67 Tulip Lane, Palo Alto, CA 94303, USA
+1 650.387.2651
marc@najork.org
marc.najork.org
najork
marcnajork

Marc Najork

Experience

Mar 2014–now | **Google Inc., Mountain View, CA**
Director, Research Engineering (Nov 2017–now)
Senior Staff Research Scientist (Mar 2014–Oct 2017)

- Since Jan. 2015, part of Google Research. Managing a group of ∼50 researchers. Current projects focus on a range of Information Retrieval topics: information extraction, document retrieval and ranking, semantic matching, document recommendation, document discovery, information quality and integrity. Virtually all of our research output is used in Google products. Past projects include infrastructure for highly scalable training of linear models; highly scalable topic modeling; scalable product recommendations; and ads personalization based on long-term user interest and user journeys. Overall revenue impact ∼$2B ARR.
- From Mar. 2014 until Dec. 2014, part of the Personal Search Infrastructure team. Worked on HappyHour, a processing and serving system for structured personal data.

Oct 2001–Mar 2014 | **Microsoft Corporation, Microsoft Research Silicon Valley, Mountain View, CA**
Principal Researcher (Mar 2006–Mar 2014)
Senior Researcher (Oct 2005–Mar 2006)
Researcher (Oct 2001–Sep 2005)

- Collaborated with Bing on various facets of social search.
- Explored link-based ranking techniques for web search results.
- Developed the Scalable Hyperlink Store, a specialized database giving extremely fast access to nodes and edges of the web graph induced by the Bing corpus.
- Explored techniques for identifying spam web pages.
- Consulted on large-scale web crawling for Bing.
- Principal contributor to PageTurner, a large-scale study of the evolution of web pages.
- Contributed to Boxwood, a distributed, scalable, and reliable B-Tree system.
- Four Microsoft Gold Star awards; three Microsoft Research Tech Transfer Awards; Microsoft Corporate Bench Program class of 2005; promoted to Partner level in 2009.

Oct 1993–Sep 2001 | **Digital Equipment Corporation (Compaq since 1998), Systems Research Center, Palo Alto, CA**
Manager, Programming Technology (Jun 2001–Sep 2001)
Senior Member of Technical Staff (Sep 1999–Jun 2001)
Software Principal Engineer (Oct 1993–Sep 1999)

- Managed a group of five Ph.D.-level researchers. Responsible for four projects.
- Principal contributor to Mercator, an extensible, high-performance web crawler. Mercator formed the web crawling component of AltaVista's Search Engine 3 product, which sold to over 1,200 enterprise customers, and it became the standard web crawler of the various AltaVista sites.
- Main contributor to JCAT, a Java-based algorithm animation system.
- Worked on tools and techniques for building distributed, collaborative, web-based applications.
- Designed and implemented Obliq-3D, a fast-turnaround, interactive 3D animation environment.

Education

January 1994 | **Ph.D. in Computer Science, University of Illinois at Urbana-Champaign**
Dissertation "*Programming in Three Dimensions*" supervised by Prof. Simon Kaplan.

May 1989 | **Diplom-Wirtschaftsinformatiker, Technical University of Darmstadt, Germany**
Program covers Computer Science, Mathematics, Business Administration, Economics, and Law.

Honors

2021 ACM SIGIR Academy

2020 IEEE Fellow. *"For contributions to web crawling and web data processing"*

2019 ACM Fellow. *"For contributions to web search and web science"*

2012 IEEE Senior Member

2008 ACM Distinguished Member

Patents

P30 | US Patent 10,970,293. *Ranking search result documents*. Inventors: Mike Bendersky, **Marc Alexander Najork**, Donald Metzler, Xuanhui Wang. Assignee: Google LLC. Filed 2019-08-26, issued 2021-04-06.

P29 | US Patent 10,824,630. *Search and retrieval of structured information cards*. Inventors: **Marc Alexander Najork**, Sujith Ravi, Michael Bendersky, Peter Shao-sen Young, Timothy Youngjin Sohn, Mingyang Zhang, Thomas Nelson, Xuanhui Wang. Assignee: Google LLC. Filed 2016-10-26, issued 2020-11-03.

P28 | US Patent 10,540,610. *Generating and applying a trained structured machine learning model for determining a semantic label for content of a transient segment of a communication*. Inventors: Jie Yang, Amr Ahmed, Luis Garcia Pueyo, Mike Bendersky, Amitabh Saikia, Marc-Allen Cartright, **Marc Alexander Najork**, MyLinh Yang, Hui Tan, Weinan Zhang, Vanja Josifovski, Alexander J. Smola. Assignee: Google LLC. Filed 2016-04-27, issued 2020-01-21.

P27 | US Patent 10,394,832. *Ranking search result documents*. Inventors: Mike Bendersky, **Marc Alexander Najork**, Donald Metzler, Xuanhui Wang. Assignee: Google LLC. Filed 2016-10-24, issued 2019-08-27.

P26 | US Patent 9,953,185. *Identifying query patterns and associated aggregate statistics among search queries*. Inventors: Mike Bendersky, Donald Metzler, **Marc Alexander Najork**, Dor Naveh, Vlad Panait, Xuanhui Wang. Assignee: Google LLC. Filed 2015-11-24, issued 2017-05-25.

P25 | US Patent 8,949,232. *Social network recommended content and recommending members for personalized search results*. Inventors: Timothy Harrington, Rajesh Shenoy, **Marc Najork**, Rina Panigrahy. Assignee: Microsoft Corporation. Filed 2011-10-04, issued 2015-02-03.

P24 | US Patent 8,856,112. *Considering document endorsements when processing queries*. Inventors: **Marc A. Najork**, Rina Panigrahy, Rajesh K. Shenoy. Assignee: Microsoft Corporation. Filed 2011-08-26, issued issued 2014-10-01.

P23 | US Patent 8,666,920. *Estimating shortest distances in graphs*. Inventors: **Marc A. Najork**, Sreenivas Gollapudi, Rina Panigrahy, Atish Das Sarma. Assignee: Microsoft Corporation. Filed 2011-08-18, issued 2014-03-04.

P22 | US Patent 8,392,366. *Changing number of machines running distributed hyperlink database*. Inventors: **Marc A. Najork**. Assignee: Microsoft Corporation. Filed 2006-08-29, issued 2013-03-05.

P21 | US Patent 8,209,305. *Incremental update scheme for hyperlink database*. Inventors: **Marc A. Najork**. Assignee: Microsoft Corporation. Filed 2007-10-25, issued 2012-06-26.

P20 | US Patent 7,962,510. *Using content analysis to detect spam web pages*. Inventors: **Marc Alexander Najork**, Dennis Craig Fetterly, Mark Steven Manasse, Alexandros Ntoulas. Assignee: Microsoft Corporation. Filed 2005-02-11, issued 2011-06-14.

P19 | US Patent 7,818,334. *Query dependent link-based ranking using authority scores*. Inventors: **Marc A. Najork**. Assignee: Microsoft Corporation. Filed 2007-10-22, issued 2010-10-19.

P18 | US Patent 7,792,854. *Query dependent link-based ranking*. Inventors: **Marc A. Najork**. Assignee: Microsoft Corporation. Filed 2007-10-22, issued 2010-09-07.

P17 | US Patent 7,783,671. *Deletion and compaction using versioned nodes*. Inventors: **Marc A. Najork**, Chandramohan A. Thekkath. Filed 2006-03-16, issued 2010-08-24.

P16 | US Patent 7,739,281. *Systems and methods for ranking documents based upon structurally interrelated information*. Inventors: **Marc A. Najork**. Assignee: Microsoft Corporation. Filed 2003-09-16, issued 2010-06-15.

P15 | US Patent 7,680,785. *Systems and methods for inferring uniform resource locator (URL) normalization rules*. Inventors: **Marc Alexander Najork**. Assignee: Microsoft Corporation. Filed 2005-03-25, issued 2010-03-16.

P14 | US Patent 7,627,777. *Fault tolerance scheme for distributed hyperlink database*. Inventors: **Marc Alexander Najork**. Assignee: Microsoft Corporation. Filed 2006-03-17, issued 2009-12-01.

P13 | US Patent 7,340,467. *System and method for maintaining a distributed database of hyperlinks*. Inventors: **Marc A. Najork**. Assignee: Microsoft Corporation. Filed 2003-04-15, issued 2008-03-04.

P12 | US Patent 7,139,747. *System and method for distributed web crawling*. Inventors: **Marc Alexander Najork**. Assignee: Hewlett-Packard Development Company. Filed 2000-11-03, issued 2006-11-21.

P11 | US Patent 7,082,438. *Algorithm for tree traversals using left links*. Inventors: **Marc A. Najork**, Chandramohan A. Thekkath. Assignee: Microsoft Corporation. Filed 2005-09-01, issued 2006-07-25.

P10 | US Patent 7,072,904. *Deletion and compaction using versioned nodes*. Inventors: **Marc A. Najork**, Chandramohan A. Thekkath. Assignee: Microsoft Corporation. Filed 2002-12-02, issued 2006-07-04.

P9 | US Patent 7,007,027. *Algorithm for tree traversals using left links*. Inventors: **Marc A. Najork**, Chandramohan A. Thekkath. Assignee: Microsoft Corporation. Filed 2002-12-02, issued 2006-02-28.

P8 | US Patent 6,952,730. *System and method for efficient filtering of data set addresses in a web crawler*. Inventors: **Marc Alexander Najork**, Clark Allan Heydon. Assignee: Hewlett-Packard Development Company. Filed 2000-06-30, issued 2005-10-04.

P7 | US Patent 6,910,077. *System and method for identifying cloaked web servers*. Inventors: **Marc A. Najork**. Assignee: Hewlett-Packard Development Company. Filed 2002-01-04, issued 2005-06-21.

P6 | US Patent 6,594,694. *System and method for near-uniform sampling of web page addresses*. Inventors: **Marc Alexander Najork**, Clark Allan Heydon, Michael Mitzenmacher, Monika H. Henzinger. Assignee: Hewlett-Packard Development Company. Filed 2000-05-12, issued 2003-07-15.

P5 | US Patent 6,377,984. *Web crawler system using parallel queues for queing data sets having common address and concurrently downloading data associated with data set in each queue*. Inventors: **Marc Alexander Najork**, Clark Allan Heydon. Assignee: Alta Vista Company. Filed 1999-11-02, issued 2002-04-23.

P4 | US Patent 6,351,755. *System and method for associating an extensible set of data with documents downloaded by a web crawler*. Inventors: **Marc Alexander Najork**, Clark Allan Heydon. Assignee: Alta Vista Company. Filed 1999-11-02, issued 2002-02-26.

P3 | US Patent 6,321,265. *System and method for enforcing politeness while scheduling downloads in a web crawler*. Inventors: **Marc Alexander Najork**, Clark Allan Heydon. Assignee: Alta Vista Company. Filed 1999-11-02, issued 2001-11-20.

P2 | US Patent 6,301,614. *System and method for efficient representation of data set addresses in a web crawler*. Inventors: **Marc Alexander Najork**, Clark Allan Heydon. Assignee: Alta Vista Company. Filed 1999-11-02, issued 2001-10-09.

P1 | US Patent 6,263,364. *Web crawler system using plurality of parallel priority level queues having distinct associated download priority levels for prioritizing document downloading and maintaining document freshness*. Inventors: **Marc Alexander Najork**, Clark Allan Heydon, Janet Lynn Wiener. Assignee: Alta Vista Company. Filed 1999-11-02, issued 2001-07-17.

For a list of non-issued published patent applications please visit my home page http://marc.najork.org

Conference and Workshop Papers

C79 | Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, **Marc Najork**. *WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning*. To appear in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 2021.

C78 | Zhen Qin, Le Yan, Honglei Zhuang, Yi Tay, Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, **Marc Najork**. *Are Neural Ranters Still Outperformed by Gradient Boosted Decision Trees?*. To appear in Proceedings of the 9th International Conference on Learning Representations, May 2021.

C77 | Rolf Jagerman, Weize Kong, Rama Kumar Pasumarthi, Zhen Qin, Michael Bendersky, **Marc Najork**. *Improving Cloud Storage Search with User Activity*. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pages 508–516, March 2021.

C76 | Jiecao Chen, Liu Yang, Karthik Raman, Michael Bendersky, Jung-Jung Yeh, Yun Zhou, **Marc Najork**, Danyang Cai, Ehsan Emadzadeh. *DiPair: Fast and Accurate Distillation for Trillion-Scale Text Matching and Pair Modeling*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 2925–2937, November 2020.

C75 | Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky and **Marc Najork**. *Beyond 512 Tokens: Siamese Multi-depth Transformer-based Hierarchical Encoder for Long-Form Document Matching*. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management, pages 1725–1734, October 2020.

C74 | Rama Kumar Pasumarthi, Honglei Zhuang, Xuanhui Wang, Michael Bendersky and **Marc Najork**. *Permutation Equivariant Document Interaction Network for Neural Learning to Rank*. In Proceedings of the 6th ACM SIGIR International Conference on the Theory of Information Retrieval, pages 145–148, September 2020.

C73 | Weize Kong, Michael Bendersky, **Marc Najork**, Brandon Vargo and Mike Colagrosso. *Learning to Cluster Documents into Workspaces Using Large Scale Activity Logs*. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2416–2424, August 2020.

C72 | Honglei Zhuang, Xuanhui Wang, Michael Bendersky, **Marc Najork**. *Feature Transformation for Neural Ranking Models*. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1649–1652, July 2020.

C71 | Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James B. Wendt, Qi Zhao, **Marc Najork**. *Representation Learning for Information Extraction from Form-like Documents*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6495–6504, July 2020.

C70 | Shuguang Han, Michael Bendersky, Przemek Gajda, Sergey Novikov, **Marc Najork**, Bernhard Brodowsky, Alexandrin Popescul. *Adversarial Bandits Policy for Crawling Commercial Web Content*. In Proceedings of the 2020 World Wide Web Conference, pages 407–417, April 2020.

C69 | Sebastian Bruch, Shuguang Han, Michael Bendersky, **Marc Najork**. *A Stochastic Treatment of Learning to Rank Scoring Functions*. In Proceedings of the 13th ACM International Conference on Web Search and Data Mining, pages 61–69, February 2020.

C68 | Ying Sheng, Nguyen Vo, James B. Wendt, Sandeep Tata, **Marc Najork**. *Migrating a Privacy-Safe Information Extraction System to a Software 2.0 Design*. In Proceedings of the 10th Conference on Innovative Data Systems Research, January 2020.

C67 | Qingyao Ai, Xuanhui Wang, Sebastian Bruch, Nadav Golbandi, Mike Bendersky, **Marc Najork**. *Learning Groupwise Multivariate Scoring Functions Using Deep Neural Networks*. In Proceedings of the 5th ACM SIGIR International Conference on the Theory of Information Retrieval, pages 85–92, October 2019.

C66 | Sebastian Bruch, Xuanhui Wang, Mike Bendersky, **Marc Najork**. *An Analysis of the Softmax Cross Entropy Loss for Learning-to-Rank with Binary Relevance*. In Proceedings of the 5th ACM SIGIR International Conference on the Theory of Information Retrieval, pages 75–78, October 2019.

C65 | Rama Kumar Pasumarthi, Sebastian Bruch, Xuanhui Wang, Cheng Li, Mike Bendersky, **Marc Najork**, Jan Pfeifer, Nadav Golbandi, Rohan Anil, Stephan Wolf. *TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank*. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2970–2978, August 2019.

C64 | Cheng Li, Mingyang Zhang, Mike Bendersky, Hongbo Deng, Don Metzler, **Marc Najork**. *Multi-view Embedding-based Synonyms for Personal Search*. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 575–584, July 2019.

C63 | Sebastian Bruch, Masrour Zoghi, Mike Bendersky, **Marc Najork**. *Revisiting Approximate Metric Optimization in the Age of Deep Neural Networks*. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1241–1244, July 2019.

C62 | Aman Agarwal, Xuanhui Wang, Cheng Li, Mike Bendersky, **Marc Najork**. *Addressing Trust Bias for Unbiased Learning-to-Rank*. In Proceedings of the 2019 World Wide Web Conference, pages 4–14, May 2019.

C61 | Shuguang Han, Bernhard Brodowsky, Przemek Gajda, Sergey Novikov, Mike Bendersky, **Marc Najork**, Robin Dua, Alexandrin Popescul. *Predictive Crawling for Commercial Web Content*. In Proceedings of the 2019 World Wide Web Conference, pages 627–637, May 2019.

C60 | Furkan Kocayusufoglu, Ying Sheng, Nguyen Ha Vo, James B. Wendt, Qi Zhao, Sandeep Tata, **Marc Najork**. *RiSER: Learning Better Representations for Richly Structured Emails*. In Proceedings of the 2019 World Wide Web Conference, pages 886–895, May 2019.

C59 | Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Mike Bendersky, Nadav Golbandi, **Marc Najork**. *Semantic Text Matching for Long-Form Documents*. In Proceedings of the 2019 World Wide Web Conference, pages 795-806, May 2019.

C58 | Qingyao Ai, Xuanhui Wang, Nadav Golbandi, Michael Bendersky, **Marc Najork**. *Learning Groupwise Scoring Functions using Deep Neural Networks*. In WSDM 2019 Workshop on Deep Matching in Practical Applications, February 2019.

C57 | Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, **Marc Najork**, Thorsten Joachims. *Estimating Position Bias without Intrusive Interventions*. In Proceedings of the 12 ACM International Conference on Web Search and Data Mining, pages 474–482, February 2019.

C56 | Manzil Zaheer, Amr Ahmed, Yuan Wang, Daniel Silva, **Marc Najork**, Yuchen Wu, Shibani Sanan, Surojit Chatterjee. *Uncovering Hidden Structure in Sequence Data via Threading Recurrent Models*. In Proceedings of the 12 ACM International Conference on Web Search and Data Mining, pages 186–194, February 2019.

C55 | Yu Sun, Luis Garcia Pueyo, James B. Wendt, **Marc Najork**, Andrei Broder. *Learning Effective Embeddings for Machine Generated Emails with Applications to Email Category Prediction*. In Proceedings of the 2018 IEEE International Conference on Big Data, pages 1846–1855, December 2018.

C54 | Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, **Marc Najork**. *Offline Comparison of Ranking Functions using Randomized Data*. In RecSys 2018 Workshop on Offline Evaluation for Recommender Systems, October 2018.

C53 | Xuanhui Wang, Cheng Li, Nadav Golbandi, Mike Bendersky, **Marc Najork**. *The LambdaLoss Framework for Ranking Metric Optimization*. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pages 1313–1322, October 2018.

C52 | Ying Sheng, Sandeep Tata, James B. Wendt, Jing Xie, Qi Zhao, **Marc Najork**. *Anatomy of a Privacy-Safe Large-Scale Information Extraction System Over Email*. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 734–743, August 2018.

C51 | Mike Bendersky, Xuanhui Wang, **Marc Najork**, Don Metzler. *Learning with Sparse and Biased Feedback for Personal Search*. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pages 5219–5223, July 2018.

C50 | John Foley, Mingyang Zhang, Mike Bendersky, **Marc Najork**. *Semantic Location in Email Query Suggestion*. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pages 977–980, July 2018.

C49 | Navneet Potti, James B. Wendt, Qi Zhao, Sandeep Tata, **Marc Najork**. *Hidden in Plain Sight: Classifying Emails Using Embedded Image Contents*. In Proceedings of the 2018 World Wide Web Conference, pages 1865–1874, April 2018.

C48 | Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, **Marc Najork**. *Position Bias Estimation for Unbiased Learning to Rank in Personal Search*. In Proceedings of the 11th ACM International Conference on Web Search and Data Mining, pages 610–618, February 2018.

C47 | Sandeep Tata, Alexandrin Popescul, **Marc Najork**, Mike Colagrosso, Julian Gibbons, Alan Green, Alexandre Mah, Michael James Smith, Divanshu Garg, Cayden Meyer, Reuben Kan. *Quick Access: Building a Smart Experience for Google Drive*. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1643–1651, August 2017.

C46 | Aston Zhang, Lluis Garcia-Pueyo, James B. Wendt, **Marc Najork**, Andrei Broder. *Email Category Prediction*. In Companion Proceedings of the 26th International Conference on World Wide Web, pages 495–503, April 2017.

C45 | Michael Bendersky, Xuanhui Wang, Don Metzler, **Marc Najork**. *Learning from User Interactions in Personal Search via Attribute Parameterization*. In Proceedings of the 10th ACM International Conference on Web Search and Data Mining, pages 791–799, February 2017.

C44 | Xuanhui Wang, Michael Bendersky, Donald Metzler, **Marc Najork**. *Learning to Rank with Selection Bias in Personal Search*. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 115–124, July 2016.

C43 | Omar Alonso, Catherine C. Marshall, **Marc Najork**. *Debugging a Crowdsourced Task with Low Inter-Rater Agreement*. In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, pages 101–110, June 2015.

C42 | Omar Alonso, Catherine C. Marshall, **Marc Najork**. *A Human-Centered Framework for Ensuring Reliability on Crowdsourced Labeling Tasks*. In Works in Progress and Demonstration Abstracts, an Adjunct to the Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing, November 2013.

C41 | Omar Alonso, Catherine C. Marshall, **Marc Najork**. *Are Some Tweets More Interesting Than Others? #HardQuestion*. In Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, October 2013.

C40 | Moises Goldszmidt, **Marc Najork**, Stelios Paparizos. *Boot-Strapping Language Identifiers for Short Colloquial Postings*. In Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases, Part II, pages 95–111, September 2013.

C39 | Nick Craswell, Bodo Billerbeck, Dennis Fetterly, **Marc Najork**. *Robust query rewriting using anchor data*. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining, pages 335–344, February 2013.

C38 | **Marc Najork**. *Detecting quilted web pages at scale*. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 385–394, August 2012.

C37 | Rina Panigrahy, **Marc Najork**, Yinglian Xie. *How user behavior is related to social affinity*. In Proceedings of the 5th ACM International Conference on Web Search and Data Mining, pages 713–722, February 2012.

C36 | **Marc Najork**, Dennis Fetterly, Alan Halverson, Krishnaram Kenthapadi, Sreenivas Gollapudi. *Of hammers and nails: an empirical comparison of three paradigms for processing large graphs*. In Proceedings of the 5th ACM International Conference on Web Search and Data Mining, pages 103–112, February 2012.

C35 | Bodo Billerbeck, Nick Craswell, Dennis Fetterly, **Marc Najork**. *Microsoft Research at TREC 2011 Web Track*. In Proceedings of the 20th Text REtrieval Conference, November 2011.

C34 | Nick Craswell, Dennis Fetterly, **Marc Najork**. *The Power of Peers*. In Proceedings of the 33rd European Conference on Advances in Information Retrieval, pages 497–502, April 2011.

C33 | Nick Craswell, Dennis Fetterly, **Marc Najork**. *Microsoft Research at TREC 2010 Web Track*. In Proceedings of the 19th Text REtrieval Conference, November 2010.

C32 | **Marc Najork**. *Querying the Web Graph (Invited Talk)*. In Proceedings of the 17th International Symposium on String Processing and Information Retrieval, pages 1–12, October 2010.

C31 | Atish Das Sarma, Sreenivas Gollapudi, **Marc Najork**, Rina Panigrahy. *A Sketch-Based Distance Oracle for Web-Scale Graphs*. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pages 401–410, February 2010.

C30 | Nick Craswell, Dennis Fetterly, **Marc Najork**, Stephen Robertson, Emine Yilmaz. *Microsoft Research at TREC 2009: Web and Relevance Feedback Track*. In Proceedings of the 18th Text REtrieval Conference, November 2009.

C29 | **Marc Najork**. *The Scalable Hyperlink Store*. In Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, pages 89–98, June 2009.

C28 | **Marc Najork**, Sreenivas Gollapudi, Rina Panigrahy. *Less is More: Sampling the Neighborhood Graph Makes SALSA Better and Faster*. In Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, pages 242–251, February 2009.

C27 | **Marc Najork**, Nick Craswell. *Efficient and Effective Link Analysis with Precomputed SALSA Maps*. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, pages 53–62, October 2008.

C26 | Frank McSherry, **Marc Najork**. *Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores*. In Proceedings of the 30th European Conference on IR Research, pages 414–421, April 2008.

C25 | Sreenivas Gollapudi, **Marc Najork**, Rina Panigrahy. *Using Bloom Filters to Speed Up HITS-Like Ranking Algorithms*. In Proceedings of the 5th International Workshop on Algorithms and Models for the Web-Graph, pages 195–201, December 2007.

C24 | **Marc Najork**. *Comparing the effectiveness of hits and salsa*. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, pages 157–164, November 2007.

C23 | **Marc Najork**, Hugo Zaragoza, Michael J. Taylor. *Hits on the web: how does it compare?*. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 471–478, July 2007.

C22 | Alexandros Ntoulas, **Marc Najork**, Mark Manasse, Dennis Fetterly. *Detecting spam web pages through content analysis*. In Proceedings of the 15th International Conference on World Wide Web, pages 83–92, May 2006.

C21 | Dennis Fetterly, Mark Manasse and **Marc Najork**. *Detecting Phrase-Level Duplication on the World-Wide Web*. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 170–177, August 2005.

C20 | John MacCormick, Nick Murphy, **Marc Najork**, Chandramohan A. Thekkath and Lidong Zhou. *Boxwood: Abstractions as the Foundation for Storage Infrastructure*. In Proceedings of the 6th Symposium on Operating Systems Design & Implementation, pages 105–120, December 2004.

C19 | Dennis Fetterly, Mark Manasse and **Marc Najork**. *Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages*. In Proceedings of the 7th International Workshop on the Web and Databases, pages 1–6, June 2004.

C18 | Dennis Fetterly, Mark Manasse and **Marc Najork**. *On the Evolution of Clusters of Near-Duplicate Web Pages*. In Proceedings of the 1st Latin American Web Congress, pages 37–45, November 2003.

C17 | Dennis Fetterly, Mark Manasse, **Marc Najork** and Janet Wiener. *A Large-Scale Study of the Evolution of Web Pages*. In Proceedings of the 12th International World Wide Web Conference, pages 669–678, May 2003.

C16 | Andrei Z. Broder, **Marc Najork**, Janet L. Wiener. *Efficient URL caching for World Wide Web crawling*. In Proceedings of the 12th International World Wide Web Conference, pages 679–689, May 2003.

C15 | **Marc Najork**. *Web-Based Algorithm Animation*. In Proceedings of the 38th Design Automation Conference, pages 506–511, June 2001.

C14 | **Marc Najork** and Janet L. Wiener. *Breadth-First Search Crawling Yields High-Quality Pages*. In Proceedings of the 10th International World Wide Web Conference, pages 114–118, May 2001.

C13 | Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher and **Marc Najork**. *On Near-Uniform URL Sampling*. In Proceedings of the 9th International World Wide Web Conference, pages 295–308, May 2000.

C12 | Allan Heydon and **Marc Najork**. *Performance Limitations of the Java Core Libraries*. In Proceedings of the ACM 1999 Conference on Java Grande, pages 35–41, June 1999.

C11 | Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher and **Marc Najork**. *Measuring Index Quality Using Random Walks on the Web*. In Proceedings of the 8th International World Wide Web Conference, pages 213–225, May 1999.

C10 | Marc H. Brown, **Marc A. Najork** and Roope Raisamo. *A Java-Based Implementation of Collaborative Active Textbooks*. In Proceedings of the 1997 IEEE Symposium on Visual Languages, pages 372–379, September 1997.

C9 | Marc H. Brown and **Marc A. Najork**. *Collaborative Active Textbooks: A Web-based Algorithm Animation System for an Electronic Classroom*. In Proceedings of the 1996 IEEE Symposium on Visual Languages, pages 266–275, September 1996.

C8 | Marc H. Brown and **Marc A. Najork**. *Distributed Active Objects*. In Proceedings of the 5th International World Wide Web Conference, pages 1037–1052, May 1996.

C7 | **Marc A. Najork** and Marc H. Brown. *A Library for Visualizing Combinatorial Structures*. In Proceedings of IEEE Visualization '94, pages 164–171, October 1994.

C6 | Marc H. Brown and **Marc A. Najork**. *Algorithm Animation Using 3D Interactive Graphics*. In Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology, pages 93–100, November 1993.

C5 | **Marc Najork** and Simon Kaplan. *Cube: Eine dreidimensionale visuelle Programmiersprache*. In Informatik, Wirtschaft, Gesellschaft, pages 340–345, October 1993.

C4 | **Marc A. Najork** and Simon M. Kaplan. *Specifying Visual Languages with Conditional Set Rewrite Systems*. In Proceedings of the 1993 IEEE Symposium on Visual Languages, pages 12–18, August 1993.

C3 | **Marc A. Najork** and Simon M. Kaplan. *A Prototype Implementation of the CUBE Language*. In Proceedings of the 1992 IEEE Workshop on Visual Languages, pages 270–272, September 1992.

C2 | **Marc A. Najork** and Simon M. Kaplan. *The CUBE Language*. In Proceedings of the 1991 IEEE Workshop on Visual Languages, pages 218–224, October 1991.

C1 | **Marc A. Najork** and Eric Golin. *Enhancing Show-and-Tell with a polymorphic type system and higher-order functions*. In Proceedings of the 1990 IEEE Workshop on Visual Languages, pages 215–220, October 1990.

Abstracts

A4 | **Marc Najork**. *Training On-Device Ranking Models from Cross-User Interactions in a Privacy-Preserving Fashion*. In Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, page 108, August 2018.

A3 | **Marc Najork**. *Using Machine Learning to Improve the Email Experience*. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, page 891, October 2016.

A2 | **Marc Najork**. *Social Search*. In Proceedings of the 14th International Conference on Web Engineering, pages 571–572, July 2014.

A1 | Marc H. Brown and **Marc A. Najork**. *Distributed Applets*. In CHI '97 Extended Abstracts on Human Factors in Computing, pages 204–205, March 1997.

Book Chapters

B2 | **Marc Najork**, Allan Heydon. *High-Performance Web Crawling*. Chapter 2 in James Abello, Panos M. Pardalos and Mauricio G.C. Resende (editors), *Handbook of Massive Data Sets*, Kluwer Academic Publishers, 2002.

B1 | Marc H. Brown and **Marc A. Najork**. *Algorithm Animation Using Interactive 3D Graphics*. Chapter 9 in John Stasko, John Domingue, Marc H. Brown and Blaine A. Price (editors), *Software Visualization – Programming as a Multimedia Experience*, MIT Press, 1998.

Journal Papers

J12 | Sandeep Tata, Navneet Potti, James Wendt, Lauro Beltrão Costa, **Marc Najork** and Beliz Gunel. *Glean: Structured Extractions from Templatic Documents*. Proceedings of the VLDB Endowment **14**(6):997–1005, February 2021.

J11 | Michael Whittaker, Nick Edmonds, Sandeep Tata, James B. Wendt and **Marc Najork**. *Online Template Induction for Machine-Generated Emails*. Proceedings of the VLDB Endowment **12**(11):1235–1248, August 2019.

J10 | Christopher Olston and **Marc Najork**. *Web Crawling*. Foundations and Trends in Information Retrieval **4**(3):175–246, 2010.

J9 | Brian D. Davison, **Marc Najork** and Tim Converse. *Adversarial Information Retrieval on the Web*. SIGIR Forum **40**(2):27–30, December 2006.

J8 | Dennis Fetterly, Mark Manasse and **Marc Najork**. *On the Evolution of Clusters of Near-duplicate Web Pages*. Journal of Web Engineering **2**(4):228–246, October 2004.

J7 | Dennis Fetterly, Mark Manasse, **Marc Najork** and Janet L. Wiener. *A Large-Scale Study of the Evolution of Web Pages*. Software: Practice & Experience **34**(2):213–237, February 2004.

J6 | Allan Heydon and **Marc Najork**. *Performance Limitations of the Java Core Libraries*. Concurrency: Practice & Experience **12**(6):363–373, May 2000.

J5 | Allan Heydon and **Marc Najork**. *Mercator: A Scalable, Extensible Web Crawler*. World Wide Web **2**(4):219–229, December 1999.

J4 | Marc H Brown and **Marc A. Najork**. *Collaborative Active Textbooks*. Journal of Visual Languages and Computing **8**(4):453–486, August 1997.

J3 | **Marc A. Najork**. *Programming in Three Dimensions*. Journal of Visual Languages and Computing **7**(2):219–242, June 1996.

J2 | **Marc A. Najork** and Marc H. Brown. *Obliq-3D: A High-Level, Fast-Turnaround 3D Animation System*. IEEE Transactions on Visualization and Computer Graphics **1**(2):175–193, June 1995.

J1 | Sharon Kuck, Roland John, Arnd Lewe and **Marc Najork**. *Roles and their role in posing recursive queries*. Information Systems **15**(2):173–186, 1990.

Technical Reports

T13 | Omar Alonso, Catherine Marshall and Marc Najork. *Crowdsourcing a Subjective Labeling Task: A Human-Centered Framework to Ensure Reliable Results*. MSR-TR-2014-91, Microsoft Research, June 2014.

T12 | **Marc Najork** and Allan Heydon. *High-Performance Web Crawling*. SRC Research Report 173, COMPAQ Systems Research Center, September 2001.

T11 | **Marc A. Najork** and Marc H. Brown. *Three-Dimensional Web-Based Algorithm Animations*. SRC Research Report 170, COMPAQ Systems Research Center, July 2001.

T10 | Marc H. Brown, Hannes Marais, **Marc A. Najork** and William E. Weihl. *Focus + Context Displays of Web Pages: Implementation Alternatives*. SRC Technical Note 1997-010, DIGITAL Systems Research Center, May 1997.

T9 | Marc H. Brown and **Marc A. Najork**. *Collaborative Active Textbooks: A Web-based Algorithm Animation System for an Electronic Classroom*. SRC Research Report 142, DIGITAL Systems Research Center, May 1996.

T8 | Marc H. Brown and **Marc A. Najork**. *Distributed Active Objects*. SRC Research Report 141 (paper & video), DIGITAL Systems Research Center, May 1996.

T7 | **Marc A. Najork**. *Obliq-3D Tutorial and Reference Manual*. SRC Research Report 129, DIGITAL Systems Research Center, December 1994.

T6 | **Marc A. Najork** and Marc H. Brown. *A Library for Visualizing Combinatorial Structures*. SRC Research Report 128 (paper & video), DIGITAL Systems Research Center, September 1994.

T5 | **Marc-Alexander Najork**. *Programming in Three Dimensions*. Technical Report UIUCDCS-R-93-1838, Department of Computer Science, University of Illinois, October 1993.

T4 | Marc H. Brown and **Marc A. Najork**. *Algorithm Animation Using 3D Interactive Graphics*. SRC Research Report 110 (paper & video), DIGITAL Systems Research Center, September 1993.

T3 | **Marc Najork**. *Funktionale, logik-basierte und objektorientierte Sprachstile und Wege zur Vereinheitlichung*. Thesis, Fachbereich Informatik, Technische Hochschule Darmstadt (Germany), March 1989.

T2 | **Marc Najork**. *Enhanced ER-Easy: A Database Scheme Designer*. Technical Report UIUCDCS-R-88-1464, Department of Computer Science, University of Illinois, May 1988.

T1 | Roland John, Sharon Kuck, Arnd Lewe and **Marc Najork**. *Roles and their Role in Posing Recursive Queries over the Universal Relation*. Technical Report UIUCDCS-R-88-1463, Department of Computer Science, University of Illinois, May 1988.

arXiv Papers

X13 | Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, **Marc Najork**. *Privacy-Adaptive BERT for Natural Language Understanding*. arXiv:2104.07504 [cs.CL], submitted on 2021-04-15.

X12 | Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, **Marc Najork**. *WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning*. arXiv:2103.01913 [cs.CV], submitted on 2021-03-02.

X11 | Nicholas Monath, Avinava Dubey, Guru Guruganesh, Manzil Zaheer, Amr Ahmed, Andrew McCallum, Gokhan Mergen, **Marc Najork**, Mert Terzihan, Bryon Tjanaka, Yuan Wang, Yuchen Wu. *Scalable Bottom-Up Hierarchical Clustering*. arXiv:2010.11821 [cs.LG], submitted on 2020-10-22.

X10 | Jiecao Chen, Liu Yang, Karthik Raman, Michael Bendersky, Jung-Jung Yeh, Yun Zhou, **Marc Najork**, Danyang Cai, Ehsan Emadzadeh. *DiPair: Fast and Accurate Distillation for Trillion-Scale Text Matching and Pair Modeling*. arXiv:2010.03099 [cs.CL], submitted on 2020-10-07.

X9 | Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, **Marc Najork**. *Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems: A Hybrid Approach*. arXiv:2010.01195 [cs.IR], submitted on 2020-10-02.

X8 | Abbas Kazerouni, Qi Zhao, Jing Xie, Sandeep Tata, **Marc Najork**. *Active Learning for Skewed Data Sets*. arXiv:2005.11442 [cs.LG], submitted on 2020-05-23.

X7 | Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, **Marc Najork**. *Beyond 512 Tokens: Siamese Multi-depth Transformer-based Hierarchical Encoder for Document Matching*. arXiv:2004.12297 [cs.IR], submitted on 2020-04-26.

X6 | Shuguang Han, Xuanhui Wang, Mike Bendersky, **Marc Najork**. *Learning-to-Rank with BERT in TF-Ranking*. arXiv:2004.08476 [cs.IR], submitted on 2020-04-17.

X5 | Rama Kumar Pasumarthi, Xuanhui Wang, Michael Bendersky, **Marc Najork**. *Self-Attentive Document Interaction Networks for Permutation Equivariant Ranking*. arXiv:1910.09676 [cs.IR], submitted on 2019-10-21.

X4 | Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, **Marc Najork** and Thorsten Joachims. *Estimating Position Bias without Intrusive Interventions*. arXiv:1812.05161 [cs.IR], submitted on 2018-12-12.

X3 | Rama Kumar Pasumarthi, Xuanhui Wang, Cheng Li, Sebastian Bruch, Michael Bendersky, **Marc Najork**, Jan Pfeifer, Nadav Golbandi, Rohan Anil and Stephan Wolf. *TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank*. arXiv:1812.00073 [cs.IR], submitted on 2018-11-30.

X2 | Qingyao Ai, Xuanhui Wang, Nadav Golbandi, Michael Bendersky and **Marc Najork**. *Learning Groupwise Scoring Functions Using Deep Neural Networks*. arXiv:1811.04415 [cs.IR], submitted on 2018-11-11.

X1 | Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky and **Marc Najork**. *Offline Comparison of Ranking Functions using Randomized Data*. arXiv:1810.05252 [cs.IR], submitted on 2018-10-11.

Popular Magazines

M2 | Marc H. Brown and **Marc A. Najork**. *Distributed Active Objects*. Dr. Dobb's Journal, pages 34–41, March 1997.

M1 | **Marc Najork**. *Visual Programming in 3-D*. Dr. Dobb's Journal, pages 18–31, December 1995.

Encyclopedia Entries

E3 | **Marc Najork**. *Web Crawler Architecture.* In Ling Liu and M. Tamer Özsu (editors), *Encyclopedia of Database Systems,* Springer, 2009.

E2 | Hugo Zaragoza and **Marc Najork**. *Web Search Relevance Ranking.* In Ling Liu and M. Tamer Özsu (editors), *Encyclopedia of Database Systems,* Springer, 2009.

E1 | **Marc Najork**. *Web Spam Detection.* In Ling Liu and M. Tamer Özsu (editors), *Encyclopedia of Database Systems,* Springer, 2009.

Editorial Boards

2007-2015 | **ACM Transactions on the Web (TWEB)**
Editor-in-Chief (2012–2015)
Associate Editor (2007–2011)

2008–2014 | **Communications of the ACM (CACM)**
Co-Chair of News Board

1996–2011 | **Journal of Visual Languages and Computing (JVLC)**
Associate Editor (2001–2011)
Visual Software Tools Editor (1996–2001)

Steering Committees

2020–now | **ACM Publications Board**

2017–now | **Conference on Design of Experimental Search and Information Retrieval Systems (DESIRES)**

2017–now | **BigData Innovators Gathering (BIG)**

2010-2018 | **ACM International Conference on Web Search and Data Mining (WSDM)** (chair from 2014–2018)

Conference & Workshop Committees

2022 WSDM (SPC)
2021 WWW **(program co-chair)** · SIGIR (SPC) · DESIRES (co-chair)
2020 WWW (SPC) · SIGIR (SPC) · BIG (co-chair)
2019 WSDM (SPC) · WWW (SPC) · SIGIR (short papers track co-chair) · KDD (SPC) · ICWE
2018 WSDM (SPC) · SIGIR · ASONAM · BIG
2017 WSDM (SPC) · ASONAM · BIG (co-chair)
2016 WSDM (SPC) · WWW
2015 WSDM (SPC) · CIKM (Industry track co-chair)
2014 –
2013 SIGIR (area chair)
2012 WSDM (SPC) · WWW · SIGIR (SPC)
2011 WSDM (tutorials co-chair) · WWW (PC & posters co-chair) · SIGIR (area chair)
2010 WSDM 2010 · WWW (Search track co-chair) · KDD (SPC)
2009 WSDM (SPC) · WWW (Industrial Practice & Experience track co-chair) · AIRWEB
2008 WSDM **(general chair)** · SIGIR · KDD · CIKM · AIRWEB · VLC
2007 WWW (tutorials & workshops co-chair) · AIRWEB · VLC
2006 WWW (Industrial Practice & Experience track co-chair) · SIGIR · AIRWEB (co-chair) · VLC · VL
2005 WWW · SIGIR · AIRWEB · VL
2004 WWW **(program co-chair)** · VLC
2003 WWW (Browsers and Tools track vice chair · IC · WAW · VLC
2002 WWW (Browsers and User Interfaces track deputy chair)
2001 WWW (Browsers and Tools track deputy chair)
2000 WWW · VL
1999 WWW · VL (PC & video chair)
1998 VL
1997 –
1996 VL