániel Fogaras, Eszter Friedman, Balázs Rácz, Tamás Sarlós, Máte Uher, and Eszter Windhager. Magyar nyelvű tartalom a világhálón (Hungarian Content on the WWW). *Információs Társadalom és Trendkutató Központ Kutatási Jelentés* 26:48-55, 2004.

[H2] András A. Benczúr, István Bíró, Károly Csalogány, Balázs Rácz, Tamás Sarlós, and Máte Uher. PageRank és azon túl: Hiperhivatkozások szerepe a keresésben (PageRank and Beyond: The Role of Hyperlinks in Search). To appear in *Magyar Tudomány*, 2006.

[H3] Miklós Kurucz, László Lukács, Dávid Siklósi, András A. Benczúr, Károly Csalogány, András Lukács Kapcsolatok és távolságok: a hazai vezetékes hívás-szokások elemzése. (Contacts and Distances: Analysis of Hungarian Landline Telephone Calls). *Magyar Tudomány*, 2009/6.