

István BÍRÓ

PERSONAL DATA

ADDRESS: 73 Swallows Croft, Reading, Berkshire, RG1 6EH
PHONE: +44 750 3794063
EMAIL: istvan.m.biro@gmail.com
HOMEPAGE: <http://www.ilab.sztaki.hu/~ibiro>

WORK EXPERIENCE

- AUG 2009 – PRESENT | Software Engineer at [CLEARSWIFT](#), Theale, United Kingdom
I am responsible for maintaining and developing back end components of different security products on Windows and Linux multi-threaded platforms. These involve researching and developing state of the art localisation methods, file content analysers, image and text classification and clustering algorithms used in special applications.
- SEP 2006 – JUL 2009 | Junior Researcher at [COMPUTER AND AUTOMATION RESEARCH INSTITUTE, HUNGARIAN ACADEMY OF SCIENCES](#), Budapest, Hungary
I took part in various research projects related to Data Mining, Machine learning and Natural Language Processing. My significant contributions are summarized as follows.
1. LARGE SCALE WEB DOCUMENT CLASSIFICATION WITH LATENT TOPIC MODELS. The research primarily focused on three issues: applying Latent Dirichlet Allocation to direct classification, scaling the Gibbs sampling algorithm responsible for the Bayesian inference to be able to process very large web corpora and modifying the Bayesian inference network to exploit the hyperlinked environment. The proposed methods were successfully applied to web spam filtering as well. Results were published in [2, 3, 4, 5, 6, 7, 15].
2. RESEARCH WITH STATISTICAL LANGUAGE MODELS. The research can be classified into 3 main topics: how to exploit the similarity information between words in order to improve the prediction performance of language models, how to apply language modeling approaches for web spam filtering and how to improve cross-language information retrieval with the help of Wikipedia. Results were published in [8, 9, 10, 14].
- MAR 2008 – JUL 2008 | Visitor researcher at NATIONAL UNIVERSITY OF THE SOUTH, Bahia Blanca, Argentina
I worked on comparing various latent topic models (Latent Semantic Analysis, Latent Dirichlet Allocation) in terms of web document classification performance.

EDUCATION

- 2003 – 2009 Ph.D. in COMPUTER SCIENCE with *summa cum laude*
Eötvös Loránd University, Budapest
Thesis: “[Document classification with Latent Dirichlet Allocation](#)”
- 1998 – 2002 B.Sc. in ENGINEERING AND MANAGEMENT
Budapest Technical College, Budapest
- 1996 – 2001 M.Sc. in PHYSICS with *honours*
Budapest University of Technology and Economics, Budapest

TECHNICAL EXPERTISE

- Software Development: C/C++, Bash, AWK, Java, Matlab, Python, SQL
Operating systems: Linux, Windows
Internet Technologies: XML, HTML, HTTP, TCP/IP
Developing Environments: Eclipse SDK, Netbeans, Microsoft Visual Studio
Miscellaneous: STL, Boost, CVS/SVN, WEKA, Agile/Scrum, Condor, multi-threading, openMP, \LaTeX

GRANTS, AWARDS

- 2009 6. place on [KDDCup 2009](#)
2007 1. place on [AIRWEB 2007 WebSpam Challenge](#)
2001 Excellent diploma award of Pro Pogressio Foundation
2000 Aschner Lipót fellowship (General Electric Lighting Tungstram Rt.)
1999 and 2000 2. place on György Hajós National Mathematics Competition

LANGUAGES

- HUNGARIAN: Native
ENGLISH: Fluent
GERMAN: Basic Knowledge

HOBBIES

Skiing, cycling, chess, strategic board games, travel, film, music.

PUBLICATIONS

- [1] M. Kurucz, I. Bíró, D. Siklósi, P. Csizsek, Z. Fekete, R. Iwatt, T. Kiss, A. Szabó. [KDD CUP 2009@ BUDAPEST: FEATURE PARTITIONING AND BOOSTING](#). In *Journal of Machine Learning Research special issue on KDD Cup 2009*, 2009
- [2] István Bíró, Jácint Szabó, András A. Benczúr, and Dávid Siklósi. [LINKED LATENT DIRICHLET ALLOCATION IN WEB SPAM FILTERING](#). In *Proceedings of the 5th international workshop on Adversarial Information Retrieval on the Web*, 2009.
- [3] István Bíró, and Jácint Szabó. [LATENT DIRICHLET ALLOCATION FOR AUTOMATIC DOCUMENT CATEGORIZATION](#). In *Proceedings of the 19th European Conference on Machine Learning and 12th Principles of Knowledge Discovery in Databases*, 2009.
- [4] István Bíró, Jácint Szabó, András A. Benczúr, and Anna. G. Maguitman. [A COMPARATIVE ANALYSIS OF LATENT VARIABLE MODELS FOR WEB PAGE CLASSIFICATION](#). *LA-WEB*, pages 23–28. IEEE Computer Society, 2008.
- [5] István Bíró, Jácint Szabó, András A. Benczúr. [LATENT DIRICHLET ALLOCATION IN WEB SPAM FILTERING](#). In *Proceedings of the 4rd international workshop on Adversarial Information Retrieval on the Web*, 2008.
- [6] D. Siklósi, A. A. Benczúr, Z. Fekete, M. Kurucz, I. Bíró, A. Pereszlényi, S. Rácz, A. Szabó, and J. Szabó. [WEB SPAM HUNTING @ BUDAPEST](#). In *Proceedings of the 4rd international workshop on Adversarial Information Retrieval on the Web*, 2008.
- [7] A. Benczúr, D. Siklósi, J. Szabó, I. Bíró, Z. Fekete, M. Kurucz, A. Pereszlényi, S. Rácz, and A. Szabó. [WEB SPAM: A SURVEY WITH VISION FOR THE ARCHIVIST](#). In *8th International Web archiving workshop.*, 2008.
- [8] Z. Szamonek and I. Bíró. [SIMILARITY BASED SMOOTHING IN LANGUAGE MODELING](#). *Acta Cybernetica*, 18(2):303–314, 2007.
- [9] I. Bíró, C. Szepesvári, and Z. Szamonek. [SEQUENCE PREDICTION EXPLOITING SIMILARITY INFORMATION](#). *IJCAI 2007: Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1576–1581, 2007.
- [10] P. Schönhofen, I. Bíró, A. A. Benczúr, and K. Csalogány. [PERFORMING CROSS LANGUAGE RETRIEVAL WITH WIKIPEDIA](#). In *Working Notes for the CLEF 2007 Workshop*, 2007.
- [11] A. A. Benczúr, I. Bíró, M. Brendel, K. Csalogány, B. Daróczy, and D. Siklósi. [CROSS-MODAL RETRIEVAL BY TEXT AND IMAGE FEATURE BICLUSTERING](#). In *Working Notes for the CLEF 2007 Workshop*, 2007.
- [12] A. A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. [WEB SPAM DETECTION VIA COMMERCIAL INTENT ANALYSIS](#). In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial Information Retrieval on the Web*, pages 89–92, 2007.
- [13] E. P. Windhager, L. Tansini, I. Bíró, and D. Dubhashi. [ITERATIVE ALGORITHMS FOR COLLABORATIVE FILTERING WITH MIXTURE MODELS](#). In *proceedings of International Workshop on Intelligent Information Access (IIIA)*, 2006.
- [14] A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher. [DETECTING NEPOTISTIC LINKS BY LANGUAGE MODEL DISAGREEMENT](#). *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, pages 939–940, 2006.
- [15] István Bíró, Jácint Szabó, and András A. Benczúr. [LARGE SCALE LINK BASED LATENT DIRICHLET ALLOCATION FOR WEB DOCUMENT CLASSIFICATION](#). Manuscript.