Jimmy Lin

Professor and David R. Cheriton Chair
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

E-mail: jimmylin@uwaterloo.ca
Homepage: https://cs.uwaterloo.ca/~jimmylin/
Google Scholar Profile: https://scholar.google.com/citations?user=0EWw1z8AAAAJ
DBLP Profile: https://dblp.org/pid/00/7739.html
ACL Anthology Profile: https://www.aclweb.org/anthology/people/j/jimmy-lin/

Prepared February 24, 2023

Contents

1 Personal Information **3**
1.1 Academic Appointments . . . 3
1.2 Education . . . 3
1.3 Employment . . . 3

2 Publications **4**
2.1 Books and Monographs . . . 4
2.2 Proceeding Volumes . . . 5
2.3 Book Chapters . . . 5
2.4 Journal Articles . . . 5
2.5 Conference Papers . . . 10
2.6 Conference Abstracts and Extended Abstracts . . . 31
2.7 Workshop Papers . . . 35
2.8 Tutorials . . . 41
2.9 TREC, TAC, and DUC Papers . . . 42
2.10 Notes and Other Articles . . . 45
2.11 Preprints and Technical Reports . . . 46

3 Patents **54**

4 Open-Source Software **54**

5 Prizes and Awards **57**

6 Presentations and Publicity **58**
6.1 Keynotes and Invited Talks . . . 58
6.2 Seminars and Other External Talks . . . 59
6.3 Internal Talks . . . 62
6.4 Panels . . . 63
6.5 Media Mentions . . . 64

7 Funding: University of Waterloo **65**

8 Funding: University of Maryland **67**

9 Teaching: University of Waterloo **71**
9.1 Courses Taught . . . 71

Cloudera Exhibit 1006 - Page 1 of 86

10 Teaching: University of Maryland **72**
10.1 Courses Taught . . . 72
10.2 Course or Curriculum Development . . . 73

11 Advising: University of Waterloo **73**
11.1 Undergraduate . . . 73
11.2 Master's . . . 74
11.3 Ph.D. . . . 76
11.4 Post-Doctoral . . . 77

12 Advising: University of Maryland **78**
12.1 Undergraduate . . . 78
12.2 Master's . . . 78
12.3 Ph.D. . . . 78
12.4 Post-Doctoral . . . 80

13 Service: Professional **80**
13.1 Editorial Boards . . . 80
13.2 Conference Chair Positions . . . 80
13.3 Reviewing: Journals . . . 81
13.4 Reviewing: Conferences . . . 82
13.5 Reviewing: Workshops . . . 82
13.6 Reviewing: Funding Agencies . . . 83
13.7 Workshop Organization . . . 83
13.8 TREC Organization . . . 83
13.9 Other . . . 84

14 Service: University of Waterloo **84**
14.1 School . . . 84
14.2 Faculty . . . 84
14.3 Campus . . . 84

15 Service: University of Maryland **85**
15.1 Department . . . 85
15.2 College: Standing Committees . . . 85
15.3 College: Ad Hoc Committees . . . 85
15.4 College: Other . . . 85
15.5 University . . . 85

1 Personal Information

1.1 Academic Appointments

University of Waterloo

05/2021—	Co-director, Waterloo AI Institute
06/2018—	Affiliate Faculty Vector Institute
08/2015—	Professor and David R. Cheriton Chair David R. Cheriton School of Computer Science

University of Maryland

08/2015—07/2018	Research Professor, College of Information Studies
08/2015—08/2015	Professor, College of Information Studies (70%)
08/2015—08/2015	Professor, Institute for Advanced Computer Studies (30%)
08/2015—08/2015	Professor, Department of Computer Science (0%)
08/2012—08/2015	Associate Dean for Research, College of Information Studies
08/2010—08/2015	Associate Professor, College of Information Studies (70%)
08/2010—08/2015	Associate Professor, Institute for Advanced Computer Studies (30%)
08/2009—08/2015	Affiliate Associate Professor, Department of Computer Science (0%)
08/2009—08/2010	Affiliate Associate Professor, Institute for Advanced Computer Studies (0%)
08/2009—08/2010	Associate Professor, College of Information Studies (100%)
07/2006—08/2009	Affiliate Assistant Professor, Department of Computer Science (0%)
10/2004—08/2009	Affiliate Assistant Professor, Institute for Advanced Computer Studies (0%)
08/2004—08/2009	Assistant Professor, College of Information Studies (100%)

1.2 Education

02/2001—08/2004	Massachusetts Institute of Technology (Cambridge, Massachusetts) Doctor of Philosophy in Electrical Engineering and Computer Science *Thesis title:* Event Structure and the Encoding of Arguments: The Syntax of the Mandarin and English Verb Phrase
02/2000—02/2001	Massachusetts Institute of Technology (Cambridge, Massachusetts) Master of Engineering in Electrical Engineering and Computer Science *Thesis title:* Indexing and Retrieving Natural Language Using Ternary Expressions
09/1997—02/2000	Massachusetts Institute of Technology (Cambridge, Massachusetts) Bachelor of Science in Electrical Engineering and Computer Science

1.3 Employment

05/2022—	Chief Technology Officer Primal Fusion Inc. (Waterloo, Ontario)
10/2018—09/2021	Chief Scientist RSVP.ai (Waterloo, Ontario)

06/2010—01/2015	Research Scientist Twitter (San Francisco, California)
06/2013—08/2013	Visiting Researcher Centrum Wiskunde & Informatica (Amsterdam, The Netherlands)
06/2007—06/2009 06/2006—08/2006	Visiting Scientist National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (Bethesda, Maryland)
06/2005—08/2005	Visiting Scientist Lister Hill National Center for Biomedical Communications, National Library of Medicine (NLM), National Institutes of Health (Bethesda, Maryland)
09/1997—08/2004	Research Assistant Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (Cambridge, Massachusetts)
06/2001—08/2001	Summer Intern Microsoft Research (Redmond, Washington)
06/1999—08/1999	Summer Intern IBM T.J. Watson Research Center (Hawthorne, New York)
01/1998—05/1998	Research Assistant Media Lab, Massachusetts Institute of Technology (Cambridge, Massachusetts)

2 Publications

Total publication count = 593
h-index 86, ∼27K citations according to Google Scholar as of February, 2023.

- In all references, my name is in bold.
- Unless otherwise indicated, the first author is the lead author.
- References marked with $^{\alpha}$ indicate that authors are listed in alphabetical order, or that all co-authors contributed equally.
- Underlined names indicate students or postdocs I have mentored, including those for whom I am/was the (co-)advisor and other cases where guidance was limited to specific projects. Typically, the lead author is one of these individuals. In some cases, I have continued to co-author with a student/postdoc after they have graduated/left; these cases are *not* underlined.
- The prototypical author order in my publications is: lead student (or co-leads), other contributing students, myself. Other contributing students are listed in order of contribution or in alphabetical order; other than the lead position (or co-leads), the order of authors typically contains low signal. In cases involving other faculty, they are typically listed after the students; there is little signal in the ordering among faculty. Especially post-tenure, my default my position is the final author. This means that any deviation from this prototype *is* intended to convey some signal.
- Publications are listed reverse chronologically and numbered in descending order per section to provide stable references from other documents.

2.1 Books and Monographs

(B4) **Jimmy Lin**, Rodrigo Nogueira, and Andrew Yates. Pretrained Transformers for Text Ranking: BERT and Beyond. Morgan & Claypool Publishers, 2021. (325 pages)

Cloudera Exhibit 1006 - Page 4 of 86

(B3) **Jimmy Lin** and Chris Dyer. Data-Intensive Text Processing with MapReduce. Morgan & Claypool Publishers, 2010. (177 pages)

(B2) **Jimmy Lin**. Event Structure and the Encoding of Arguments: The Syntax of the Mandarin and English Verb Phrase. Ph.D. Thesis, Massachusetts Institute of Technology, 2004. (194 pages)

(B1) **Jimmy Lin**. Indexing and Retrieving Natural Language Using Ternary Expressions. Master's Thesis, Massachusetts Institute of Technology, 2001. (58 pages)

2.2 Proceeding Volumes

(P3) $^{\alpha}$Ryan Clancy, Nicola Ferro, Claudia Hauff, **Jimmy Lin**, Tetsuya Sakai, Ze Zhong Wu. (Editors) Proceedings of the Open-Source IR Replicability Challenge (OSIRRC 2019): CEUR Workshop Proceedings Vol-2409, July 2019, Paris, France.

(P2) $^{\alpha}$**Jimmy Lin**, Jian Pei, Xiaohua Hu, Wo Chang, Raghunath Nambiar, Charu Aggarwal, Nick Cercone, Vasant Honavar, Jun Huan, Bamshad Mobasher, and Saumyadipta Pyne (editors). Proceedings of the 2014 IEEE International Conference on Big Data. New Jersey: IEEE Press, 2014.

(P1) $^{\alpha}$David Cheung, Il-Yeol Song, Wesley Chu, Xiaohua Hu, and **Jimmy Lin** (editors). Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2009.

2.3 Book Chapters

(H4) $^{\alpha}$Philip Resnik and **Jimmy Lin**. Evaluation of NLP Systems. In Alex Clark, Chris Fox, and Shalom Lappin, editors, Computational Linguistics and Natural Language Processing Handbook. Oxford, England: Blackwell Publishers, 2010, pages 271–295.

(H3) **Jimmy Lin**. Summarization. In Tamer M. Özsu and Ling Liu (editors), Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009, pages 2884–2889.

(H2) Boris Katz, Sue Felshin, **Jimmy Lin**, and Gregory Marton. Viewing the Web as a Virtual Database for Question Answering. In Mark Maybury, editor, New Directions in Question Answering. Cambridge, Massachusetts: MIT Press, 2004, pages 215–226.

(H1) Boris Katz, **Jimmy Lin**, Chris Stauffer, and Eric Grimson. Answering Questions about Moving Objects in Videos. In Mark Maybury, editor, New Directions in Question Answering. Cambridge, Massachusetts: MIT Press, 2004, pages 113–128.

2.4 Journal Articles

The right column lists either (1) the impact factor from Journal Citation Reports by Clarivate, for the year in which the article was published or the most recent impact factor available, or (2) the acceptance rate, if applicable.

(J56)	Sheng-Chieh Lin and **Jimmy Lin**. A Dense Representation Framework for Lexical and Semantic Matching. *ACM Transactions on Information Systems*, *in press*, 2023. (DOI: `10.1145/3582426`)	4.657 (2021)
(J55)	Joel Mackenzie, Andrew Trotman, and **Jimmy Lin**. Efficient Document-at-a-Time and Score-at-a-Time Query Evaluation for Learned Sparse Representations. *ACM Transactions on Information Systems*, *in press*, 2022. (DOI: `10.1145/3576922`)	4.657 (2021)

(J54) Mashrekur Rahman, Jonathan M. Frame, **Jimmy Lin**, and Grey S. Nearing. Hydrology Research Articles Are Becoming More Topically Diverse. *Journal of Hydrology*, 614:128551, 2022. (DOI: 10.1016/j.jhydrol.2022.128551) 6.708 (2021)

(J53) Bhaleka D. Persaud, Krysha A. Dukacz, Gopal C. Saha, Amber Peterson, Laleh Moradi, Stephen O'Hearn, Erin Clary, Juliane Mai, Michael Steeleworthy, Jason J. Venkiteswaran, Homa Kheyrollah Pour, Brent B. Wolfe, Sean K. Carey, John W. Pomeroy, Chris M. DeBeer, James M. Waddington, Philippe Van Cappellen, and **Jimmy Lin**. Ten Best Practices to Strengthen Stewardship and Sharing of Water Science Data in Canada. *Hydrological Processes*, 35(11):e14385, 2021. (DOI: 10.1002/hyp.14385) 3.784 (2021)

(J52) Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and **Jimmy Lin**. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. *ACM Transactions on Information Systems*, 39(4), Article No. 48, 2021. (29 pages) (DOI: 10.1145/3446426) 4.657 (2021)

(J51) Juliane Mai, Bryan A. Tolson, Hongren Shen, Étienne Gaborit, Vincent Fortin, Nicolas Gasset, Hervé Awoye, Tricia A. Stadnyk, Lauren M. Fry, Emily A. Bradley, Frank Seglenieks, André G. Temgoua, Daniel G. Princz, Shervan Gharari, Amin Haghnegahdar, Mohamed E. Elshamy, Saman Razavi, Martin Gauch, **Jimmy Lin**, Xiaojing Ni, Yongping Yuan, Meghan McLeod, Nandita Basu, Rohini Kumar, Oldrich Rakovec, Luis Samaniego, Sabine Attinger, Narayan K. Shrestha, Prasad Daggupati, Tirthankar Roy, Sungwook Wi, Tim Hunter, James R. Craig, and Alain Pietroniro. Great Lakes Runoff Intercomparison Project Phase 3: Lake Erie (GRIP-E). *Journal of Hydrologic Engineering*, 26(9):05021020, 2021. (19 pages) (2023 Award for Best Case Study) (DOI: 10.1061/(ASCE)HE.1943-5584.0002097) 2.439 (2021)

(J50) Martin Gauch, Frederik Kratzert, Daniel Klotz, Grey Nearing, **Jimmy Lin**, and Sepp Hochreiter. Rainfall–Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network. *Hydrology and Earth System Sciences*, 25(4):2045–2062, 2021. (DOI: 10.5194/hess-25-2045-2021) 6.617 (2021)

(J49) Samantha Fritz, Ian Milligan, Nick Ruest, and **Jimmy Lin**. Fostering Community Engagement through Datathon Events: The Archives Unleashed Experience. *Digital Humanities Quarterly*, 15(1), 2021. (14 pages)

(J48) Nick Ruest, Samantha Fritz, Ryan Deschamps, **Jimmy Lin**, and Ian Milligan. From Archive to Analysis: Accessing Web Archives at Scale Through a Cloud-Based Interface. *International Journal of Digital Humanities*, 2021. (20 pages)

(J47) Martin Gauch, Juliane Mai, and **Jimmy Lin**. The Proper Care and Feeding of CAMELS: How Limited Training Data Affects Streamflow Prediction. *Environmental Modelling & Software*, 135:104926, 2021. (9 pages) (DOI: 10.1016/j.envsoft.2020.104926) 5.288 (2020)

(J46) Rodrigo Nogueira, Zhiying Jiang, Kyunghyun Cho, and **Jimmy Lin**. Navigation-Based Candidate Expansion and Pretrained Language Models for Citation Recommendation. *Scientometrics*, 125(3):3001–3016, 2020. (DOI: 10.1007/s11192-020-03718-9) 2.867 (2019)

(J45) Samantha Fritz, Ian Milligan, Nick Ruest, and **Jimmy Lin**. Building Community at Distance: A Datathon During COVID-19. *Digital Library Perspectives*, 36(4):415–428, 2020. (DOI: 10.1108/DLP-04-2020-0024)

(J44) Siddhartha Sahu, Amine Mhedhbi, Semih Salihoglu, **Jimmy Lin**, and M. Tamer Özsu. The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing: Extended Survey. *The VLDB Journal*, 29(2–3):595–618, 2020. (DOI: 10.1007/s00778-019-00548-x) 2.868 (2020)

(J43) Frank Hopfgartner, Allan Hanbury, Henning Müller, Ivan Eggel, Krisztian Balog, Torben Brodt, Gordon V. Cormack, **Jimmy Lin**, Jayashree Kalpathy-Cramer, Noriko Kando, Makoto P. Kato, Anastasia Krithara, Tim Gollub, Martin Potthast, Evelyne Viega, and Simon Mercer. Evaluation-as-a-Service for the Computational Sciences: Overview and Outlook. *Journal of Data and Information Quality*, 10(4), Article 15, 2018. (32 pages) (DOI: 10.1145/3239570)

(J42) Peilin Yang, Hui Fang, and **Jimmy Lin**. Anserini: Reproducible Ranking Baselines Using Lucene. *Journal of Data and Information Quality*, 10(4), Article 16, 2018. (20 pages) (DOI: 10.1145/3239571)

(J41) Siddhartha Sahu, Amine Mhedhbi, Semih Salihoglu, **Jimmy Lin**, M. Tamer Özsu. The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing. *Proceedings of the VLDB Endowment*, 11(4):420-431, 2017. (Best Paper) (DOI: 10.1145/3164135.3164139)

(J40) **Jimmy Lin**, Ian Milligan, Jeremy Wiebe, and Alice Zhou. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *ACM Journal on Computing and Cultural Heritage*, 10(4), Article 22, 2017. (30 pages) (DOI: 10.1145/3097570) 0.529 (2017)

(J39) $^{\alpha}$**Jimmy Lin** and Andrew Trotman. The Role of Index Compression in Score-at-a-Time Query Evaluation. *Information Retrieval*, 20(3):199–220, 2017. (DOI: 10.1007/s10791-016-9291-5) 1.488 (2017)

(J38) Aneesh Sharma, Jerry Jiang, Praveen Bommannavar, Brian Larson, and **Jimmy Lin**. GraphJet: Real-Time Content Recommendations at Twitter. *Proceedings of the VLDB Endowment*, 9(13):1281–1292, 2016. 33%

(J37) Ahmed El-Roby, Khaled Ammar, Ashraf Aboulnaga, and **Jimmy Lin**. Sapphire: Querying RDF Data Made Simple. *Proceedings of the VLDB Endowment*, 9(13):1481–1484, 2016. 37%

(J36) Abdul Quamar, Amol Deshpande, and **Jimmy Lin**. NScale: Neighborhood-Centric Large-Scale Graph Analytics in the Cloud. *The VLDB Journal*, 25(2):125–150, 2016. (DOI: 10.1007/s00778-015-0405-2) 4.269 (2016)

(J35) **Jimmy Lin**. On Building Better Mousetraps and Understanding the Human Condition: Reflections on Big Data in the Social Sciences. *The Annals of the American Academy of Political and Social Science*, 659(1):33–47, May 2015. (DOI: 10.1177/0002716215569174) 1.708 (2015)

(J34) Hua He, **Jimmy Lin**, and Adam Lopez. Gappy Pattern Matching on GPUs for On-Demand Extraction of Hierarchical Translation Grammars. *Transactions of the Association for Computational Linguistics*, 3:87–100, 2015. (DOI: 10.1162/tacl_a_00124)

(J33) Ferhan Ture, **Jimmy Lin**. Exploiting Representations from Statistical Machine Translation for Cross-Language Information Retrieval. *ACM Transactions on Information Systems*, 32(4), Article 19, 2014. (32 pages) (DOI: 10.1145/2644807) 1.021 (2014)

(J32) Nima Asadi, **Jimmy Lin**, and Arjen P. de Vries. Runtime Optimizations for Tree-Based Machine Learning Models. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2281–2292, 2014. (DOI: 10.1109/TKDE.2013.73) 2.067 (2014)

(J31) Pankaj Gupta, Venu Satuluri, Ajeet Grewal, Siva Gurumurthy, Volodymyr Zhabiuk, Quannan Li, and **Jimmy Lin**. Real-Time Twitter Recommendation: Online Motif Detection in Large Dynamic Graphs. *Proceedings of the VLDB Endowment*, 7(13):1379–1380, 2014. 34%

(J30) Oscar Boykin, Sam Ritchie, Ian O'Connell, and **Jimmy Lin**. Summingbird: A Framework for Integrating Batch and Online MapReduce Computations. *Proceedings of the VLDB Endowment*, 7(13):1441–1451, 2014. 34%

(J29) Abdul Quamar, Amol Deshpande, and **Jimmy Lin**. NScale: Neighborhood-centric Analytics on Large Graphs. *Proceedings of the VLDB Endowment*, 7(13):1673–1676, 2014. 37%

(J28) Nima Asadi and **Jimmy Lin**. Document Vector Representations for Feature Extraction in Multi-Stage Document Ranking. *Information Retrieval*, 16(6):747–768, 2013. (DOI: `10.1007/s10791-012-9217-9`) 0.625 (2013)

(J27) K. Ashwin Kumar, Jonathan Gluck, Amol Deshpande, and **Jimmy Lin**. Hone: "Scaling Down" Hadoop on Shared-Memory Systems. *Proceedings of the VLDB Endowment*, 6(12):1354–1357, 2013. 49%

(J26) Nima Asadi and **Jimmy Lin**. Fast Candidate Generation for Real-Time Tweet Search with Bloom Filter Chains. *ACM Transactions on Information Systems*, 31(3), Article 13, 2013. (36 pages) (DOI: `10.1145/2493175.2493178`) 1.300 (2013)

(J25) $^{\alpha}$George Lee, **Jimmy Lin**, Chuang Liu, Andrew Lorek, and Dmitriy Ryaboy. The Unified Logging Infrastructure for Data Analytics at Twitter. *Proceedings of the VLDB Endowment*, 5(12):1771–1780, 2012. 26%

(J24) $^{\alpha}$Gregory Chockler, Eliezer Dekel, Joseph JaJa, and **Jimmy Lin**. Special Issue on Cloud Computing. *Journal of Parallel and Distributed Computing*, 71(6):731, 2011. (DOI: `10.1016/j.jpdc.2011.04.001`) 0.859 (2011)

(J23) Ben Langmead, Michael C. Schatz, **Jimmy Lin**, Mihai Pop, and Steven L. Salzberg. Searching for SNPs with Cloud Computing. *Genome Biology*, 10:R134, 2009. (10 pages) (DOI: `10.1186/gb-2009-10-11-r134`) 6.626 (2009)

(J22) Allison Druin, Paul T. Jaeger, Kenneth R. Fleischmann, Jennifer Golbeck, **Jimmy Lin**, Yan Qu, Ping Wang, and Bo Xie. The Maryland Modular Method: An Approach to Doctoral Education in Information Studies. *Journal of Education for Library and Information Science*, 50(4):293–301, 2009.

(J21) Timothy Hawes, **Jimmy Lin**, and Philip Resnik. Elements of a Computational Model for Multi-Party Discourse: The Turn-Taking Behavior of Supreme Court Justices. *Journal of the American Society for Information Science and Technology*, 60(8):1607–1615, 2009. (DOI: `10.1002/asi.21087`) 1.783 (2009)

(J20) **Jimmy Lin** and W. John Wilbur. Modeling Actions of PubMed Users with N-Gram Language Models. *Information Retrieval*, 12(4):487–503, 2009. (DOI: `10.1007/s10791-008-9067-7`) 1.841 (2009)

(J19) Paul T. Jaeger, **Jimmy Lin**, Justin Grimes, and Shannon Simmons. Where Is the Cloud? Geography, Economics, Environment, and Jurisdiction in Cloud Computing. *First Monday*, 14(5), 2009. (Online, 12 pages printed)

(J18) Judith L. Klavans, Carolyn Sheffield, Eileen Abels, **Jimmy Lin**, Rebecca Passonneau, Tandeep Sidhu, and Dagobert Soergel. Computational Linguistics for Metadata Building (CLiMB): Using Text Mining for the Automatic Identification, Categorization, and Disambiguation of Subject Terms for Image Metadata. *Journal of Multimedia Tools and Applications*, 42(1):115–138, 2009. (DOI: `10.1007/s11042-008-0253-9`)

(J17) **Jimmy Lin**, G. Craig Murray, Bonnie J. Dorr, Jan Hajič, and Pavel Pecina. A Cost-effective Lexical Acquisition Process for Large-Scale Thesaurus Translation. *Language Resource and Evaluation*, 43(1):27–40, 2009. (DOI: 10.1007/s10579-008-9074-8) 0.634 (2009)

(J16) **Jimmy Lin**. Is Searching Full Text More Effective Than Searching Abstracts? *BMC Bioinformatics*, 10:46 (3 Feb 2009). (15 pages) (DOI: 10.1186/1471-2105-10-46) 3.428 (2009)

(J15) **Jimmy Lin**, Philip Wu, and Eileen Abels. Towards Automatic Facet Analysis and Need Negotiation: Lessons from Mediated Search. *ACM Transactions on Information Systems*, 27(1), Article 6, 2008. (42 pages) (DOI: 10.1145/1416950.1416956) 1.472 (2008)

(J14) Paul T. Jaeger, **Jimmy Lin**, and Justin Grimes. Cloud Computing and Information Policy: Computing in a Policy Cloud? *Journal of Information Technology & Politics*, 5(3):269–283, 2008. (DOI: 10.1080/19331680802425479)

(J13) **Jimmy Lin**, Michael DiCuccio, Vahan Grigoryan, and W. John Wilbur. Navigating Information Spaces: A Case Study of Related Article Search in PubMed. *Information Processing & Management*, 44(5):1771–1783, 2008. (DOI: 10.1016/j.ipm.2008.04.002) 1.852 (2008)

(J12) David Zajic, Bonnie Dorr, and **Jimmy Lin**. Single-Document and Multi-Document Summarization Techniques for Email Threads Using Sentence Compression. *Information Processing & Management*, 44(4):1600–1610, 2008. (DOI: 10.1016/j.ipm.2007.09.007) 1.852 (2008)

(J11) **Jimmy Lin**. PageRank without Hyperlinks: Reranking with PubMed Related Article Networks for Biomedical Text Retrieval. *BMC Bioinformatics*, 9:270 (6 Jun 2008). (12 pages) (DOI: 10.1186/1471-2105-9-270) 3.781 (2008)

(J10) Michael Evans, Wayne McIntosh, **Jimmy Lin**, and Cynthia Cates. Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research. *Journal of Empirical Legal Studies*, 4(4):1007–1039, 2007. (DOI: 10.1111/j.1740-1461.2007.00113.x)

(J9) **Jimmy Lin** and W. John Wilbur. PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity. *BMC Bioinformatics*, 8:423 (30 October 2007). (14 pages) (DOI: 10.1186/1471-2105-8-423) 3.493 (2007)

(J8) **Jimmy Lin** and W. John Wilbur. Syntactic Sentence Compression in the Biomedical Domain: Facilitating Access to Related Articles. *Information Retrieval*, 10(4–5):393–414, 2007. (DOI: 10.1007/s10791-007-9029-5) 0.770 (2007)

(J7) David Zajic, Bonnie J. Dorr, **Jimmy Lin**, Richard Schwartz. Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks. *Information Processing & Management*, 43(6):1549–1570, 2007. (DOI: 10.1016/j.ipm.2007.01.016) 1.500 (2007)

(J6) **Jimmy Lin**. An Exploration of the Principles Underlying Redundancy-Based Factoid Question Answering. *ACM Transactions on Information Systems*, 25(2):1–55, 2007. (DOI: 10.1145/1229179.1229180) 1.969 (2007)

(J5) Dina Demner-Fushman and **Jimmy Lin**. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103, 2007. (DOI: 10.1162/coli.2007.33.1.63) 2.367 (2007)

(J4) **Jimmy Lin**. User Simulations for Evaluating Answers to Question Series. *Information Processing & Management*, 43(3):717–729, 2007. (DOI: 10.1016/j.ipm.2006.06.006) 1.500 (2007)

(J3) **Jimmy Lin** and Dina Demner-Fushman. Methods for Automatically Evaluating Answers to Complex Questions. *Information Retrieval*, 9(5):565–587, 2006. (DOI: 10.1007/s10791-006-9003-7) 1.744 (2006)

(J2) **Jimmy Lin** and Boris Katz. Building a Reusable Test Collection for Question Answering. *Journal of the American Society for Information Science and Technology*, 57(7):851–861, 2006. (DOI: `10.1002/asi.20348`) 1.555 (2006)

(J1) Joyce Chai, **Jimmy Lin**, Wlodek Zadrozny, Timing Ye, Margo Stys-Budzikowska, Veronika Horvath, Nanda Kambhatla, and Catherine Wolf. The Role of a Natural Language Conversational Interface in Online Sales: A Case Study. *International Journal of Speech Technology*, 4(3–4):285–295, 2001. (DOI: `10.1023/A:1011316909641`)

2.5 Conference Papers

The conference abbreviation appears in bold. The exact classification of the paper (e.g., full, short, etc.) is provided in cases where there may be ambiguity. The right column shows the acceptance rate if available.

(C214) Manveer Singh Tamber, Ronak Pradeep, and **Jimmy Lin**. Pre-Processing Matters! Improved Wikipedia Corpora for Open-Domain Question Answering. Proceedings of the 45th European Conference on Information Retrieval (**ECIR 2023**), April 2023, Dublin, Ireland.

(C213) Ronak Pradeep, Haonan Chen, Lingwei Gu, Manveer Singh Tamber, and **Jimmy Lin**. PyGaggle: A Gaggle of Resources for Open-Domain Question Answering. Proceedings of the 45th European Conference on Information Retrieval (**ECIR 2023**), April 2023, Dublin, Ireland.

(C212) Hang Li, Shengyao Zhuang, Xueguang Ma, **Jimmy Lin**, and Guido Zuccon. Pseudo-Relevance Feedback with Dense Retrievers in Pyserini. Proceedings of the 26th Australasian Document Computing Symposium (**ADCS 2022**), December 2022, Adelaide, Australia.

(C211) Minghan Li, Xinyu Zhang, Ji Xin, Hongyang Zhang, and **Jimmy Lin**. Certified Error Control of Candidate Set Pruning for Two-Stage Relevance Ranking. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2022**), pages 333–345, December 2022, Abu Dhabi, United Arab Emirates.

(C210) Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and **Jimmy Lin**. AfriCLIRMatrix: Enabling Cross-Lingual Information Retrieval for African Languages. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2022**), pages 8721–8728, December 2022, Abu Dhabi, United Arab Emirates.

(C209) Raphael Tang, Karun Kumar, Gefei Yang, Akshat Pandey, Yajie Mao, Vladislav Belyaev, Madhuri Emmadi, Craig Murray, Ferhan Ture, and **Jimmy Lin**. SpeechNet: Weakly Supervised, End-to-End Speech Recognition at Industrial Scale. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2022**), December 2022, Abu Dhabi, United Arab Emirates.

(C208) Wei Zhong, Jheng-Hong Yang, Yuqing Xie, and **Jimmy Lin**. Evaluating Token-Level and Passage-Level Dense Retrieval Models for Math Information Retrieval. Findings of the Association for Computational Linguistics: **EMNLP 2022**, pages 1092–1102, December 2022, Abu Dhabi, United Arab Emirates.

(C207) Peng Shi, Rui Zhang, He Bai, and **Jimmy Lin**. XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing. Findings of the Association for Computational Linguistics: **EMNLP 2022**, pages 5248–5259, December 2022, Abu Dhabi, United Arab Emirates.

(C206) Peng Shi, Linfeng Song, Lifeng Jin, Haitao Mi, He Bai, **Jimmy Lin**, and Dong Yu. Cross-Lingual Text-to-SQL Semantic Parsing with Representation Mixup. Findings of the Association for Computational Linguistics: **EMNLP 2022**, pages 5296–5306, December 2022, Abu Dhabi, United Arab Emirates.

(C205) Zhiying Jiang, Yiqin Dai, Ji Xin, Ming Li, and **Jimmy Lin**. Few-Shot Non-Parametric Learning with Deep Latent Variable Model. Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems (**NeurIPS 2022**), December 2022.

(C204) Wei Zhong, Yuqing Xie, and **Jimmy Lin**. Applying Structural and Dense Semantic Matching for the ARQMath Lab 2022, CLEF. Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum: CEUR Workshop Proceedings Vol-3180, pages 147–170, Bologna, Italy, September 2022. (Best paper from ARQMath-3 @ CLEF 2022)

(C203) Chris Kamphuis, Faegheh Hasibi, **Jimmy Lin**, and Arjen de Vries. REBL: Entity Linking at Scale. Proceedings of the 3rd International Conference on Design of Experimental Search & Information REtrieval Systems (**DESIRES 2022**), San Jose, California, August 2022.

(C202) Ronak Pradeep, Yilin Li, Yuetong Wang, and **Jimmy Lin**. Neural Query Synthesis and Domain-Specific Ranking Templates for Multi-Stage Clinical Trial Matching. Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2022**), pages 2325–2330, July 2022, Madrid, Spain. (Short Paper) (DOI: `10.1145/3477495.3531853`) 25%

(C201) Hang Li, Shuai Wang, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, **Jimmy Lin**, and Guido Zuccon. To Interpolate or not to Interpolate: PRF, Dense and Sparse Retrievers. Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2022**), pages 2495–2500, July 2022, Madrid, Spain. (Short Paper) (DOI: `10.1145/3477495.3531884`) 25%

(C200) Yuqi Liu, Chengcheng Hu, and **Jimmy Lin**. Another Look at Information Retrieval as Statistical Translation. Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2022**), pages 2749–2754, July 2022, Madrid, Spain. (Reproducibility Paper) (DOI: `10.1145/3477495.3531717`) 28%

(C199) **Jimmy Lin**, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Fostering Coopetition While Plugging Leaks: The Design and Implementation of the MS MARCO Leaderboards. Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2022**), pages 2939–2948, July 2022, Madrid, Spain. (Resource Paper) (DOI: `10.1145/3477495.3531725`) 41%

(C198) Ellen M. Voorhees, Nick Craswell, and **Jimmy Lin**. Too Many Relevants: Whither Cranfield Test Collections? Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2022**), pages 2970–2980, July 2022, Madrid, Spain. (Resource Paper) (DOI: `10.1145/3477495.3531728`) 41%

(C197) Xueguang Ma, Ronak Pradeep, Rodrigo Nogueira, and **Jimmy Lin**. Document Expansions and Learned Sparse Lexical Representations for MS MARCO V1 and V2. Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2022**), pages 3187–3197, July 2022, Madrid, Spain. (Resource Paper) (DOI: `10.1145/3477495.3531749`) 41%

(C196) Andrew Trotman, Joel Mackenzie, Pradeesh Parameswaran, and **Jimmy Lin**. A Common Framework for Exploring Document-at-a-Time and Score-at-a-Time Retrieval Methods. Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2022**), pages 3229–3234, July 2022, Madrid, Spain. (Demonstration Paper) (DOI: 10.1145/3477495.3531657) 52%

(C195) Matthew Y. R. Yang, Siwen Yang, and **Jimmy Lin**. Integration of Text and Geospatial Search for Hydrographic Datasets Using the Lucene Search Library. Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (**JCDL 2022**), article no. 36, pages 1–5, June 2022, Cologne, Germany. (Demonstration Paper) (DOI: 10.1145/3529372.3533280)

(C194) Raphael Tang, Karun Kumar, Ji Xin, Piyush Vyas, Wenyan Li, Gefei Yang, Yajie Mao, Craig Murray, and **Jimmy Lin**. Temporal Early Exiting for Streaming Speech Commands Recognition. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP 2022**), pages 7567–7571, May 2022, Singapore. (DOI: 10.1109/ICASSP43922.2022.9746863) 45%

(C193) Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, **Jimmy Lin**, and Guido Zuccon. Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback: A Reproducibility Study. Proceedings of the 44th European Conference on Information Retrieval (**ECIR 2022**), Part I, pages 599–612, April 2021, Stavanger, Norway. (Reproducibility Paper) (DOI: 10.1007/978-3-030-99736-6_40)

(C192) Xueguang Ma, Kai Sun, Ronak Pradeep, Minghan Li, and **Jimmy Lin**. Another Look at DPR: Reproduction of Training and Replication of Retrieval. Proceedings of the 44th European Conference on Information Retrieval (**ECIR 2022**), Part I, pages 613–626, April 2021, Stavanger, Norway. (Reproducibility Paper) (DOI: 10.1007/978-3-030-99736-6_41)

(C191) Ronak Pradeep, Yuqi Liu, Xinyu Zhang, Yilin Li, Andrew Yates, and **Jimmy Lin**. Squeezing Water from a Stone: A Bag of Tricks for Further Improving Cross-Encoder Effectiveness for Reranking. Proceedings of the 44th European Conference on Information Retrieval (**ECIR 2022**), Part I, pages 655–670, April 2021, Stavanger, Norway. (Reproducibility Paper) (DOI: 10.1007/978-3-030-99736-6_44)

(C190) Josh Devins, Julie Tibshirani, and **Jimmy Lin**. Aligning the Research and Practice of Building Search Applications: Elasticsearch and Pyserini. Proceedings of the 15th ACM International Conference on Web Search and Data Mining (**WSDM 2022**), pages 1573–1576, February 2022. (Demonstration Paper) (DOI: 10.1145/3488560.3502186)

(C189) Raphael Tang, Karun Kumar, Kendra Chalkley, Ji Xin, Liming Zhang, Wenyan Li, Gefei Yang, Yajie Mao, Junho Shin, Geoffrey Murray, and **Jimmy Lin**. Voice Query Auto Completion. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2021**), pages 900–906, November 2021, Online and Punta Cana, Dominican Republic. (Short Paper) 18%

(C188) Sheng-Chieh Lin, Jheng-Hong Yang, and **Jimmy Lin**. Contextualized Query Embeddings for Conversational Search. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2021**), pages 1004–1015, November 2021, Online and Punta Cana, Dominican Republic. (Long Paper) 26%

(C187) Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and **Jimmy Lin**. Simple and Effective Unsupervised Redundancy Elimination to Compress Dense Vectors for Passage Retrieval. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2021**), pages 2854–2859, November 2021, Online and Punta Cana, Dominican Republic. (Short Paper) 18%

(C186) Minghan Li, Ming Li, Kun Xiong, and **Jimmy Lin**. Multi-Task Dense Retrieval via Model Uncertainty Fusion for Open-Domain Question Answering. Findings of the Association for Computational Linguistics: **EMNLP 2021**, pages 274–287, November 2021, Punta Cana, Dominican Republic. 37%

(C185) Anup Anand Deshmukh, Qianqiu Zhang, Ming Li, **Jimmy Lin**, and Lili Mou. Unsupervised Chunking as Syntactic Structure Induction with a Knowledge-Transfer Approach. Findings of the Association for Computational Linguistics: **EMNLP 2021**, pages 3626–3634, November 2021, Punta Cana, Dominican Republic. 29%

(C184) Ogundepo Odunayo, Naveela N. Sookoo, Gautam Bathla, Anthony Cavallin, Bhaleka D. Persaud, Kathy Szigeti, Philippe Van Cappellen, **Jimmy Lin**. Rescuing Historical Climate Observations to Support Hydrological Research: A Case Study of Solar Radiation Data. Proceedings of the 21st ACM Symposium on Document Engineering (**DocEng '21**), Article No. 19, August 2021. (Short Paper) (DOI: `10.1145/3469096.3474929`) 49%

(C183) Ji Xin, Raphael Tang, Yaoliang Yu, and **Jimmy Lin**. The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (**ACL 2021**), pages 1040–1051, August 2021. (DOI: `10.18653/v1/2021.acl-long.84`) 21%

(C182) Kelvin Jiang, Ronak Pradeep, and **Jimmy Lin**. Exploring Listwise Evidence Reasoning with T5 for Fact Verification. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (**ACL 2021**), pages 402–410, August 2021. (DOI: `10.18653/v1/2021.acl-short.51`) 21%

(C181) Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, **Jimmy Lin**, and Allan Hanbury. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2021**), pages 113–122, July 2021. (Full Paper) (DOI: `https://doi.org/10.1145/3404835.3462891`) 21%

(C180) Nick Craswell, Bhaskar Mitra, Daniel Campos, Emine Yilmaz, **Jimmy Lin**. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2021**), pages 1566–1576, July 2021. (Perspectives Paper) (DOI: `10.1145/3404835.3462804`) 27%

(C179) Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and **Jimmy Lin**. Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2021**), pages 2066–2070, July 2021. (Short Paper) (DOI: `10.1145/3404835.3463120`) 28%

(C178) **Jimmy Lin**, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2021**), pages 2283—2287, July 2021. (Short Paper) (DOI: `10.1145/3404835.3463034`) 28%

(C177) **Jimmy Lin**, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2021**), pages 2356–2362, July 2021. (Resource Paper) (DOI: 10.1145/3404835.3463238) 45%

(C176) Edwin Zhang, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, Rodrigo Nogueira, and **Jimmy Lin**. Chatty Goose: A Python Framework for Conversational Search. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2021**), pages 2521–2525, July 2021. (Demonstration Paper) (DOI: 10.1145/3404835.3462782) 54%

(C175) Wei Zhong and **Jimmy Lin**. PyA0: A Python Toolkit for Accessible Math-Aware Search. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2021**), pages 2541–2545, July 2021. (Demonstration Paper) (DOI: 10.1145/3404835.3462794) 54%

(C174) Xiao Han, Yuqi Liu, and **Jimmy Lin**. The Simplest Thing That Can Possibly Work: (Pseudo-)Relevance Feedback via Text Classification. Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (**ICTIR 2021**), pages 123–129, July 2021. (DOI: 10.1145/3471158.3472261) 51%

(C173) Ji Xin, Raphael Tang, Yaoliang Yu, and **Jimmy Lin**. BERxiT: Early Exiting for *BERT with Better Fine-Tuning and Extension to Regression. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (**EACL 2021**), pages 91–104, April 2021. (Long Paper) (DOI: 10.18653/v1/2021.eacl-main.8) 25%

(C172) Mohan Zhang, Luchen Tan, Zhengkai Tu, Zihang Fu, Kun Xiong, Ming Li, and **Jimmy Lin**. Don't Change Me! User-Controllable Selective Paraphrase Generation. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (**EACL 2021**), pages 3522–3527, April 2021. (Short Paper) (DOI: 10.18653/v1/2021.eacl-main.307) 25%

(C171) Xinyu Zhang, Andrew Yates, and **Jimmy Lin**. Comparing Score Aggregation Approaches for Document Retrieval with Pretrained Transformers. Proceedings of the 43rd European Conference on Information Retrieval (**ECIR 2021**), Part II, pages 150–163, March 2021. (Reproducibility Paper) (DOI: 10.1007/978-3-030-72240-1_11)

(C170) He Bai, Peng Shi, **Jimmy Lin**, Yuqing Xie, Luchen Tan, Kun Xiong, Wen Gao, and Ming Li. Segatron: Segment-Aware Transformer for Language Modeling and Understanding. Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (**AAAI-21**), pages 12526–12534, February 2021. 21%

(C169) Jheng-Hong Yang, Sheng-Chieh Lin, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and **Jimmy Lin**. Designing Templates for Eliciting Commonsense Knowledge from Pretrained Sequence-to-Sequence Models. Proceedings of the 28th International Conference on Computational Linguistics (**COLING 2020**), pages 3449–3453, December 2020. (Short Paper) 26%

(C168) Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and **Jimmy Lin**. Document Ranking with a Pretrained Sequence-to-Sequence Model. Findings of the Association for Computational Linguistics: **EMNLP 2020**, pages 708–718, November 2020. (DOI: 10.18653/v1/2020.findings-emnlp.63) 38%

(C167) Peng Shi, He Bai, and **Jimmy Lin**. Cross-Lingual Training of Neural Models for Document Ranking. Findings of the Association for Computational Linguistics: **EMNLP 2020**, pages 2768–2773, November 2020. (DOI: 10.18653/v1/2020.findings-emnlp.249) 38%

(C166) Zhiying Jiang, Raphael Tang, Ji Xin, and **Jimmy Lin**. Inserting Information Bottlenecks for Attribution in Transformers. Findings of the Association for Computational Linguistics: **EMNLP 2020**, pages 3850–3857, November 2020. (DOI: 10.18653/v1/2020.findings-emnlp.343) 38%

(C165) Andrew Yates, Kevin Martin Jose, Xinyu Zhang, **Jimmy Lin**. Flexible IR Pipelines with Capreolus. Proceedings of the 29th International Conference on Information and Knowledge Management (**CIKM 2020**), pages 3181–3188, October 2020. (Resource Paper) (DOI: 10.1145/3340531.3412780)

(C164) Zhengkai Tu, Wei Yang, Zihang Fu, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and **Jimmy Lin**. Approximate Nearest Neighbor Search and Lightweight Dense Vector Reranking in Multi-Stage Retrieval Architectures. Proceedings of the 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (**ICTIR 2020**), pages 97–100, September 2020. (Short Paper) (DOI: 10.1145/3409256.3409818) 57%

(C163) Venu Satuluri, Yao Wu, Xun Zheng, Yilei Qian, Brian Wichers, Qieyun Dai, Gui Ming Tang, Jerry Jiang, and **Jimmy Lin**. SimClusters: Community-Based Representations for Heterogeneous Recommendations at Twitter. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (**KDD 2020**), pages 3183–3193, August 2020. (Applied Data Science Paper) (DOI: 10.1145/3394486.3403370) 16%

(C162) Nick Ruest, **Jimmy Lin**, Ian Milligan, and Samantha Fritz. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. Proceedings of the 20th ACM/IEEE Joint Conference on Digital Libraries (**JCDL 2020**), pages 157–166, August 2020. (Full Paper) (DOI: 10.1145/3383583.3398513) 31%

(C161) Zeynep Akkalyoncu Yilmaz, Charles L. A. Clarke, and **Jimmy Lin**. A Lightweight Environment for Learning Experimental IR Research Practices. Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2020**), pages 2113–2116, July 2020. (Demonstration Paper) (DOI: 10.1145/3397271.3401395) 58%

(C160) **Jimmy Lin**, Joel Mackenzie, Chris Kamphuis, Craig Macdonald, Antonio Mallia, Michał Siedlaczek, Andrew Trotman, and Arjen de Vries. Supporting Interoperability Between Open-Source Search Engines with the Common Index File Format. Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2020**), pages 2149–2152, July 2020. (Demonstration Paper) (DOI: 10.1145/3397271.3401404) 58%

(C159) Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and **Jimmy Lin**. DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (**ACL 2020**), pages 2246–2251, July 2020. (Short Paper) (DOI: 10.18653/v1/2020.acl-main.204) 18%

(C158) Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and **Jimmy Lin**. Showing Your Work Doesn't Always Work. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (**ACL 2020**), pages 2766–2772, July 2020. (Short Paper) (DOI: 10.18653/v1/2020.acl-main.246) 18%

(C157) Hamidreza Shahidi, Ming Li, and **Jimmy Lin**. Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (**ACL 2020**), pages 3864–3870, July 2020. (Short Paper) (DOI: 10.18653/v1/2020.acl-main.355) 18%

(C156) Yuqing Xie, Wei Yang, Luchen Tan, Kun Xiong, Nicholas Jing Yuan, Baoxing Huai, Ming Li, and **Jimmy Lin**. Distant Supervision for Multi-Stage Fine-Tuning in Retrieval-Based Question Answering. Proceedings of The Web Conference 2020 (**WWW '20**), pages 2934–2940, April 2020. (DOI: 10.1145/3366423.3380060) 24%

(C155) Adrien Grand, Robert Muir, Jim Ferenczi, and **Jimmy Lin**. From MaxScore to Block-Max WAND: The Story of How Lucene Significantly Improved Query Evaluation Performance. Proceedings of the 42nd European Conference on Information Retrieval, Part II (**ECIR 2020**), pages 20–27, April 2020. (Reproducibility Paper) (DOI: 10.1007/978-3-030-45442-5_3) 38%

(C154) Chris Kamphuis, Arjen de Vries, Leonid Boytsov, and **Jimmy Lin**. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. Proceedings of the 42nd European Conference on Information Retrieval, Part II (**ECIR 2020**), pages 28–34, April 2020. (Reproducibility Paper) (DOI: 10.1007/978-3-030-45442-5_4) 38%

(C153) **Jimmy Lin** and Qian Zhang. Reproducibility is a Process, not an Achievement: The Replicability of IR Reproducibility Experiments. Proceedings of the 42nd European Conference on Information Retrieval, Part II (**ECIR 2020**), pages 43–49, April 2020. (Reproducibility Paper) (DOI: 10.1007/978-3-030-45442-5_6) 38%

(C152) **Jimmy Lin**, Ian Milligan, Douglas Oard, Nick Ruest, and Katie Shilton. We Could, but Should We? Ethical Considerations for Providing Access to GeoCities and Other Historical Digital Collections. Proceedings of the Fifth ACM SIGIR Conference on Human Information Interaction and Retrieval (**CHIIR 2020**), pages 135–144, March 2020. (Perspectives Paper) (DOI: 10.1145/3343413.3377980) 34%

(C151) Royal Sequiera, Luchen Tan, Yinan Zhang, and **Jimmy Lin**. Update Delivery Mechanisms for Prospective Information Needs: A Reproducibility Study. Proceedings of the Fifth ACM SIGIR Conference on Human Information Interaction and Retrieval (**CHIIR 2020**), pages 308–312, March 2020. (Short Paper) (DOI: 10.1145/3343413.3377988) 34%

(C150) Andrew Yates, Siddhant Arora, Xinyu Zhang, Wei Yang, Kevin Martin Jose, and **Jimmy Lin**. Capreolus: A Toolkit for End-to-End Neural Ad Hoc Retrieval. Proceedings of the 13th ACM International Conference on Web Search and Data Mining (**WSDM 2020**), pages 861–864, February 2020, Houston, Texas. (Demonstration Paper) (DOI: 10.1145/3336191.3371868)

(C149) Allison McCoy, Dean Sittig, **Jimmy Lin**, and Adam Wright. Identification and Ranking of Biomedical Informatics Researcher Citation Statistics through a Google Scholar Scraper. Proceedings of the 2019 Annual Symposium of the American Medical Informatics Association (**AMIA 2019**), pages 655–663, November 2019, Washington, D.C. (Full Paper)

(C148) Linqing Liu, Wei Yang, Jinfeng Rao, Raphael Tang, and **Jimmy Lin**. Incorporating Contextual and Syntactic Structures Improves Semantic Similarity Modeling. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP 2019**), pages 1204–1209, November 2019, Hong Kong, China. (Short Paper) (DOI: 10.18653/v1/D19-1114) 21%

(C147) Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and **Jimmy Lin**. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP 2019**), pages 3490–3496, November 2019, Hong Kong, China. (Short Paper) (DOI: 10.18653/v1/D19-1352) 21%

(C146) Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, **Jimmy Lin**, and Xu Sun. Aligning Cross-Lingual Entities with Multi-Aspect Information. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP 2019**), pages 4431–4441, November 2019, Hong Kong, China. (Long Paper) (DOI: 10.18653/v1/D19-1451) 26%

(C145) Jinfeng Rao, Linqing Liu, Yi Tay, Wei Yang, Peng Shi, and **Jimmy Lin**. Bridging the Gap Between Relevance Matching and Semantic Matching for Short Text Similarity Modeling. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP 2019**), pages 5370–5381, November 2019, Hong Kong, China. (Long Paper) (DOI: 10.18653/v1/D19-1540) 26%

(C144) Ji Xin, **Jimmy Lin**, and Yaoliang Yu. What Part of the Neural Network Does This? Understanding LSTMs by Measuring and Dissecting Neurons. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP 2019**), pages 5823—5830, November 2019, Hong Kong, China. (Short Paper) (DOI: 10.18653/v1/D19-1591) 21%

(C143) Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and **Jimmy Lin**. Applying BERT to Document Retrieval with Birch. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP 2019**): System Demonstrations, pages 19–24, November 2019, Hong Kong, China. (DOI: 10.18653/v1/D19-3004) 40%

(C142) Jaejun Lee, Raphael Tang, and **Jimmy Lin**. Honkling: In-Browser Personalization for Ubiquitous Keyword Spotting. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP 2019**): System Demonstrations, pages 91–96, November 2019, Hong Kong, China. (DOI: 10.18653/v1/D19-3016) 40%

(C141) Martin Gauch, Juliane Mai, Shervan Gharari, and **Jimmy Lin**. Data-Driven vs. Physically-Based Streamflow Prediction Models. Proceedings of the 9th International Workshop on Climate Informatics, October 2019, Paris, France.

(C140) Wei Yang, Kuang Lu, Peilin Yang, and **Jimmy Lin**. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2019**), pages 1129–1132, July 2019, Paris, France. (Short Paper) (Best Short Paper Honorable Mention) (DOI: 10.1145/3331184.3331340) 24%

(C139) **Jimmy Lin** and Peilin Yang. The Impact of Score Ties on Repeatability in Document Ranking. Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2019**), pages 1125–1128, July 2019, Paris, France. (Short Paper) (DOI: 10.1145/3331184.3331339) 24%

(C138) Raphael Tang, Ferhan Ture, and **Jimmy Lin**. Yelling at Your TV: An Analysis of Speech Recognition Errors and Subsequent User Behavior on Entertainment Systems. Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2019**), pages 853–856, July 2019, Paris, France. (Short Paper) (DOI: 10.1145/3331184.3331271) 24%

(C137) Ryan Clancy, Toke Eskildsen, Nick Ruest, and **Jimmy Lin**. Solr Integration in the Anserini Information Retrieval Toolkit. Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2019**), pages 1285–1288, July 2019, Paris, France. (Demonstration Paper) (DOI: 10.1145/3331184.3331401)

(C136) Ryan Clancy, Jaejun Lee, Zeynep Akkalyoncu Yilmaz, and **Jimmy Lin**. Information Retrieval Meets Scalable Text Analytics: Solr Integration with Spark. Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2019**), pages 1313–1316, July 2019, Paris, France. (Demonstration Paper) (DOI: 10.1145/3331184.3331395)

(C135) Peng Shi, Jinfeng Rao, and **Jimmy Lin**. Simple Attention-Based Representation Learning for Ranking Short Social Media Posts. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (**NAACL 2019**), pages 2212–2217, June 2019, Minneapolis, Minnesota. (Short Paper) (DOI: 10.18653/v1/N19-1229) 21%

(C134) Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and **Jimmy Lin**. Rethinking Complex Neural Network Architectures for Document Classification. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (**NAACL 2019**), pages 4046–4051, June 2019, Minneapolis, Minnesota. (Short Paper) (DOI: 10.18653/v1/N19-1408) 21%

(C133) Wei Yang, Luchen Tan, Chunwei Lu, Anqi Cui, Han Li, Xi Chen, Kun Xiong, Muzi Wang, Ming Li, Jian Pei, and **Jimmy Lin**. Detecting Customer Complaint Escalation with Recurrent Neural Networks and Manually-Engineered Features. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers) (**NAACL 2019**), pages 56–63, June 2019, Minneapolis, Minnesota. (DOI: 10.18653/v1/N19-2008) 28%

(C132) Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and **Jimmy Lin**. End-to-End Open-Domain Question Answering with BERTserini. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (**NAACL 2019**), pages 72–77, June 2019, Minneapolis, Minnesota. (DOI: 10.18653/v1/N19-4013) 38%

(C131) Ryan Deschamps, Samantha Fritz, **Jimmy Lin**, Ian Milligan, and Nick Ruest. The Cost of a WARC: Analyzing Web Archives in the Cloud. Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (**JCDL 2019**), pages 261–264, June 2019, Urbana-Champaign, Illinois. (Short Paper) (DOI: 10.1109/JCDL.2019.00043) 23%

(C130) Ian Milligan, Nathalie Casemajor, Samantha Fritz, **Jimmy Lin**, Nick Ruest, Matthew S. Weber, and Nicholas Worby. Building Community and Tools for Analyzing Web Archives through Datathons. Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (**JCDL 2019**), pages 265–268, June 2019, Urbana-Champaign, Illinois. (Short Paper) (DOI: 10.1109/JCDL.2019.00044) 23%

(C129) Ruifan Yu, Yuhao Xie, and **Jimmy Lin**. Simple Techniques for Cross-Collection Relevance Feedback. Proceedings of the 41st European Conference on Information Retrieval, Part I (**ECIR 2019**), pages 369–381, April 2019, Cologne, Germany. (Reproducibility Paper) (Best Reproducibility Paper — Honourable Mention) (DOI: 10.1007/978-3-030-15712-8_26)

(C128) Peilin Yang and **Jimmy Lin**. Reproducing and Generalizing Semantic Term Matching in Axiomatic Information Retrieval. Proceedings of the 41st European Conference on Information Retrieval, Part I (**ECIR 2019**), pages 397–409, April 2019, Cologne, Germany. (Reproducibility Paper) (DOI: 10.1007/978-3-030-15712-8_24)

(C127) Jinfeng Rao, Wei Yang, Yuhao Zhang, Ferhan Ture, and **Jimmy Lin**. Multi-Perspective Relevance Matching with Hierarchical ConvNets for Social Media Search. Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (**AAAI-19**), pages 232–240, January 2019, Honolulu, Hawaii. (Full Paper) (DOI: 10.1609/aaai.v33i01.3301232) 16%

(C126) Michael Azmy, Peng Shi, Ihab Ilyas, and **Jimmy Lin**. Farewell Freebase: Migrating the SimpleQuestions Dataset to DBpedia. Proceedings of the 27th International Conference on Computational Linguistics (**COLING 2018**), pages 2093–2103, August 2018, Santa Fe, New Mexico. (Resource Paper) 37%

(C125) Jinfeng Rao, Ferhan Ture, and **Jimmy Lin**. Multi-Task Learning with Neural Networks for Voice Query Understanding on an Entertainment Platform. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (**KDD 2018**), pages 635–645, August 2018, London, United Kingdom. (Applied Data Science Track Paper) (DOI: 10.1145/3219819.3219870) 23%

(C124) $^{\alpha}$**Jimmy Lin**, Salman Mohammed, Royal Sequiera, and Luchen Tan. Update Delivery Mechanisms for Prospective Information Needs: An Analysis of Attention in Mobile Users. Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2018**), pages 785–794, July 2018, Ann Arbor, Michigan. (Full Paper) (DOI: 10.1145/3209978.3210018) 21%

(C123) Jinfeng Rao, Ferhan Ture, and **Jimmy Lin**. What Do Viewers Say to Their TVs? An Analysis of Voice Queries to Entertainment Systems. Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2018**), pages 1213–1216, July 2018, Ann Arbor, Michigan. (Short Paper) (DOI: 10.1145/3209978.3210140) 30%

(C122) Ajeet Grewal, Jerry Jiang, Gary Lam, Tristan Jung, Lohith Vuddemarri, Quannan Li, Aaditya Landge, and **Jimmy Lin**. RecService: Distributed Real-Time Graph Processing at Twitter. Proceedings of the 10th USENIX Workshop on Hot Topics in Cloud Computing (**HotCloud '18**), July 2018, Boston, Massachusetts. 35%

(C121) Peilin Yang, Srikanth Thiagarajan, and **Jimmy Lin**. Robust, Scalable, Real-Time Event Time Series Aggregation at Twitter. Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data (**SIGMOD 2018**), pages 595–599, June 2018, Houston, Texas. (Industry Track Paper) (DOI: 10.1145/3183713.3190663) 38%

(C120) Salman Mohammed, Peng Shi, and **Jimmy Lin**. Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (**NAACL 2018**), pages 291–296, May 2018, New Orleans, Louisiana. (DOI: 10.18653/v1/N18-2047) 29%

(C119) Zhucheng Tu, Mengping Li, and **Jimmy Lin**. Pay-Per-Request Deployment of Neural Network Models Using Serverless Architectures. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (**NAACL 2018**), pages 6–10, May 2018, New Orleans, Louisiana. (DOI: 10.18653/v1/N18-5002) 51%

(C118) Yiyun Liang, Zhucheng Tu, Laetitia Huang, and **Jimmy Lin**. CNNs for NLP in the Browser: Client-Side Deployment and Visualization Opportunities. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (**NAACL 2018**), pages 61–65, May 2018, New Orleans, Louisiana. (DOI: 10.18653/v1/N18-5013) 51%

(C117) Raphael Tang, Weijie Wang, Zhucheng Tu, and **Jimmy Lin**. An Experimental Analysis of the Power Consumption of Convolutional Neural Networks for Keyword Spotting. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP 2018**), pages 5479–5483, April 2018, Calgary, Alberta, Canada. (DOI: 10.1109/ICASSP.2018.8461624) 50%

(C116) Raphael Tang and **Jimmy Lin**. Deep Residual Learning for Small-Footprint Keyword Spotting. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP 2018**), pages 5484–5488, April 2018, Calgary, Alberta, Canada. (DOI: 10.1109/ICASSP.2018.8462688) 50%

(C115) Joel Mackenzie, Shane Culpepper, Roi Blanco, Matt Crane, Charles L. A. Clarke, and **Jimmy Lin**. Query Driven Algorithm Selection in Early Stage Retrieval. Proceedings of the 11th ACM International Conference on Web Search and Data Mining (**WSDM 2018**), pages 396–404, February 2018, Marina Del Rey, California. (DOI: 10.1145/3159652.3159676) 16%

(C114) Gaurav Baruah, Richard McCreadie, and **Jimmy Lin**. A Comparison of Nuggets and Clusters for Evaluating Timeline Summaries. Proceedings of the 2017 International Conference on Information and Knowledge Management (**CIKM 2017**), pages 67–76, November 2017, Singapore. (Full Paper) (DOI: 10.1145/3132847.3133000) 21%

(C113) Jinfeng Rao, Ferhan Ture, Hua He, Oliver Jojic, and **Jimmy Lin**. Talking to Your TV: Context-Aware Voice Search with Hierarchical Recurrent Neural Networks. Proceedings of the 2017 International Conference on Information and Knowledge Management (**CIKM 2017**), pages 557–566, November 2017, Singapore. (Full Paper) (DOI: 10.1145/3132847.3132893) 21%

(C112) Jinfeng Rao, Ferhan Ture, Xing Niu, and **Jimmy Lin**. Mining the Temporal Statistics of Query Terms for Searching Social Media Posts. Proceedings of the 3rd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2017**), pages 133–140, October 2017, Amsterdam, The Netherlands. (Full Paper) (DOI: 10.1145/3121050.3121052) 50%

(C111) Matt Crane and **Jimmy Lin**. An Exploration of Serverless Architectures for Information Retrieval. Proceedings of the 3rd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2017**), pages 241–244, October 2017, Amsterdam, The Netherlands. (Short Paper) (DOI: 10.1145/3121050.3121086) 58%

(C110) Gaurav Baruah and **Jimmy Lin**. The Pareto Frontier of Utility Models as a Framework for Evaluating Push Notification Systems. Proceedings of the 3rd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2017**), pages 253–256, October 2017, Amsterdam, The Netherlands. (Short Paper) (Best Short Paper) (DOI: 10.1145/3121050.3121089) 58%

(C109) Salman Mohammed, Matt Crane, and **Jimmy Lin**. Quantization in Append-Only Collections. Proceedings of the 3rd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2017**), pages 265–268, October 2017, Amsterdam, The Netherlands. (Short Paper) (DOI: 10.1145/3121050.3121092) 58%

(C108) Hua He, Kris Ganjam, Navendu Jain, Jessica Lundin, Ryen White, and **Jimmy Lin**. An Insight Extraction System on BioMedical Literature with Deep Neural Networks. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2017**), pages 2691–2701, September 2017, Copenhagen, Denmark. (Long Paper) (DOI: 10.18653/v1/D17-1285) 26%

(C107) Adam Roegiest, Luchen Tan, and **Jimmy Lin**. Online In-Situ Interleaved Evaluation of Real-Time Push Notification Systems. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2017**), pages 415–424, August 2017, Tokyo, Japan. (Full Paper) (DOI: 10.1145/3077136.3080808) 22%

(C106) Luchen Tan, Gaurav Baruah, and **Jimmy Lin**. On the Reusability of "Living Labs" Test Collections: A Case Study of Real-Time Summarization. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2017**), pages 793–796, August 2017, Tokyo, Japan. (Short Paper) (DOI: 10.1145/3077136.3080644) 30%

(C105) Haotian Zhang, Jinfeng Rao, **Jimmy Lin**, and Mark D. Smucker. Automatically Extracting High-Quality Negative Examples for Answer Selection in Question Answering. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2017**), pages 797–800, August 2017, Tokyo, Japan. (Short Paper) (DOI: 10.1145/3077136.3080645) 30%

(C104) Jinfeng Rao, Hua He, and **Jimmy Lin**. Experiments with Convolutional Neural Network Models for Answer Selection. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2017**), pages 1217–1220, August 2017, Tokyo, Japan. (Short Paper) (DOI: 10.1145/3077136.3080648) 30%

(C103) Royal Sequiera and **Jimmy Lin**. Finally, a Downloadable Test Collection of Tweets. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2017**), pages 1225–1228, August 2017, Tokyo, Japan. (Short Paper) (DOI: 10.1145/3077136.3080667) 30%

(C102) Peilin Yang, Hui Fang, and **Jimmy Lin**. Anserini: Enabling the Use of Lucene for Information Retrieval Research. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2017**), pages 1253–1256, August 2017, Tokyo, Japan. (Short Paper) (DOI: 10.1145/3077136.3080721) 30%

(C101) Nimesh Ghelani, Salman Mohammed, Shine Wang, and **Jimmy Lin**. Event Detection on Curated Tweet Streams. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2017**), pages 1325–1328, August 2017, Tokyo, Japan. (Demonstration Paper) (DOI: 10.1145/3077136.3084141) 47%

(C100) Kareem El Gebaly and **Jimmy Lin**. In-Browser Interactive SQL Analytics with Afterburner. Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data (**SIGMOD 2017**), pages 1623–1626, May 2017, Chicago, Illinois. (Demonstration Paper) (DOI: 10.1145/3035918.3058736) 34%

(C99) [α]Charles L. A. Clarke, Gordon V. Cormack, **Jimmy Lin**, and Adam Roegiest. Ten Blue Links on Mars. Proceedings of the 26th International World Wide Web Conference (**WWW 2017**), pages 273–281, April 2017, Perth, Australia. (Full Paper) (DOI: 10.1145/3038912.3052625) 17%

(C98) Yulu Wang and **Jimmy Lin**. Partitioning and Segment Organization Strategies for Real-Time Selective Search on Document Streams. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (**WSDM 2017**), pages 221–230, February 2017, Cambridge, the United Kingdom. (Full Paper) (DOI: 10.1145/3018661.3018727) 16%

(C97) [α]Matt Crane, J. Shane Culpepper, **Jimmy Lin**, Joel Mackenzie, Andrew Trotman. A Comparison of Document-at-a-Time and Score-at-a-Time Query Evaluation. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (**WSDM 2017**), pages 201–210, February 2017, Cambridge, the United Kingdom. (Full Paper) (DOI: 10.1145/3018661.3018726) 16%

(C96) Andrew Trotman and **Jimmy Lin**. In Vacuo and In Situ Evaluation of SIMD Codecs. Proceedings of the 21st Australasian Document Computing Symposium (**ADCS 2016**), pages 1–8, November 2016, Melbourne, Australia. (Full Paper) (DOI: 10.1145/3015022.3015023) 56%

(C95) J. Shane Culpepper, Charles L. A. Clarke and **Jimmy Lin**. Dynamic Cutoff Prediction in Multi-Stage Retrieval Systems. Proceedings of the 21st Australasian Document Computing Symposium (**ADCS 2016**), pages 17–24, November 2016, Melbourne, Australia. (Full Paper) (Best Paper) (DOI: 10.1145/3015022.3015026) 56%

(C94) Jinfeng Rao, Hua He, and **Jimmy Lin**. Noise-Contrastive Estimation for Answer Selection with Deep Neural Networks. Proceedings of the 25th International Conference on Information and Knowledge Management (**CIKM 2016**), pages 1913–1916, October 2016, Indianapolis, Indiana. (Short Paper) (DOI: 10.1145/2983323.2983872) 24%

(C93) Gaurav Baruah, Haotian Zhang, Rakesh Guttikonda, **Jimmy Lin**, Mark D. Smucker, and Olga Vechtomova. Optimizing Nugget Annotations with Active Learning. Proceedings of the 25th International Conference on Information and Knowledge Management (**CIKM 2016**), pages 2359–2364, October 2016, Indianapolis, Indiana. (Extended Short Paper) (DOI: 10.1145/2983323.2983694) 24%

(C92) [α]Charles L. A. Clarke, Gordon V. Cormack, **Jimmy Lin**, and Adam Roegiest. Total Recall: Blue Sky on Mars. Proceedings of the 2nd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2016**), pages 45–48, September 2016, Newark, Delaware. (Short Paper) (DOI: 10.1145/2970398.2970430) 54%

(C91) Jiaul H. Paik and **Jimmy Lin**. Retrievability in API-Based "Evaluation as a Service". Proceedings of the 2nd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2016**), pages 91–94, September 2016, Newark, Delaware. (Short Paper) (Best Short Paper) (DOI: 10.1145/2970398.2970427) 54%

(C90) Ahmed Elbagoury, Matt Crane, and **Jimmy Lin**. Rank-at-a-Time Query Processing. Proceedings of the 2nd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2016**), pages 229–232, September 2016, Newark, Delaware. (Short Paper) (DOI: 10.1145/2970398.2970434) 54%

(C89) Jinfeng Rao and **Jimmy Lin**. Temporal Query Expansion Using a Continuous Hidden Markov Model. Proceedings of the 2nd ACM International Conference on the Theory of Information Retrieval (**ICTIR 2016**), pages 295–298, September 2016, Newark, Delaware. (Short Paper) (DOI: 10.1145/2970398.2970424) 54%

(C88) Xin Qian, **Jimmy Lin**, and Adam Roegiest. Interleaved Evaluation for Retrospective Summarization and Prospective Notification on Document Streams. Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2016**), pages 175–184, July 2016, Pisa, Italy. (Full Paper) (DOI: 10.1145/2911451.2911494) 18%

(C87) Luchen Tan, Adam Roegiest, **Jimmy Lin**, and Charles L. A. Clarke. An Exploration of Evaluation Metrics for Mobile Push Notifications. Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2016**), pages 741–744, July 2016, Pisa, Italy. (Short Paper) (DOI: 10.1145/2911451.2914694) 31%

(C86) Cody Buntain and **Jimmy Lin**. Burst Detection in Social Media Streams for Tracking Interest Profiles in Real Time. Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2016**), pages 777–780, July 2016, Pisa, Italy. (Short Paper) (DOI: 10.1145/2911451.2914733) 31%

(C85) Haotian Zhang, **Jimmy Lin**, Gordon V. Cormack, and Mark D. Smucker. Sampling Strategies and Active Learning for Volume Estimation. Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2016**), pages 981–984, July 2016, Pisa, Italy. (Short Paper) (DOI: 10.1145/2911451.2914685) 31%

(C84) Luchen Tan, Adam Roegiest, Charles L. A. Clarke, and **Jimmy Lin**. Simple Dynamic Emission Strategies for Microblog Filtering. Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2016**), pages 1009–1012, July 2016, Pisa, Italy. (Short Paper) (DOI: 10.1145/2911451.2914704) 31%

(C83) Adam Roegiest, Luchen Tan, **Jimmy Lin**, and Charles L. A. Clarke. A Platform for Streaming Push Notifications to Mobile Assessors. Proceedings of the 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2016**), pages 1077–1080, July 2016, Pisa, Italy. (Demonstration Paper) (DOI: 10.1145/2911451.2911463) 60%

(C82) $^{\alpha}$Andrew Jackson, **Jimmy Lin**, Ian Milligan, and Nick Ruest. Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (**JCDL 2016**), pages 103–106, June 2016, Newark, New Jersey. (Short Paper) (DOI: 10.1145/2910896.2910912) 35%

(C81) Ian Milligan, Nick Ruest, and **Jimmy Lin**. Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses. Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (**JCDL 2016**), pages 107–110, June 2016, Newark, New Jersey. (Short Paper) (DOI: 10.1145/2910896.2910913) 35%

(C80) Hua He and **Jimmy Lin**. Pairwise Word Interaction Modeling with Neural Networks for Semantic Similarity Measurement. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (**NAACL–HLT 2016**), pages 937–948, June 2016, San Diego, California. (Long Paper) (DOI: 10.18653/v1/N16-1108) 25%

(C79) Praveen Bommannavar, **Jimmy Lin**, and Anand Rajaraman. Estimating Topical Volume in Social Media Streams. Proceedings of the 31st ACM Symposium on Applied Computing (**SAC 2016**), pages 1096–1101, April 2016, Pisa, Italy. (Full Paper) (DOI: 10.1145/2851613.2851810) 24%

(C78) **Jimmy Lin**, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, Sebastiano Vigna. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. Proceedings of the 38th European Conference on Information Retrieval (**ECIR 2016**), pages 408–420, March 2016, Padua, Italy. (Full Paper, Reproducible IR Track) (DOI: 10.1007/978-3-319-30671-1_30)

(C77) Jinfeng Rao, Xing Niu, and **Jimmy Lin**. Compressing and Decoding Term Statistics Time Series. Proceedings of the 38th European Conference on Information Retrieval (**ECIR 2016**), pages 675–681, March 2016, Padua, Italy. (Short Paper) (DOI: 10.1007/978-3-319-30671-1_52) 30%

(C76) Cody Buntain, **Jimmy Lin**, and Jennifer Golbeck. Discovering Key Moments in Social Media Streams. Proceedings of the 13th Annual IEEE Consumer Communications & Networking Conference (**CCNC 2016**), pages 373–381, January 2016, Las Vegas, Nevada. (Full Paper) (DOI: 10.1109/CCNC.2016.7444808)

(C75) $^{\alpha}$**Jimmy Lin** and Andrew Trotman. Anytime Ranking for Impact-Ordered Indexes. Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (**ICTIR 2015**), pages 301–304, September 2015, Northampton, Massachusetts. (Short Paper) (DOI: 10.1145/2808194.2809477) 51%

(C74) **Jimmy Lin**. Building a Self-Contained Search Engine in the Browser. Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (**ICTIR 2015**), pages 309–312, September 2015, Northampton, Massachusetts. (Short Paper) (DOI: 10.1145/2808194.2809478) 51%

(C73) Yulu Wang and **Jimmy Lin**. The Feasibility of Brute Force Scans for Real-Time Tweet Search. Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval (**ICTIR 2015**), pages 321–324, September 2015, Northampton, Massachusetts. (Short Paper) (DOI: 10.1145/2808194.2809489) 51%

(C72) Hua He, Kevin Gimpel, and **Jimmy Lin**. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2015**), pages 1576–1586, September 2015, Lisbon, Portugal. (Long Paper) (DOI: 10.18653/v1/D15-1181) 26%

(C71) Yulu Wang, Garrick Sherman, **Jimmy Lin**, and Miles Efron. Assessor Differences and User Preferences in Tweet Timeline Generation. Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2015**), pages 615–624, August 2015, Santiago, Chile. (Full Paper) (DOI: 10.1145/2766462.2767699) 20%

(C70) Sarah Weissman, Samet Ayhan, Joshua Bradley, and **Jimmy Lin**. Identifying Duplicate and Contradictory Information in Wikipedia. Proceedings of the 2015 IEEE/ACM Joint Conference on Digital Libraries (**JCDL 2015**), pages 57–60, June 2015, Knoxville, Tennessee. (Short Paper) (DOI: 10.1145/2756406.2756947) 50%

(C69) Jinfeng Rao, **Jimmy Lin**, and Miles Efron. Reproducible Experiments on Lexical and Temporal Feedback for Tweet Search. Proceedings of the 37th European Conference on Information Retrieval (**ECIR 2015**), pages 755–767, March 2015, Vienna, Austria. (Full Paper, Reproducible IR Track) (DOI: 10.1007/978-3-319-16354-3_82) 23%

(C68) Krist Wongsuphasawat and **Jimmy Lin**. Using Visualizations to Monitor Changes and Harvest Insights from a Global-Scale Logging Infrastructure at Twitter. Proceedings of the 2014 IEEE Conference on Visual Analytics Science and Technology (**VAST 2014**), pages 113–122, November 2014, Paris, France. (Full Paper, Conference Track) (DOI: 10.1109/VAST.2014.7042487) 38%

(C67) Miles Efron, **Jimmy Lin**, Jiyin He, and Arjen de Vries. Temporal Feedback for Tweet Search with Non-Parametric Density Estimation. Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2014**), pages 33–42, July 2014, Gold Coast, Australia. (Full Paper) (DOI: 10.1145/2600428.2609575) 21%

(C66) Hannes Mühleisen, Thaer Samar, **Jimmy Lin**, and Arjen de Vries. Old Dogs Are Great at New Tricks: Column Stores for IR Prototyping. Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2014**), pages 863–866, July 2014, Gold Coast, Australia. (Short Paper) (DOI: 10.1145/2600428.2609460) 39%

(C65) Ellen M. Voorhees, **Jimmy Lin**, and Miles Efron. On Run Diversity in "Evaluation as a Service". Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2014**), pages 959–962, July 2014, Gold Coast, Australia. (Short Paper) (DOI: 10.1145/2600428.2609484) 39%

(C64) Yulu Wang and **Jimmy Lin**. The Impact of Future Term Statistics in Real-Time Tweet Search. Proceedings of the 36th European Conference on Information Retrieval (**ECIR 2014**), pages 567–572, April 2014, Amsterdam, The Netherlands. (Short Paper) (DOI: 10.1007/978-3-319-06028-6_58) 30%

(C63) Seth A. Myers, Aneesh Sharma, Pankaj Gupta, and **Jimmy Lin**. Information Network or Social Network? The Structure of the Twitter Follow Graph. Proceedings of the 23rd International World Wide Web Conference Companion (**WWW 2014**), pages 493–498, April 2014, Seoul, South Korea. (Web Science Track, Full Paper) (DOI: 10.1145/2567948.2576939) 25%

(C62) **Jimmy Lin** and Miles Efron. Infrastructure Support for Evaluation as a Service. Proceedings of the 23rd International World Wide Web Conference Companion (**WWW 2014**), pages 79–82, April 2014, Seoul, South Korea. (Demonstration Paper) (DOI: 10.1145/2567948.2577014) 33%

(C61) K. Ashwin Kumar, Jonathan Gluck, Amol Deshpande, and **Jimmy Lin**. Optimization Techniques for "Scaling Down" Hadoop on Multi-Core, Shared-Memory Systems. Proceedings of the 17th International Conference on Extending Database Technology (**EDBT 2014**), pages 13–24, March 2014, Athens, Greece. (Full Paper) (DOI: 10.5441/002/edbt.2014.03) 20%

(C60) Nima Asadi, **Jimmy Lin**, and Michael Busch. Dynamic Memory Allocation Policies for Postings in Real-Time Twitter Search. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**SIGKDD 2013**), pages 1186–1194, August 2013, Chicago, Illinois. (Industry/Government Track, Full Paper) (DOI: 10.1145/2487575.2488221) 25%

(C59) Vladimir Eidelman, Ke Wu, Ferhan Ture, Philip Resnik, and **Jimmy Lin**. Mr. MIRA: Open-Source Large-Margin Structured Learning on MapReduce. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations (**ACL 2013**), pages 199–204, August 2013, Sofia, Bulgaria. (Demonstration Paper)

(C58) Nima Asadi and **Jimmy Lin**. Effectiveness/Efficiency Tradeoffs for Candidate Generation in Multi-Stage Retrieval Architectures. Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2013**), pages 997–1000, July 2013, Dublin, Ireland. (Short Paper) (DOI: 10.1145/2484028.2484132) 34%

(C57) Ferhan Ture and **Jimmy Lin**. Flat vs. Hierarchical Phrase-Based Translation Models for Cross-Language Information Retrieval. Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2013**), pages 813–816, July 2013, Dublin, Ireland. (Short Paper) (DOI: 10.1145/2484028.2484137) 34%

(C56) Miguel Rios and **Jimmy Lin**. Visualizing the "Pulse" of World Cities on Twitter. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (**ICWSM 2013**), pages 717–720, July 2013, Boston, Massachusetts. (Poster Paper)

(C55) Gilad Mishne, Jeff Dalton, Zhenghua Li, Aneesh Sharma, and **Jimmy Lin**. Fast Data in the Era of Big Data: Twitter's Real-Time Related Query Suggestion Architecture. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (**SIGMOD 2013**), pages 1147–1158, June 2013, New York, New York. (Industry Track, Full Paper) (DOI: 10.1145/2463676.2465290)

(C54) Hua He, **Jimmy Lin**, and Adam Lopez. Massively Parallel Suffix Array Queries and On-Demand Phrase Extraction for Statistical Machine Translation Using GPUs. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (**NAACL–HLT 2013**), pages 325–334, June 2013, Atlanta, Georgia. (Long Paper) 30%

(C53) Pankaj Gupta, Ashish Goel, **Jimmy Lin**, Aneesh Sharma, Dong Wang, and Reza Zadeh. WTF: The Who to Follow Service at Twitter. Proceedings of the 22th International World Wide Web Conference (**WWW 2013**), May 2013, pages 505–514, Rio de Janeiro, Brazil. (Practice and Experience Track, Full Paper) (DOI: 10.1145/2488388.2488433) 15%

(C52) Nima Asadi and **Jimmy Lin**. Training Efficient Tree-Based Models for Document Ranking. Proceedings of the 34th European Conference on Information Retrieval (**ECIR 2013**), pages 146–157, March 2013, Moscow, Russia. (Full Paper) (DOI: 10.1007/978-3-642-36973-5_13) 16%

(C51) Ferhan Ture, **Jimmy Lin**, and Douglas W. Oard. Combining Statistical Translation Techniques for Cross-Language Information Retrieval. Proceedings of the 24th International Conference on Computational Linguistics (**COLING 2012**), pages 2685–2702, December 2012, Mumbai, India. (Full Paper) 17%

(C50) Nima Asadi and **Jimmy Lin**. Fast Candidate Generation for Two-Phase Document Ranking: Postings List Intersection with Bloom Filters. Proceedings of the 21st International Conference on Information and Knowledge Management (**CIKM 2012**), pages 2419–2422, October 2012, Maui, Hawaii. (Poster Paper) (DOI: 10.1145/2396761.2398656) 47%

(C49) Ferhan Ture and **Jimmy Lin**. Why Not Grab a Free Lunch? Mining Large Corpora for Parallel Sentences to Improve Translation Modeling. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (**NAACL–HLT 2012**), pages 626–630, June 2012, Montreal, Quebec, Canada. (Short Paper) 35%

(C48) Ian Soboroff, Dean McCullough, **Jimmy Lin**, Craig Macdonald, Iadh Ounis, and Richard McCreadie. Evaluating Real-Time Search over Tweets. Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (**ICWSM 2012**), pages 579–582, June 2012, Dublin, Ireland. (Poster Paper)

(C47) **Jimmy Lin** and Gilad Mishne. A Study of "Churn" in Tweets and Real-Time Search Queries. Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (**ICWSM 2012**), pages 503–506, June 2012, Dublin, Ireland. (Poster Paper)

(C46) **Jimmy Lin** and Alek Kolcz. Large-Scale Machine Learning at Twitter. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (**SIGMOD 2012**), pages 793–804, May 2012, Scottsdale, Arizona. (Industry Track, Full Paper) (DOI: 10.1145/2213836.2213958)

(C45) Michael Busch, Krishna Gade, Brian Larson, Patrick Lok, Samuel Luckenbill, and **Jimmy Lin**. Earlybird: Real-Time Search at Twitter. Proceedings of the 28th International Conference on Data Engineering (**ICDE 2012**), pages 1360–1369, April 2012, Washington, D.C. (Industry Track, Full Paper) (DOI: 10.1109/ICDE.2012.149)

(C44) Florian Leibert, Jake Mannix, **Jimmy Lin**, and Babak Hamadani. Automatic Management of Partitioned, Replicated Search Services. Proceedings of the 2nd ACM symposium on Cloud computing (**SoCC 2011**), Article 27, October 2011, Cascais, Portugal. (8 pages) (Short Paper) (DOI: 10.1145/2038916.2038943) 17%

(C43) Tamer Elsayed, **Jimmy Lin**, and Don Metzler. When Close Enough Is Good Enough: Approximate Positional Indexes for Efficient Ranked Retrieval. Proceedings of the 20th International Conference on Information and Knowledge Management (**CIKM 2011**), pages 1993–1996, October 2011, Glasgow, Scotland. (Poster Paper) (DOI: 10.1145/2063576.2063873) 35%

(C42) **Jimmy Lin**, Rion Snow, and William Morgan. Smoothing Techniques for Adaptive Online Language Models: Topic Tracking in Tweet Streams. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**SIGKDD 2011**), pages 422–429, August 2011, San Diego, California. (Industry/Government Track, Full Paper) (DOI: 10.1145/2020408.2020476) 30%

(C41) Lidan Wang, **Jimmy Lin**, and Donald Metzler. A Cascade Ranking Model for Efficient Ranked Retrieval. Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2011**), pages 105–114, July 2011, Beijing, China. (Full Paper) (DOI: 10.1145/2009916.2009934) 20%

(C40) Ferhan Ture, Tamer Elsayed, and **Jimmy Lin**. No Free Lunch: Brute Force vs. Locality-Sensitive Hashing for Cross-lingual Pairwise Similarity. Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2011**), pages 943–952, July 2011, Beijing, China. (Full Paper) (DOI: 10.1145/2009916.2010042) 20%

(C39) Nima Asadi, Donald Metzler, Tamer Elsayed, and **Jimmy Lin**. Pseudo Test Collections for Learning Web Search Ranking Functions. Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2011**), pages 1073–1082, July 2011, Beijing, China. (Full Paper) (DOI: 10.1145/2009916.2010058) 20%

(C38) Di-Wei Huang and **Jimmy Lin**. Scaling Populations of a Genetic Algorithm for Job Shop Scheduling Problems using MapReduce. Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science (**CLOUDCOM 2010**), pages 780–785, November 2010, Indianapolis, Indiana. (Full Paper) (DOI: 10.1109/CloudCom.2010.18) 23%

(C37) Lidan Wang, Donald Metzler, and **Jimmy Lin**. Ranking under Temporal Constraints. Proceedings of the 19th International Conference on Information and Knowledge Management (**CIKM 2010**), pages 79–88, October 2010, Toronto, Ontario, Canada. (Full Paper) (DOI: 10.1145/1871437.1871452) 13%

(C36) Lidan Wang, **Jimmy Lin**, and Donald Metzler. Learning to Efficiently Rank. Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2010**), pages 138–145, July 2010, Geneva, Switzerland. (Full Paper) (DOI: 10.1145/1835449.1835475) 17%

(C35) **Jimmy Lin**, Nitin Madnani, and Bonnie Dorr. Putting the User in the Loop: Interactive Maximal Marginal Relevance for Query-Focused Summarization. Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (**NAACL/HLT 2010**), pages 305–308, June 2010, Los Angeles, California. (Short Paper) 35%

(C34) **Jimmy Lin**. Brute Force and Indexed Approaches to Pairwise Document Similarity Comparisons with MapReduce. Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2009**), pages 155–162, July 2009, Boston, Massachusetts. (Full Paper) (DOI: 10.1145/1571941.1571970) 16%

(C33) Michael Lieberman and **Jimmy Lin**. You Are Where You Edit: Locating Wikipedia Users Through Edit Histories. Proceedings of the Third International Conference on Weblogs and Social Media (**ICWSM 2009**), pages 106–113, May 2009, San Jose, California. (Full Paper) 17%

(C32) **Jimmy Lin**. Scalable Language Processing Algorithms for the Masses: A Case Study in Computing Word Co-occurrence Matrices with MapReduce. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2008**), pages 419–428, Honolulu, Hawaii, October 2008. (Full Paper) 30%

(C31) $^{\alpha}$**Jimmy Lin** and Mark D. Smucker. How Do Users Find Things with PubMed? Towards Automatic Utility Evaluation with User Simulations. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2008**), pages 19–26, Singapore, July 2008. (Full Paper) (DOI: 10.1145/1390334.1390340) 17%

(C30) Tamer Elsayed, **Jimmy Lin**, and Douglas W. Oard. Pairwise Document Similarity in Large Collections with MapReduce. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Companion Volume (**ACL 2008**), pages 265–268, June 2008, Columbus, Ohio. (Short Paper) 25%

(C29) **Jimmy Lin** and Dina Demner-Fushman. Semantic Clustering of Answers to Clinical Questions. Proceedings of the 2007 Annual Symposium of the American Medical Informatics Association (**AMIA 2007**), pages 458–462, November 2007, Chicago, Illinois. (Full Paper) 45%

(C28) **Jimmy Lin** and Pengyi Zhang. Deconstructing Nuggets: The Stability and Reliability of Complex Question Answering Evaluation. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2007**), pages 327–334, July 2007, Amsterdam, The Netherlands. (Full Paper) (DOI: 10.1145/1277741.1277799) 17%

(C27) $^{\alpha}$Hoa Trang Dang and **Jimmy Lin**. Different Structures for Evaluating Answers to Complex Questions: Pyramids Are Stable, and So Are Human Assessors. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (**ACL 2007**), pages 768–775, June 2007, Prague, Czech Republic. (Full Paper) 22%

(C26) **Jimmy Lin**. Is Question Answering Better Than Information Retrieval? Towards a Task-Based Evaluation Framework for Question Series. Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (**HLT/NAACL 2007**), pages 212–219, April 2007, Rochester, New York. (Full Paper) 24%

(C25) Xiaoli Huang, **Jimmy Lin**, and Dina Demner-Fushman. Evaluation of PICO as a Knowledge Representation for Clinical Questions. Proceedings of the 2006 Annual Symposium of the American Medical Informatics Association (**AMIA 2006**), pages 359–363, November 2006, Washington, D.C. (Full Paper) 41%

(C24) **Jimmy Lin** and Dina Demner-Fushman. The Role of Knowledge in Conceptual Retrieval: A Study in the Domain of Clinical Medicine. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2006**), pages 99–106, August 2006, Seattle, Washington. (Full Paper) (DOI: `10.1145/1148170.1148191`) 18%

(C23) **Jimmy Lin**, Philip Wu, Dina Demner-Fushman, and Eileen Abels. Exploring the Limits of Single-Iteration Clarification Dialogs. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2006**), pages 469–476, August 2006, Seattle, Washington. (Full Paper) (DOI: `10.1145/1148170.1148251`) 18%

(C22) **Jimmy Lin**. The Role of Information Retrieval in Answering Complex Questions. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics **(COLING/ACL 2006)**, Main Conference Poster Sessions, pages 523–530, July 2006, Sydney, Australia. (Full Poster Paper) 23%

(C21) G. Craig Murray, Bonnie J. Dorr, **Jimmy Lin**, Jan Hajic, and Pavel Pecina. Leveraging Reusability: Cost-Effective Lexical Acquisition for Large-scale Ontology Translation. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (**COLING/ACL 2006**), pages 945–952, July 2006, Sydney, Australia. (Full Paper) (DOI: `10.3115/1220175.1220294`) 24%

(C20) Dina Demner-Fushman and **Jimmy Lin**. Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (**COLING/ACL 2006**), pages 841–848, July 2006, Sydney, Australia. (Full Paper) (DOI: `10.3115/1220175.1220281`) 24%

(C19) G. Craig Murray, Bonnie J. Dorr, **Jimmy Lin**, Jan Hajic, and Pavel Pecina. Leveraging Recurrent Phrase Structure in Large-scale Ontology Translation. Proceedings of the 11th Annual conference of the European Association for Machine Translation (**EAMT 2006**), June 2006, Oslo, Norway. (10 pages) (Full Paper)

(C18) **Jimmy Lin** and Dina Demner-Fushman. Will Pyramids Built of Nuggets Topple Over? Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (**HLT/NAACL 2006**), pages 383–390, June 2006, New York, New York. (Full Paper) 25%

(C17) **Jimmy Lin** and Dina Demner-Fushman. Automatically Evaluating Answers to Definition Questions. Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (**HLT/EMNLP 2005**), pages 931–938, October 2005, Vancouver, British Columbia, Canada. (Full Paper) 25%

(C16) **Jimmy Lin**. Evaluation of Resources for Question Answering Evaluation. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2005**), pages 392–399, August 2005, Salvador, Brazil. (Full Paper) (DOI: 10.1145/1076034.1076102) 19%

(C15) Michael Evans, Wayne McIntosh, Cynthia L. Cates, and **Jimmy Lin**. Recounting the Courts? Toward A Text-Centered Computational Approach to Understanding the Dynamics of the Judicial System. Proceedings of the 2005 Annual Meeting of the Midwest Political Science Association, April 2005, Chicago, Illinois. (29 pages) (Full Paper)

(C14) **Jimmy Lin**. Are Degree Achievements Really Achievements? Proceedings of the 9th International Symposium on Chinese Languages and Linguistics, November 2004, Taipei, Taiwan. (12 pages) (Full Paper)

(C13) $^{\alpha}$Wesley Hildebrandt, Boris Katz, and **Jimmy Lin**. Answering Definition Questions with Multiple Knowledge Sources. Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (**HLT/NAACL 2004**), pages 49–56, May 2004, Boston, Massachusetts. (Full Paper) 26%

(C12) **Jimmy Lin** and Boris Katz. Question Answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques. Proceedings of the Twelfth International Conference on Information and Knowledge Management (**CIKM 2003**), pages 116–123, November 2003, New Orleans, Louisiana. (Full Paper) (DOI: 10.1145/956863.956886) 15%

(C11) Boris Katz and **Jimmy Lin**. Organizing and Accessing a Comprehensive Knowledge Base Using the World Wide Web. Proceedings of the International Conference on Integration of Knowledge Intensive Multi-Agent System (**KIMAS 2003**), pages 529–534, September 2003, Boston, Massachusetts. (DOI: 10.1109/KIMAS.2003.1245096)

(C10) **Jimmy Lin**, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. What Makes a Good Answer? The Role of Context in Question Answering. Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (**INTERACT 2003**), pages 25–32, September 2003, Zurich, Switzerland. (Full Paper) 33%

(C9) Stefanie Tellex, Boris Katz, **Jimmy Lin**, Gregory Marton, and Aaron Fernandes. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2003**), pages 41–47, July 2003, Toronto, Ontario, Canada. (Full Paper) (Best Student Paper) (DOI: 10.1145/860435.860445) 17%

(C8) Boris Katz, Roger Hurwitz, **Jimmy Lin**, Özlem Uzuner. Better Public Policy Through Natural Language Information Access. Proceedings of the National Conference on Digital Government Research (**DG.O 2003**), pages 73–78, May 2003, Boston, Massachusetts. (Full Paper)

(C7) Boris Katz, **Jimmy Lin**, Chris Stauffer, and Eric Grimson. Answering Questions about Moving Objects in Surveillance Videos. Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering, pages 145–152, March 2003, Palo Alto, California. (Full Paper)

(C6) Boris Katz, **Jimmy Lin**, and Sue Felshin. The START Multimedia Information System: Current Technology and Future Directions. Proceedings of the International Workshop on Multimedia Information Systems (**MIS 2002**), pages 117–123, October 2002, Tempe, Arizona. (Short Paper)

(C5) Boris Katz, **Jimmy Lin**, and Dennis Quan. Natural Language Annotations for the Semantic Web. Proceedings of the International Conference on Ontologies, Databases, and Application of Semantics (**ODBASE 2002**), pages 1317–1331, September 2002, Irvine, California. (Full Paper) (DOI: 10.1007/3-540-36124-3_83)

(C4) Susan Dumais, Michele Banko, Eric Brill, **Jimmy Lin**, and Andrew Ng. Web Question Answering: Is More Always Better? Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2002**), pages 291–298, August 2002, Tampere, Finland. (Full Paper) (DOI: 10.1145/564376.564428) 20%

(C3) Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, **Jimmy Lin**, Gregory Marton, Alton Jerome McFarland, Baris Temelkuran. Omnibase: Uniform Access to Heterogeneous Data for Question Answering. Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems (**NLDB 2002**), pages 230–234, June 2002, Stockholm, Sweden. (Short Paper) (DOI: 10.1007/3-540-36271-1_23) 58%

(C2) **Jimmy Lin**. The Web as a Resource for Question Answering: Perspective and Challenges. Proceedings of the Third International Conference on Language Resources and Evaluation (**LREC 2002**), pages 2120–2127, May 2002, Canary Islands, Spain. (Full Paper)

(C1) Joyce Chai, **Jimmy Lin**, Wlodek Zadrozny, Timing Ye, Margo Budzikowska, Veronika Horvath, Nanda Kambhatla, Catherine Wolf. Comparative Evaluation of a Natural Language Dialog Based System and a Menu Driven System for Information Access: A Case Study. Proceedings of the International Conference on Multimedia Information Retrieval – Volume 2 (**RIAO 2000**), pages 1590–1600, April 2000, Paris, France. (Full Paper)

2.6 Conference Abstracts and Extended Abstracts

(A40) Josh Seltzer, Kathy Cheng, Shi Zong, and **Jimmy Lin**. Flipping the Script: Inverse Information Seeking Dialogues for Market Research. Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), pages 3380–3383, July 2022, Madrid, Spain. (DOI: 10.1145/3477495.3536326)

(A39) Tobi Adewoye, Xiao Han, Nick Ruest, Ian Milligan, Samantha Fritz, and **Jimmy Lin**. Content-Based Exploration of Archival Images Using Neural Networks. Proceedings of the 20th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020), pages 489–490, August 2020. (DOI: 10.1145/3383583.3398577)

(A38) Martin Gauch, James Bai, Juliane Mai, and **Jimmy Lin**. An Open-Source Interface to the Canadian Surface Prediction Archive. Proceedings of the 20th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020), pages 529–530, August 2020. (DOI: 10.1145/3383583.3398626)

(A37) Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and **Jimmy Lin**. Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset. Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, July 2020.

(A36) Ferhan Ture, Jinfeng Rao, Raphael Tang, and **Jimmy Lin**. Challenges and Opportunities in Understanding Spoken Queries Directed at Modern Entertainment Platforms. Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), pages 1375–1376, July 2019, Paris, France. (DOI: 10.1145/3331184.3331433)

(A35) $^{\alpha}$Ryan Clancy, Nicola Ferro, Claudia Hauff, **Jimmy Lin**, Tetsuya Sakai, and Ze Zhong Wu. The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), pages 1432–1434, July 2019, Paris, France. (DOI: 10.1145/3331184.3331647)

(A34) Ryan Deschamps, Nick Ruest, **Jimmy Lin**, Samantha Fritz, and Ian Milligan. The Archives Unleashed Notebook: Madlibs for Jumpstarting Scholarly Exploration of Web Archives. Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2019), pages 337–338, June 2019, Urbana-Champaign, Illinois. (DOI: 10.1109/JCDL.2019.00059)

(A33) Hsiu-Wei Yang, Linqing Liu, Ian Milligan, Nick Ruest, and **Jimmy Lin**. Scalable Content-Based Analysis of Images in Web Archives with TensorFlow and the Archives Unleashed Toolkit. Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2019), pages 436–437, June 2019, Urbana-Champaign, Illinois. (DOI: 10.1109/JCDL.2019.00107)

(A32) Nick Ruest, Ian Milligan, and **Jimmy Lin**. Warclight: A Rails Engine for Web Archive Discovery. Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2019), pages 442–443, June 2019, Urbana-Champaign, Illinois. (DOI: 10.1109/JCDL.2019.00110)

(A31) Ryan Deschamps, **Jimmy Lin**, Nick Ruest, Samantha Fritz, Ian Milligan. Usability, Accessibility, and Performance: Striking the Right Balance with the Archives Unleashed Toolkit. CSDH/SCHN Digital Humanities Conference 2018, May 2018, Toronto, Ontario, Canada.

(A30) Jaejun Lee, Raphael Tang, and **Jimmy Lin**. Universal Voice-Enabled User Interfaces Using JavaScript. Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion (IUI '19), pages 81–82, March 2019, Marina del Ray, California. (DOI: 10.1145/3308557.3308693)

(A29) Ajeet Grewal and **Jimmy Lin**. The Evolution of Content Analysis for Personalized Recommendations at Twitter. Proceedings of the 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), pages 1355–1356, July 2018, Ann Arbor, Michigan. (DOI: 10.1145/3209978.3210206)

(A28) Ian Milligan, **Jimmy Lin**, Jeremy Wiebe, and Alice Zhou. Exploring and Discovering Archive-It Collections with Warcbase. Digital Humanities 2016, July 2016, Krakow, Poland.

(A27) Dean F. Sittig, Allison B. McCoy, Adam Wright, and **Jimmy Lin**. Developing an Open-Source Bibliometric Ranking Website Using Google Scholar Citation Profiles for Researchers in the Field of Biomedical Informatics. Proceedings of the 14th World Congress on Medical and Health Informatics, page 1004, August 2015, São Paulo, Brazil.

(A26) **Jimmy Lin**. The Sum of All Human Knowledge in Your Pocket: Full-Text Searchable Wikipedia on a Raspberry Pi. Proceedings of the 2015 IEEE/ACM Joint Conference on Digital Libraries (JCDL 2015), pages 85–86, June 2015, Knoxville, Tennessee. (DOI: 10.1145/2756406.2756938)

(A25) **Jimmy Lin**, Kari Kraus, and Ricardo Punzalan. Supporting "Distant Reading" for Web Archives. Digital Humanities 2014, pages 239–241, July 2014, Lausanne, Switzerland.

(A24) Hannes Mühleisen, Thaer Samar, **Jimmy Lin**, and Arjen P. de Vries. Column Stores as an IR Prototyping Tool. Proceedings of the 36th European Conference on Information Retrieval (ECIR 2014), pages 789–792, April 2014, Amsterdam, The Netherlands. (DOI: 10.1007/978-3-319-06028-6_97)

(A23) Zhengzheng Xu, Dan Goldwasser, Benjamin B. Bederson, and **Jimmy Lin**. Visual Analytics of MOOCs at Maryland. Proceedings of the First ACM Conference on Learning at Scale, pages 195–196, March 2014, Atlanta, Georgia. (DOI: 10.1145/2556325.2567878)

(A22) Alan Said, Alejandro Bellogín, **Jimmy Lin**, Arjen P. de Vries. Do Recommendations Matter? News Recommendation in Real Life. Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2014), pages 237–240, February 2014, Baltimore, Maryland. (DOI: 10.1145/2556420.2556510)

(A21) Ferhan Ture, **Jimmy Lin**, and Douglas W. Oard. Looking Inside the Box: Context-Sensitive Translation for Cross-Language Information Retrieval. Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), pages 1105–1106, August 2012, Portland, Oregon. (DOI: 10.1145/2348283.2348491)

(A20) Richard McCreadie, Ian Soboroff, **Jimmy Lin**, Craig Macdonald, Iadh Ounis, and Dean McCullough. On Building a Reusable Twitter Corpus. Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), pages 1113–1114, August 2012, Portland, Oregon. (DOI: 10.1145/2348283.2348495)

(A19) Gilad Mishne and **Jimmy Lin**. Twanchor Text: A Preliminary Study of the Value of Tweets as Anchor Text. Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), pages 1159–1160, August 2012, Portland, Oregon. (DOI: 10.1145/2348283.2348518)

(A18) Nima Asadi, Donald Metzler, and **Jimmy Lin**. Cross-Corpus Relevance Projection. Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011), pages 1163–1164, July 2011, Beijing, China. (DOI: 10.1145/2009916.2010100)

(A17) Earl J. Wagner and **Jimmy Lin**. In-Depth Accounts and Passing Mentions in the News: Connecting Readers to the Context of a News Event. Proceedings of the 2011 iConference, pages 790–791, February 2011, Seattle, Washington. (DOI: 10.1145/1940761.1940906)

(A16) Michael C. Schatz, Ben Langmead, **Jimmy Lin**, Mihai Pop, and Steven L. Salzberg. Whole Genome Resequencing Analysis in the Clouds. Supercomputing 2009, November 2009, Portland, Oregon. (Poster)

(A15) Ben Langmead, Michael C. Schatz, **Jimmy Lin**, Mihai Pop, and Steven L. Salzberg. Human SNPs from Short Reads in Hours using Cloud Computing. The 9th Workshop on Algorithms in Bioinformatics, September 2009, Baltimore, Maryland. (Poster)

(A14) Judith L. Klavans, Eileen Abels, **Jimmy Lin**, Rebecca Passonneau, Carolyn Sheffield, and Dagobert Soergel. Mining Texts for Image Terms: The CLiMB Project. Digital Humanities 2009, pages 184-186, June 2009, College Park, Maryland.

(A13) G. Craig Murray, **Jimmy Lin**, John Wilbur, and Zhiyong Lu. Users' Adjustments to Unsuccessful Queries in Biomedical Search. Proceedings of the 2009 ACM/IEEE-CS Joint International Conference on Digital Libraries (JCDL 2009), pages 433–434, June 2009, Austin, Texas. (DOI: 10.1145/1555400.1555499)

(A12) Justin Grimes, Paul T. Jaeger, and **Jimmy Lin**. Weathering the Storm: The Policy Implications of Cloud Computing. Proceedings of the 2009 iConference, 2009, Chapel Hill, North Carolina. (3 pages)

(A11) Judith L. Klavans, Carolyn Sheffield, **Jimmy Lin**, and Tandeep Sidhu. Computational Linguistics for Metadata Building. Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008), page 427, Pittsburgh, Pennsylvania. (DOI: 10.1145/1378889.1378974)

(A10) G. Craig Murray, **Jimmy Lin**, and Abdur Chowdhury. Identification of User Sessions with Hierarchical Agglomerative Clustering. Proceedings of the 2006 Annual Meeting of the American Society for Information Science and Technology (ASIST 2006), November 2006, Austin, Texas. (4 pages)

(A9) G. Craig Murray, **Jimmy Lin**, and Abdur Chowdhury. Action Modeling: Using Language Models to Predict Query Behavior. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006), pages 681–682, August 2006, Seattle, Washington. (DOI: 10.1145/1148170.1148315)

(A8) **Jimmy Lin** and Dina Demner-Fushman. "Bag of Words" Is Not Enough for Strength of Evidence Classification. Proceedings of the 2005 Annual Symposium of the American Medical Informatics Association (AMIA 2005), page 1031, October 2005, Washington, D.C.

(A7) **Jimmy Lin** and G. Craig Murray. Assessing the Term Independence Assumption in Blind Relevance Feedback. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), pages 635–636, August 2005, Salvador, Brazil. (DOI: 10.1145/1076034.1076165)

(A6) Boris Katz, Roger Hurwitz, **Jimmy Lin**, Özlem Uzuner. START: A Framework for Facilitating E-Rulemaking. Proceedings of the National Conference on Digital Government Research (DG.O 2003), page 285, May 2003, Boston, Massachusetts.

(A5) **Jimmy Lin**, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, David R. Karger. The Role of Context in Question Answering Systems. Extended abstracts of the 2003 Conference on Human Factors in Computing Systems (CHI 2003), pages 1006–1007, April 2003, Fort Lauderdale, Florida. (DOI: 10.1145/765891.766119)

(A4) $^{\alpha}$David Karger, Boris Katz, **Jimmy Lin**, and Dennis Quan. Sticky Notes for the Semantic Web. Proceedings of the 2003 International Conference on Intelligent User Interfaces (IUI 2003), pages 254–256, January 2003, Miami, Florida. (DOI: 10.1145/604045.604091)

(A3) Michele Banko, Eric Brill, Susan Dumais, and **Jimmy Lin**. AskMSR: Question Answering Using the Worldwide Web. Proceedings of the 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, pages 7–8, March 2002, Palo Alto, California. (2 pages)

(A2) Boris Katz, Deniz Yuret, **Jimmy Lin**, Sue Felshin, Rebecca Schulman, Adnan Ilik, Ali Ibrahim, Philip Osafo-Kwaako. Integrating Large Lexicons and Web Resources into a Natural Language Query System. Proceedings of the International Conference on Multimedia Computing and Systems (ICMCS 1999), Volume 2, pages 255–261, June 1999, Florence, Italy. (DOI: 10.1109/MMCS.1999.778343)

(A1) Boris Katz, Deniz Yuret, **Jimmy Lin**, Sue Felshin, Rebecca Schulman, Adnan Ilik. Blitz: A Preprocessor for Detecting Context-Independent Linguistic Structures. Proceedings of the 5th Pacific Rim Conference on Artificial Intelligence (PRICAI 1998), November 1998, Singapore. (6 pages)

2.7 Workshop Papers

(W77) Minghan Li, Xueguang Ma, and **Jimmy Lin**. An Encoder Attribution Analysis for Dense Passage Retriever in Open-Domain Question Answering. Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), pages 1–11, July 2022, Seattle, Washington.

(W76) Odunayo Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and **Jimmy Lin**. AfriTeVA: Extending "Small Data" Pretraining Approaches to Sequence-to-Sequence Models. Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 126–135, July 2022, Seattle, Washington.

(W75) Akintunde Oladipo, Odunayo Ogundepo, Kelechi Ogueji, and **Jimmy Lin**. An Exploration of Vocabulary Size and Transfer Effects in Multilingual Language Models for African Languages. Proceedings of the 3rd Workshop on African Natural Language Processing (AfricaNLP 2022), April 2022.

(W74) Kelechi Ogueji, Yuxin Zhu, and **Jimmy Lin**. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resource Languages. Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 116–126, November 2021, Punta Cana, Dominican Republic.

(W73) Xinyu Zhang, Xueguang Ma, Peng Shi, and **Jimmy Lin**. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. Proceedings of 1st Workshop on Multilingual Representation Learning, pages 127–137, November 2021, Punta Cana, Dominican Republic.

(W72) Peng Shi, Rui Zhang, He Bai, and **Jimmy Lin**. Cross-Lingual Training of Dense Retrievers for Document Retrieval. Proceedings of 1st Workshop on Multilingual Representation Learning, pages 251–253, November 2021, Punta Cana, Dominican Republic.

(W71) Yue Zhang, Chengcheng Hu, Yuqi Liu, Hui Fang, and **Jimmy Lin**. Learning to Rank in the Age of Muppets: Effectiveness-Efficiency Tradeoffs in Multi-Stage Ranking. Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, pages 64–73, November 2021.

(W70) Zhiying Jiang, Raphael Tang, Ji Xin, and **Jimmy Lin**. How Does BERT Rerank Passages? An Attribution Analysis with Information Bottlenecks. Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 496–509, November 2021, Punta Cana, Dominican Republic.

(W69) Wei Zhong, Xinyu Zhang, Ji Xin, **Jimmy Lin**, and Richard Zanibbi. Approach Zero and Anserini at the CLEF-2021 ARQMath Track: Applying Substructure Search and BM25 on Operator Tree Path Tokens. Proceedings of the Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum: CEUR Workshop Proceedings Vol-2936, pages 133–156, September 2021.

(W68) Mayank Anand, Jiarui Zhang, Shane Ding, Ji Xin, and **Jimmy Lin**. Serverless BM25 Search and BERT Reranking. Proceedings of the 2nd International Conference on Design of Experimental Search & Information REtrieval Systems (DESIRES 2021): CEUR Workshop Proceedings Vol-2950, pages 3–9, September 2021.

(W67) **Jimmy Lin**, Xueguang Ma, Joel Mackenzie, and Antonio Mallia. On the Separation of Logical and Physical Ranking Models for Text Retrieval Applications. Proceedings of the 2nd International Conference on Design of Experimental Search & Information REtrieval Systems (DESIRES 2021): CEUR Workshop Proceedings Vol-2950, pages 176–178, September 2021.

(W66) Sheng-Chieh Lin, Jheng-Hong Yang, and **Jimmy Lin**. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 163–173, August 2021.

(W65) Xinyu Zhang, Ji Xin, Andrew Yates, and **Jimmy Lin**. Bag-of-Words Baselines for Semantic Code Search. Proceedings of the 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021), pages 88–94, August 2021.

(W64) He Bai, Peng Shi, **Jimmy Lin**, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu, and Ming Li. Semantics of the Unwritten: The Effect of End of Paragraph and Sequence Tokens on Text Generation with GPT2. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 148–162, August 2021. (DOI: `10.18653/v1/2021.acl-srw.16`)

(W63) Rodrigo Nogueira, Zhiying Jiang, and **Jimmy Lin**. Investigating the Limitations of Transformers with Simple Arithmetic Tasks. Proceedings of the 1st Mathematical Reasoning in General Artificial Intelligence Workshop at ICLR 2021, May 2021.

(W62) Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and **Jimmy Lin**. Scientific Claim Verification with VerT5erini. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 94–103, April 2021.

(W61) Martin Gauch, Daniel Klotz, Frederik Kratzert, Grey Nearing, Sepp Hochreiter, and **Jimmy Lin**. A Machine Learner's Guide to Streamflow Prediction. NeurIPS 2020 Workshop on AI for Earth Sciences, December 2020.

(W60) Ji Xin, Rodrigo Nogueira, Yaoliang Yu, and **Jimmy Lin**. Early Exiting BERT for Efficient Document Ranking. Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pages 83–88, November 2020. (DOI: `10.18653/v1/2020.sustainlp-1.11`)

(W59) Xinyu Zhang, Andrew Yates, and **Jimmy Lin**. A Little Bit Is Worse Than None: Ranking with Limited Training Data. Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pages 107–122, November 2020. (DOI: `10.18653/v1/2020.sustainlp-1.14`)

(W58) Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, and **Jimmy Lin**. Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset. Proceedings of the 1st Workshop on Scholarly Document Processing, pages 31–41, November 2020. (DOI: `10.18653/v1/2020.sdp-1.5`)

(W57) Shane Ding, Edwin Zhang, and **Jimmy Lin**. Cydex: Neural Search Infrastructure for the Scholarly Literature. Proceedings of the 1st Workshop on Scholarly Document Processing, pages 168–173, November 2020. (DOI: `10.18653/v1/2020.sdp-1.19`)

(W56) Raphael Tang, Jaejun Lee, Afsaneh Razi, Julia Cambre, Ian Bicking, Jofish Kaye, and **Jimmy Lin**. Howl: A Deployed, Open-Source Wake Word Detection System. Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pages 61–65, November 2020. (DOI: `10.18653/v1/2020.nlposs-1.9`)

(W55) Ashutosh Adhikari, Achyudh Ram, Raphael Tang, William L. Hamilton, and **Jimmy Lin**. Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification with DocBERT. Proceedings of the 5th Workshop on Representation Learning for NLP, pages 72–77, July 2020. (DOI: `10.18653/v1/2020.repl4nlp-1.10`)

(W54) Rodrigo Nogueira, Zhiying Jiang, Kyunghyun Cho, and **Jimmy Lin**. Evaluating Pretrained Transformer Models for Citation Recommendation. Proceedings of the 10th International Workshop on Bibliometric-enhanced Information Retrieval (BIR) at ECIR 2020: CEUR Workshop Proceedings Vol-2591, pages 89–100, April 2020.

(W53) Martin Gauch, Juliane Mai, Shervan Gharari, and **Jimmy Lin**. Streamflow Prediction with Limited Spatially-Distributed Input Data. NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning, December 2019, Vancouver, British Columbia, Canada.

(W52) Ryan Clancy, Ihab F. Ilyas, and **Jimmy Lin**. Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond. Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 39–46, November 2019, Hong Kong, China. (DOI: 10.18653/v1/D19-6607)

(W51) Raphael Tang, Yao Lu, and **Jimmy Lin**. Natural Language Generation for Effective Knowledge Distillation. Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 202–208, November 2019, Hong Kong, China. (DOI: 10.18653/v1/D19-6122)

(W50) Ryan Clancy, Nicola Ferro, Claudia Hauff, **Jimmy Lin**, Tetsuya Sakai, Ze Zhong Wu. Overview of the 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). Proceedings of the Open-Source IR Replicability Challenge (OSIRRC 2019): CEUR Workshop Proceedings Vol-2409, pages 1–7, July 2019, Paris, France.

(W49) Ryan Clancy, Zeynep Akkalyoncu Yilmaz, Ze Zhong Wu, **Jimmy Lin**. University of Waterloo Docker Images for OSIRRC at SIGIR 2019. Proceedings of the Open-Source IR Replicability Challenge (OSIRRC 2019): CEUR Workshop Proceedings Vol-2409, page 36, July 2019, Paris, France.

(W48) Raphael Tang, Ashutosh Adhikari, and **Jimmy Lin**. FLOPs as a Direct Optimization Objective for Learning Sparse Neural Networks. Proceedings of the NIPS 2018 Workshop on Compact Deep Neural Networks with Industrial Applications, December 2018, Montreal, Quebec, Canada.

(W47) Youngbin Kim and **Jimmy Lin**. Serverless Data Analytics with Flint. Proceedings of the 2018 IEEE 11th International Conference on Cloud Computing (CLOUD 2018), pages 451–455, July 2018, San Francisco, California. (DOI: 10.1109/CLOUD.2018.00063)

(W46) Kareem El Gebaly, Lukasz Golab, and **Jimmy Lin**. Portable In-Browser Data Cube Exploration. Proceedings of the KDD 2017 Workshop on Interactive Data Exploration and Analytics (IDEA), pages 35–39, August 2017, Halifax, Nova Scotia, Canada.

(W45) Zhucheng Tu, Matt Crane, Royal Sequiera, Junchen Zhang, and **Jimmy Lin**. An Exploration of Approaches to Integrating Neural Reranking Models in Multi-Stage Ranking Architectures. Proceedings of the SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR'17), August 2017, Tokyo, Japan. (6 pages) (Published as arXiv:1707.08275)

(W44) Royal Sequiera, Gaurav Baruah, Zhucheng Tu, Salman Mohammed, Jinfeng Rao, Haotian Zhang, and **Jimmy Lin**. Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End Question Answering. Proceedings of the SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR'17), August 2017, Tokyo, Japan. (5 pages) (Published as arXiv:1707.07804)

(W43) Jinfeng Rao, Hua He, Haotian Zhang, Ferhan Ture, Royal Sequiera, Salman Mohammed, and **Jimmy Lin**. Integrating Lexical and Temporal Signals in Neural Ranking Models for Social Media Search. Proceedings of the SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR'17), August 2017, Tokyo, Japan. (5 pages) (Published as arXiv:1707.07792)

(W42) Leif Azzopardi, Matt Crane, Hui Fang, Grant Ingersoll, **Jimmy Lin**, Yashar Moshfeghi, Harrisen Scells, Peilin Yang, and Guido Zuccon. The Lucene for Information Access and Retrieval Research (LIARR) Workshop at SIGIR 2017. Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), pages 1429–1430, August 2017, Tokyo, Japan. (DOI: 10.1145/3077136.3084374)

(W41) Amy Wickner, Katie Shilton, Doug Oard, and **Jimmy Lin**. Protecting Secrets in Email: Archival Views on Challenges and Opportunities. PC4DS: The First International Workshop on Privacy-sensitive Collections for Digital Scholarship, August 2017, Montreal, Quebec, Canada.

(W40) Ziquan Wang, Borui Lin, Ian Milligan, and **Jimmy Lin**. Topic Shifts Between Two US Presidential Administrations. JCDL 2017 Workshop on Web Archiving and Digital Libraries, June 2017, Toronto, Ontario, Canada.

(W39) Anil Pacaci, Alice Zhou, **Jimmy Lin**, M. Tamer Özsu. Do We Need Specialized Graph Databases? Benchmarking Real-Time Social Networking Applications. Proceedings of the Fifth International Workshop on Graph Data-management Experiences & Systems (GRADES'17), Article no. 12, May 2017, Chicago, Illinois. (7 pages) (DOI: 10.1145/3078447.3078459)

(W38) **Jimmy Lin**, Zhucheng Tu, Michael Rose, and Patrick White. Prizm: A Wireless Access Point for Proxy-Based Web Lifelogging. Proceedings of the First Workshop on Lifelogging Tools and Applications (LTA 2016), pages 19–25, October 2016, Amsterdam, The Netherlands. (DOI: 10.1145/2983576.2983581)

(W37) Hua He, John Wieting, Kevin Gimpel, Jinfeng Rao, and **Jimmy Lin**. UMD–TTIC–UW at SemEval-2016 Task 1: Attention-Based Multi-Perspective Convolutional Neural Networks for Textual Similarity Measurement. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 662–667, June 2016, San Diego, California. (DOI: 10.18653/v1/S16-1170)

(W36) Douglas Oard, Katie Shilton, and **Jimmy Lin**. Evaluating Search Among Secrets. Proceedings of the 7th International Workshop on Evaluating Information Access (EVIA 2016), pages 21–24, June 2016, Tokyo, Japan.

(W35) Jiaul H. Paik and **Jimmy Lin**. Do Multiple Listeners to the Public Twitter Sample Stream Receive the Same Tweets? Proceedings of the SIGIR 2015 Workshop on Temporal, Social and Spatially-Aware Information Access, August 2015, Santiago, Chile. (4 pages)

(W34) $^{\alpha}$Jaime Arguello, Fernando Diaz, **Jimmy Lin**, and Andrew Trotman. SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015), pages 1147–1148, August 2015, Santiago, Chile. (DOI: 10.1145/2766462.2767858)

(W33) **Jimmy Lin**. Scaling Down Distributed Infrastructure on Wimpy Machines for Personal Web Archiving. Proceedings of the 24th International World Wide Web Conference Companion (WWW 2015), May 2015, pages 1351–1355, Florence, Italy. (Temporal Web Analytics Workshop 2015) (DOI: 10.1145/2740908.2741695)

(W32) Jinfeng Rao, **Jimmy Lin**, and Hanan Samet. Partitioning Strategies for Spatio-Textual Similarity Join. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial '14), November 2014, pages 40–49, Dallas, Texas. (DOI: 10.1145/2676536.2676542)

(W31) Gebrekirstos G. Gebremeskel, Jiyin He, Arjen P. de Vries, and **Jimmy Lin**. Cumulative Citation Recommendation: A Feature-aware Comparisons of Approaches. Proceedings of the 11th International Workshop on Text-based Information Retrieval at DEXA 2014, September 2014, Munich, Germany. (5 pages)

(W30) **Jimmy Lin**, Milad Gholami, and Jinfeng Rao. Infrastructure for Supporting Exploration and Discovery in Web Archives. Proceedings of the 23rd International World Wide Web Conference Companion (WWW 2014), pages 851–855, April 2014, Seoul, South Korea. (Temporal Web Analytics Workshop 2014) (DOI: 10.1145/2567948.2579045)

(W29) Nima Asadi and **Jimmy Lin**. An Exploration of Postings List Contiguity in Main-Memory Incremental Indexing. Proceedings of the WSDM 2014 Workshop on Large-Scale and Distributed Systems for Information Retrieval, February 2014, New York, New York. (6 pages)

(W28) Alan Said, **Jimmy Lin**, Alejandro Bellogín, Arjen P. de Vries. A Month in the Life of a Production News Recommender System. Proceedings of the CIKM 2013 Workshop on Living Labs for Information Retrieval Evaluation (LivingLab '13), pages 7–10, November 2013, San Francisco, California. (DOI: 10.1145/2513150.2513159)

(W27) **Jimmy Lin** and Miles Efron. Temporal Relevance Profiles for Tweet Search. Proceedings of the SIGIR 2013 Workshop on Time-Aware Information Access, August 2013, Dublin, Ireland. (4 pages)

(W26) Vladimir Eidelman, Ke Wu, Ferhan Ture, Philip Resnik, and **Jimmy Lin**. Towards Efficient Large-Scale Feature-Rich Statistical Machine Translation. Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 128–133, August 2013, Sofia, Bulgaria.

(W25) Miguel Rios and **Jimmy Lin**. Distilling Massive Amounts of Data into Simple Visualizations: Twitter Case Studies. Proceedings of the Workshop on Social Media Visualization (SocMedVis) at ICWSM 2012, pages 22–25, June 2012, Dublin, Ireland.

(W24) **Jimmy Lin**, Dmitriy Ryaboy, and Kevin Weil. Full-Text Indexing for Optimizing Selection Operations in Large-Scale Data Analytics. Proceedings of the Second International Workshop on MapReduce and Its Applications (MAPREDUCE 2011), pages 59–66, June 2011, San Jose, California. (DOI: 10.1145/1996092.1996105)

(W23) Brandyn White, Tom Yeh, **Jimmy Lin**, and Larry Davis. Web-Scale Computer Vision using MapReduce for Multimedia Data Mining. Proceedings of the Tenth International Workshop on Multimedia Data Mining (MDMKDD 2010), pages 9:1–9:10, July 2010, Washington, D.C. (DOI: 10.1145/1814245.1814254)

(W22) **Jimmy Lin** and Michael Schatz. Design Patterns for Efficient Graph Algorithms in MapReduce. Proceedings of the Eighth Workshop on Mining and Learning with Graphs Workshop (MLG-2010), pages 78–85, July 2010, Washington, D.C. (DOI: 10.1145/1830252.1830263)

(W21) **Jimmy Lin**. The Curse of Zipf and Limits to Parallelization: A Look at the Stragglers Problem in MapReduce. Proceedings of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR'09) at SIGIR 2009, July 2009, Boston, Massachusetts. (6 pages)

(W20) Chris Dyer, Aaron Cordova, Alex Mont, and **Jimmy Lin**. Fast, Easy, and Cheap: Construction of Statistical Machine Translation Models with MapReduce. Proceedings of the Third Workshop on Statistical Machine Translation, pages 199–207, June 2008, Columbus, Ohio.

(W19) **Jimmy Lin**. Exploring Large-Data Issues in the Curriculum: A Case Study with MapReduce. Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics, pages 54–61, June 2008, Columbus, Ohio.

(W18) Judith Klavans, Carolyn Sheffield, Eileen Abels, Joan Beaudoin, Laura Jenemann, **Jimmy Lin**, Tom Lippincott, Rebecca Passonneau, Tandeep Sidhu, Dagobert Soergel, Tae Yano. Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access. Proceedings of the 2nd Workshop on Language Resources for Content-Based Image Retrieval (OntoImage 2008) at LREC 2008, pages 42–47, May 2008, Marrakech, Morocco.

(W17) Judith L. Klavans, Tandeep Sidhu, Carolyn Sheffield, Dagobert Soergel, **Jimmy Lin**, Eileen Abels, and Rebecca Passonneau. Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata. Proceedings of the VISAPP 2008 Workshop on Metadata Mining for Image Understanding, pages 3–12, January 2008, Funchal, Madeira, Portugal.

(W16) Tandeep Sidhu, Judith Klavans, and **Jimmy Lin**. Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources. Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 25–32, June 2007, Prague, Czech Republic.

(W15) Georg Apitz and **Jimmy Lin**. Interfaces to Support the Scholarly Exploration of Text Collections. Proceedings of the CHI 2007 Workshop for Exploratory Search and HCI, pages 60–63, April 2007, San Jose, California.

(W14) Dina Demner-Fushman and **Jimmy Lin**. Situated Question Answering in the Clinical Domain: Selecting the Best Drug Treatment for Diseases. Proceedings of the Workshop on Task-Focused Summarization and Question Answering, pages 24–31, July 2006, Sydney, Australia

(W13) **Jimmy Lin**, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. Generative Content Models for Structural Analysis of Medical Abstracts. Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology, pages 65–72, June 2006, New York, New York.

(W12) **Jimmy Lin** and Dina Demner-Fushman. Representation of Information Needs and the Elements of Context: A Case Study in the Domain of Clinical Medicine. Proceedings of the SIGIR 2005 Workshop on Information Retrieval in Context (IRiX 2005), pages 51–53, August 2005, Salvador, Brazil.

(W11) Dina Demner-Fushman and **Jimmy Lin**. Knowledge Extraction for Clinical Question Answering: Preliminary Results. Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains, July 2005, Pittsburgh, Pennsylvania. (9 pages)

(W10) **Jimmy Lin** and Dina Demner-Fushman. Evaluating Summaries and Answers: Two Sides of the Same Coin? Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 41–48, June 2005, Ann Arbor, Michigan.

(W9) $^{\alpha}$Matthew W. Bilotti, Boris Katz, and **Jimmy Lin**. What Works Better for Question Answering: Stemming or Morphological Query Expansion? Proceedings of the SIGIR 2004 Workshop on Information Retrieval for Question Answering (IR4QA), July 2004, Sheffield, England. (7 pages)

(W8) **Jimmy Lin**. A Computational Framework for Non-Lexicalist Semantics. Proceedings of the Student Research Workshop at HLT-NAACL 2004, pages 19–24, May 2004, Boston, Massachusetts.

(W7) **Jimmy Lin**. Fine-Grained Lexical Semantic Representations and Compositionally-Derived Events in Mandarin Chinese. Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004, pages 100–107, May 2004, Boston, Massachusetts.

(W6) $^{\alpha}$Ali Ibrahim, Boris Katz, and **Jimmy Lin**. Extracting Structural Paraphrases from Aligned Monolingual Corpora. Proceedings of the Second International Workshop on Paraphrasing, pages 57–64, July 2003, Sapporo, Japan.

(W5) $^{\alpha}$Boris Katz and **Jimmy Lin**. Selectively Using Relations to Improve Precision in Question Answering. Proceedings of the EACL 2003 Workshop on Natural Language Processing for Question Answering, pages 43–50, April 2003, Budapest, Hungary.

(W4) $^{\alpha}$Boris Katz and **Jimmy Lin**. Annotating the Semantic Web Using Natural Language. COLING-02: The 2nd Workshop on NLP and XML (NLPXML-2002), September 2002, Taipei, Taiwan. (8 pages)

(W3) Boris Katz, **Jimmy Lin**, and Sue Felshin. Gathering Knowledge for a Question Answering System from Heterogeneous Information Sources. Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management, July 2001, Toulouse, France. (8 pages)

(W2) $^{\alpha}$Boris Katz and **Jimmy Lin**. REXTOR: A System for Generating Relations from Natural Language. ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, pages 67–77, October 2000, Hong Kong, China. (DOI: `10.3115/1117755.1117764`)

(W1) Joyce Chai, **Jimmy Lin**, Wlodek Zadrozny, Timing Ye, Margo Budzikowska, Veronika Horvath, Nanda Kambhatla, Catherine Wolf. Evaluation of a Natural Language Dialog Based Web Navigation System—A Case Study. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000) Interactivity Workshop, May 2000, Athens, Greece. (4 pages)

2.8 Tutorials

(T11) Andrew Yates, Rodrigo Nogueira, and **Jimmy Lin**. Pretrained Transformers for Text Ranking: BERT and Beyond. Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021), pages 2666–2668, July 2021. [Material based on (B4)]

(T10) Andrew Yates, Rodrigo Nogueira, and **Jimmy Lin**. Pretrained Transformers for Text Ranking: BERT and Beyond. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials, pages 1–4, June 2021. [Material based on (B4)]

(T9) Andrew Yates, Rodrigo Nogueira, and **Jimmy Lin**. Pretrained Transformers for Text Ranking: BERT and Beyond. Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM 2021), pages 1154–1156, March 2021. [Material based on (B4)]

(T8) Lidan Wang, **Jimmy Lin**, Donald Metzler, and Jiawei Han. Learning to Efficiently Rank on Big Data. Proceedings of the 23rd International World Wide Web Conference Companion (WWW 2014), pages 209–210, April 2014, Seoul, South Korea. (165 slides)

(T7) **Jimmy Lin**. MapReduce Algorithm Design. Tutorial at 22th International World Wide Web Conference (WWW 2013), May 2013, Rio de Janeiro, Brazil. (137 slides) [Material based on (B3)]

(T6) **Jimmy Lin**. Large-Scale Data Processing with MapReduce. Tutorial at 25th AAAI Conference on Artificial Intelligence (AAAI-11), August 2011, San Francisco, California. (213 slides) [Material based on (B3)]

(T5) **Jimmy Lin** and Chris Dyer. Data-Intensive Text Processing with MapReduce. Tutorial at the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2010), June 2010, Los Angeles, California. (131 slides) [Material based on (B3)]

(T4) **Jimmy Lin**. Care and Feeding of Hadoop Clusters. Tutorial at the 23rd Large Installation System Administration Conference (LISA 2009), November 2009, Baltimore, Maryland. (173 slides)

(T3) **Jimmy Lin**. Data-Intensive Text Processing with MapReduce. Tutorial at the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), July 2009, Boston, Massachusetts. (207 slides)

(T2) **Jimmy Lin** and Chris Dyer. Data-Intensive Text Processing with MapReduce. Tutorial Presentation at the 2009 Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2009), May 2009, Boulder, Colorado. (178 slides)

(T1) **Jimmy Lin** and Boris Katz. Question Answering Techniques for the World Wide Web. Tutorial at the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL 2003), April 2003, Budapest, Hungary. (253 slides)

2.9 TREC, TAC, and DUC Papers

TREC (Text REtrieval Conference), TAC (Text Analysis Conferences), and DUC (Document Understanding Conference) are community-wide evaluations sponsored by the U.S. National Institute for Standards and Technology (NIST). These papers are not refereed.

(R34) Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, **Jimmy Lin**, Ellen M. Voorhees, and Ian Soboroff. Overview of the TREC 2022 Deep Learning Track. Proceedings of the Thirty-First Text REtrieval Conference (TREC 2022), November 2022, Gaithersburg, Maryland.

(R33) Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and **Jimmy Lin**. Overview of the TREC 2021 Deep Learning Track. Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021), November 2021, Gaithersburg, Maryland.

(R32) Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and **Jimmy Lin**. H_2oloo at TREC 2020: When all you got is a hammer... Deep Learning, Health Misinformation, and Precision Medicine. Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020), November 2020, Gaithersburg, Maryland.

(R31) Jheng-Hong Yang, Sheng-Chieh Lin, **Jimmy Lin**, Ming-Feng Tsai, and Chuan-Ju Wang. Query and Answer Expansion from Conversation History. Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019), November 2019, Gaithersburg, Maryland.

(R30) Zeynep Akkalyoncu Yilmaz, Shengjin Wang, and **Jimmy Lin**. H_2oloo at TREC 2019: Combining Sentence and Document Evidence in the Deep Learning Track. Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019), November 2019, Gaithersburg, Maryland.

(R29) Ruifan Yu, Yuhao Xie, and **Jimmy Lin**. H_2oloo at TREC 2018: Cross-Collection Relevance Transfer for the Common Core Track. Proceedings of the Twenty-Seventh Text REtrieval Conference (TREC 2018), November 2018, Gaithersburg, Maryland.

(R28) Peilin Yang and **Jimmy Lin**. Anserini at TREC 2018: CENTRE, Common Core, and News Tracks. Proceedings of the Twenty-Seventh Text REtrieval Conference (TREC 2018), November 2018, Gaithersburg, Maryland.

(R27) Royal Sequiera, Mustafa Abualsaud, Luchen Tan, and **Jimmy Lin**. Overview of the TREC 2018 Real-Time Summarization Track. Proceedings of the Twenty-Seventh Text REtrieval Conference (TREC 2018), November 2018, Gaithersburg, Maryland.

(R26) **Jimmy Lin**, Salman Mohammed, Royal Sequiera, Luchen Tan, Nimesh Ghelani, Mustafa Abualsaud, Richard McCreadie, Dmitrijs Milajevs, and Ellen Voorhees. Overview of the TREC 2017 Real-Time Summarization Track. Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017), November 2017, Gaithersburg, Maryland. (17 pages)

(R25) **Jimmy Lin**, Adam Roegiest, Luchen Tan, Richard McCreadie, Ellen Voorhees, and Fernando Diaz. Overview of the TREC 2016 Real-Time Summarization Track. Proceedings of the Twenty-Fifth Text REtrieval Conference (TREC 2016), November 2016, Gaithersburg, Maryland. (14 pages)

(R24) Cody Buntain and **Jimmy Lin**. Burst Detection in Social Media Streams for Tracking Interest Profiles in Real Time. Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015), November 2015, Gaithersburg, Maryland. (6 pages)

(R23) **Jimmy Lin**, Miles Efron, Yulu Wang, Garrick Sherman, and Ellen Voorhees. Overview of the TREC-2015 Microblog Track. Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015), November 2015, Gaithersburg, Maryland. (7 pages)

(R22) **Jimmy Lin**, Miles Efron, Yulu Wang, and Garrick Sherman. Overview of the TREC-2014 Microblog Track. Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014), November 2014, Gaithersburg, Maryland. (8 pages)

(R21) $^{\alpha}$Alejandro Bellogín, Gebrekirstos G. Gebremeskel, Jiyin He, **Jimmy** Lin, Alan Said, Thaer Samar, Arjen P. de Vries, and Jeroen B. P. Vuurens. CWI and TU Delft at TREC 2013: Contextual Suggestion, Federated Web Search, KBA, and Web Tracks. Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013), November 2013, Gaithersburg, Maryland. (13 pages)

(R20) $^{\alpha}$**Jimmy Lin** and Miles Efron. Overview of the TREC-2013 Microblog Track. Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013), November 2013, Gaithersburg, Maryland. (5 pages)

(R19) $^{\alpha}$Ian Soboroff, Iadh Ounis, Craig Macdonald, and **Jimmy Lin**. Overview of the TREC-2012 Microblog Track. Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012), November 2012, Gaithersburg, Maryland. (7 pages)

(R18) $^{\alpha}$Iadh Ounis, Craig Macdonald, **Jimmy Lin**, and Ian Soboroff. Overview of the TREC-2011 Microblog Track. Proceedings of the Twentieth Text REtrieval Conference (TREC 2011), November 2011, Gaithersburg, Maryland. (5 pages)

(R17) Tamer Elsayed, Nima Asadi, Donald Metzler, Lidan Wang, and **Jimmy Lin**. UMD and USC/ISI: TREC 2010 Web Track Experiments with Ivory. Proceedings of the Nineteenth Text REtrieval Conference (TREC 2010), November 2010, Gaithersburg, Maryland. (7 pages)

(R16) **Jimmy Lin**, Donald Metzler, Tamer Elsayed, and Lidan Wang. Of Ivory and Smurfs: Loxodontan MapReduce Experiments for Web Search. Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009), November 2009, Gaithersburg, Maryland. (10 pages)

(R15) Saif Mohammad, Bonnie J. Dorr, Melissa Egan, Nitin Madnani, David Zajic, and **Jimmy Lin**. Multiple Alternative Sentence Compressions and Word-Pair Antonymy for Automatic Text Summarization and Recognizing Textual Entailment. Proceedings of the First Text Analysis Conference (TAC 2008), November 2008, Gaithersburg, Maryland. (8 pages)

(R14) $^{\alpha}$Hoa Trang Dang, Diane Kelly, and **Jimmy Lin**. Overview of the TREC 2007 Question Answering Track. Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007), November 2007, Gaithersburg, Maryland. (18 pages)

(R13) Nitin Madnani, **Jimmy Lin**, and Bonnie Dorr. TREC 2007 ciQA Task: University of Maryland. Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007), November 2007, Gaithersburg, Maryland. (6 pages)

(R12) Nitin Madnani, David Zajic, Bonnie Dorr, Necip Fazil Ayan, and **Jimmy Lin**. Multiple Alternative Sentence Compressions for Automatic Text Summarization. Proceedings of the 2007 Document Understanding Conference (DUC-2007) at NAACL/HLT 2007, April 2007, Rochester, New York. (8 pages)

(R11) $^{\alpha}$Hoa Trang Dang, **Jimmy Lin**, Diane Kelly. Overview of the TREC 2006 Question Answering Track. Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), pages 99–116, November 2006, Gaithersburg, Maryland.

(R10) $^{\alpha}$Douglas Oard, Tamer Elsayed, Jianqiang Wang, Yejun Wu, Pengyi Zhang, Eileen Abels, **Jimmy Lin**, and Dagobert Soergel. TREC-2006 at Maryland: Blog, Enterprise, Legal and QA Tracks. Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), November 2006, Gaithersburg, Maryland. (16 pages)

(R9) David Zajic, Bonnie Dorr, **Jimmy Lin**, Dianne O'Leary, John Conroy, Judith Schlesinger. Sentence Trimming and Selection: Mixing and Matching. Proceedings of the 2006 Document Understanding Conference (DUC-2006) at HLT/NAACL 2006, June 2006, New York, New York. (7 pages)

(R8) David Zajic, Bonnie Dorr, **Jimmy Lin**, and Richard Schwartz. Sentence Compression as a Component of a Multi-Document Summarization System. Proceedings of the 2006 Document Understanding Conference (DUC-2006) at HLT/NAACL 2006, June 2006, New York, New York. (7 pages)

(R7) $^{\alpha}$**Jimmy Lin**, Eileen Abels, Dina Demner-Fushman, Douglas W. Oard, Philip Wu, and Yejun Wu. A Menagerie of Tracks at Maryland: HARD, Enterprise, QA, and Genomics, Oh My! Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), November 2005, Gaithersburg, Maryland. (17 pages)

(R6) $^{\alpha}$Alan R. Aronson, Dina Demner-Fushman, Susanne M. Humphrey, **Jimmy Lin**, Hongfang Liu, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, and W. John Wilbur. Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents. Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), November 2005, Gaithersburg, Maryland. (10 pages)

(R5) David Zajic, Bonnie Dorr, **Jimmy Lin**, Christof Monz, and Richard Schwartz. A Sentence-Trimming Approach to Multi-Document Summarization. Proceedings of the 2005 Document Understanding Conference (DUC-2005) at HLT/EMNLP 2005, October 2005, Vancouver, British Columbia, Canada. (7 pages)

(R4) Boris Katz, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Wesley Hildebrandt, Roni Katzir, **Jimmy Lin**, Daniel Loreto, Gregory Marton, Federico Mora, Ozlem Uzuner. Answering Multiple Questions on a Topic From Heterogeneous Resources. Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004), November 2004, Gaithersburg, Maryland. (10 pages)

(R3) Boris Katz, **Jimmy Lin**, Daniel Loreto, Wesley Hildebrandt, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Gregory Marton, and Federico Mora. Integrating Web-based and Corpus-based Techniques for Question Answering. Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), pages 426–435, November 2003, Gaithersburg, Maryland.

(R2) **Jimmy Lin**, Aaron Fernandes, Boris Katz, Gregory Marton, and Stefanie Tellex. Extracting Answers from the Web Using Knowledge Annotation and Knowledge Mining Techniques. Proceedings of the Eleventh Text REtrieval Conference (TREC 2002), pages 447–456, November 2002, Gaithersburg, Maryland.

(R1) Eric Brill, **Jimmy Lin**, Michele Banko, Susan Dumais, and Andrew Ng. Data-Intensive Question Answering. Proceedings of the Tenth Text REtrieval Conference (TREC 2001), pages 393–400, November 2001, Gaithersburg, Maryland.

2.10 Notes and Other Articles

(N17) **Jimmy Lin.** A Proposed Conceptual Framework for a Representational Approach to Information Retrieval. SIGIR Forum, 55(2), Article No. 4, pages 1–29, 2021. (Not refereed)

(N16) **Jimmy Lin.** The Neural Hype, Justified! A Recantation. SIGIR Forum, 53(2):88–93, 2019. (Not refereed)

(N15) **Jimmy Lin**. The Neural Hype and Comparisons Against Weak Baselines. SIGIR Forum, 52(2):40–51, 2018. (Not refereed)

(N14) **Jimmy Lin**. Computing without Servers, V8, Rocket Ships, and Other Batsh*t Crazy Ideas in Data Systems. Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, pages 3–6, August 2018, Bertinoro, Italy. (Abstract for Keynote)

(N13) **Jimmy Lin**. Scale Up or Scale Out for Graph Processing? IEEE Internet Computing, 22(3):72–78, 2018. (Regular Column)

(N12) **Jimmy Lin**. The Lambda and the Kappa. IEEE Internet Computing, 21(5):60–66, 2017. (Regular Column)

(N11) **Jimmy Lin**. In Defense of MapReduce. IEEE Internet Computing, 21(3):94–98, 2017. (Regular Column)

(N10) **Jimmy Lin** and Kareem El Gebaly. The Future of Big Data Is… JavaScript? IEEE Internet Computing, 20(5):82–88, 2016. (Regular Column)

(N9) **Jimmy Lin**, Charles L. A. Clarke, and Gaurav Baruah. Searching from Mars. IEEE Internet Computing, 20(1):78–82, 2016. (Regular Column)

(N8) Jaime Arguello, Matt Crane, Fernando Diaz, **Jimmy Lin**, Andrew Trotman. Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). SIGIR Forum, 49(2):107–116, 2015. (Not refereed)

(N7) **Jimmy Lin**. Is Big Data a Transient Problem? IEEE Internet Computing, 19(5):86–90, 2015. (Regular Column)

(N6) Frank Hopfgartner, Allan Hanbury, Henning Müller, Noriko Kando, Simon Mercer, Jayashree Kalpathy-Cramer, Martin Potthast, Tim Gollub, Anastasia Krithara, **Jimmy Lin**, Krisztian Balog, and Ivan Eggel. Report on the Evaluation-as-a-Service (EaaS) Expert Workshop. SIGIR Forum, 49(1):57–65, 2015. (Not refereed)

(N5) **Jimmy Lin** and Miles Efron. Evaluation as a Service for Information Retrieval. SIGIR Forum, 47(2):8–14, 2013. (Not refereed)

(N4) **Jimmy Lin**. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! Big Data, 1(1):28–37, 2013. (Invited)

(N3) **Jimmy Lin** and Dmitriy Ryaboy. Scaling Big Data Mining Infrastructure: The Twitter Experience. SIGKDD Explorations, 14(2):6–19, 2012. (Invited)

(N2) $^{\alpha}$Paul Kantor and **Jimmy Lin**. Presentation Schemes for Component Analysis in IR Experiments. SIGIR Forum, 41(1):34–39, 2007. (Not refereed)

(N1) $^{\alpha}$Diane Kelly and **Jimmy Lin**. Overview of the TREC 2006 ciQA Task. SIGIR Forum, 41(1):107–116, 2007. (Not refereed)

2.11 Preprints and Technical Reports

(E133) Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, **Jimmy Lin**, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. How to Train Your DRAGON: Diverse Augmentation Towards Generalizable Dense Retrieval. arXiv:2302.07452, February 2023.

(E132) Xinyu Zhang, Minghan Li, and **Jimmy Lin**. Improving Out-of-Distribution Generalization of Neural Rerankers with Contextualized Late Interaction. arXiv:2302.06589, February 2023.

(E131) Minghan Li, Sheng-Chieh Lin, Xueguang Ma, and **Jimmy Lin**. SLIM: Sparsified Late Interaction for Multi-Vector Retrieval with Inverted Indexes. arXiv:2302.06587, February 2023.

(E130) Shi Zong, Josh Seltzer, Jiahua (Fiona) Pan, Kathy Cheng, and **Jimmy Lin**. Which Model Shall I Choose? Cost/Quality Trade-offs for Text Classification Tasks. arXiv:2301.07006, January 2023.

(E129) **Jimmy Lin**. Building a Culture of Reproducibility in Academic Research. arXiv:2212.13534, December 2022.

(E128) Luyu Gao, Xueguang Ma, **Jimmy Lin**, and Jamie Callan. Precise Zero-Shot Dense Retrieval without Relevance Labels. arXiv:2212.10496, December 2022.

(E127) Raphael Tang, Karun Kumar, Gefei Yang, Akshat Pandey, Yajie Mao, Vladislav Belyaev, Madhuri Emmadi, Craig Murray, Ferhan Ture, and **Jimmy Lin**. SpeechNet: Weakly Supervised, End-to-End Speech Recognition at Industrial Scale. arXiv:2211.11740, November 2022.

(E126) Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, **Jimmy Lin**, Yashar Mehdad, Wen-tau Yih, Xilun Chen. CITADEL: Conditional Token Interaction via Dynamic Lexical Routing for Efficient and Effective Multi-Vector Retrieval. arXiv:2211.10411, November 2022.

(E125) Peng Shi, Rui Zhang, He Bai, and **Jimmy Lin**. XRICL: Cross-lingual Retrieval-Augmented In-Context Learning for Cross-lingual Text-to-SQL Semantic Parsing. arXiv:2210.13693, October 2022.

(E124) Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and **Jimmy Lin**. Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages. arXiv:2210.09984, October 2022.

(E123) Odunayo Ogundepo, Xinyu Zhang, and **Jimmy Lin**. Better Than Whitespace: Information Retrieval for Languages without Custom Tokenizers. arXiv:2210.05481, October 2022.

(E122) Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, **Jimmy Lin**, and Ferhan Ture. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. arXiv:2210.04885, October 2022.

(E121) Sheng-Chieh Lin, Minghan Li, and **Jimmy Lin**. Aggretriever: A Simple Approach to Aggregate Textual Representation for Robust Dense Passage Retrieval. arXiv:2208.00511, July 2022.

(E120) Ji Xin, Raphael Tang, Zhiying Jiang, Yaoliang Yu, and **Jimmy Lin**. Building an Efficiency Pipeline: Commutativity and Cumulativeness of Efficiency Operators for Transformers. arXiv:2208.00483, July 2022.

(E119) Zhiying Jiang, Yiqin Dai, Ji Xin, Ming Li, and **Jimmy Lin**. Few-Shot Non-Parametric Learning with Deep Latent Variable Model. arXiv:2206.11573, June 2022.

(E118) Sheng-Chieh Lin and **Jimmy Lin**. A Dense Representation Framework for Lexical and Semantic Matching. arXiv:2206.09912, May 2022.

(E117) Nandan Thakur, Nils Reimers, and **Jimmy Lin**. Domain Adaptation for Memory-Efficient Dense Retrieval. arXiv:2205.11498, May 2022.

(E116) Minghan Li, Xinyu Zhang, Ji Xin, Hongyang Zhang, and **Jimmy Lin**. Certified Error Control of Candidate Set Pruning for Two-Stage Relevance Ranking. arXiv:2205.09638, May 2022.

(E115) Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and **Jimmy Lin**. Towards Best Practices for Training Multilingual Dense Retrieval Models. arXiv:2204.02363, April 2022.

(E114) Luyu Gao, Xueguang Ma, **Jimmy Lin**, and Jamie Callan. Tevatron: An Efficient and Flexible Toolkit for Dense Retrieval. arXiv:2203.05765, March 2022.

(E113) Wei Zhong, Jheng-Hong Yang, and **Jimmy Lin**. Evaluating Token-Level and Passage-Level Dense Retrieval Models for Math Information Retrieval. arXiv:2203.11163, March 2022.

(E112) Ellen M. Voorhees, Ian Soboroff, and **Jimmy Lin**. Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? arXiv:2201.11086, January 2022.

(E111) Jheng-Hong Yang, Xueguang Ma, and **Jimmy Lin**. Sparsifying Sparse Representations for Passage Retrieval by Top-k Masking. arXiv:2112.09628, December 2021.

(E110) Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, **Jimmy Lin**, and Guido Zuccon. Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback: A Reproducibility Study. arXiv:2112.06400, December 2021.

(E109) Sheng-Chieh Lin and **Jimmy Lin**. Densifying Sparse Representations for Passage Retrieval by Representational Slicing. arXiv:2112.04666, December 2021.

(E108) Joel Mackenzie, Andrew Trotman, and **Jimmy Lin**. Wacky Weights in Learned Sparse Representations and the Revenge of Score-at-a-Time Query Evaluation. arXiv:2110.11540, October 2021.

(E107) Minghan Li and **Jimmy Lin**. Encoder Adaptation of Dense Passage Retrieval for Open-Domain Question Answering. arXiv:2110.01599, October 2021.

(E106) **Jimmy Lin**. A Proposed Conceptual Framework for a Representational Approach to Information Retrieval. arXiv:2110.01529, October 2021.

(E105) Peng Shi, Rui Zhang, He Bai, and **Jimmy Lin**. Cross-Lingual Training with Dense Retrieval for Document Retrieval. arXiv:2109.01628, September 2021.

(E104) Xinyu Zhang, Xueguang Ma, Peng Shi, and **Jimmy Lin**. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. arXiv:2108.08787, August 2021.

(E103) **Jimmy Lin** and Xueguang Ma. A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques. arXiv:2106.14807, June 2021. [See additional details in (C197)]

(E102) Nick Craswell, Bhaskar Mitra, Daniel Campos, Emine Yilmaz, and **Jimmy Lin**. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. arXiv:2105.04021, May 2021. [Later appears as (C180)]

(E101) Sheng-Chieh Lin, Jheng-Hong Yang, and **Jimmy Lin**. Contextualized Query Embeddings for Conversational Search. arXiv:2104.08707, April 2021. [Later appears as (C188)]

(E100) Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, **Jimmy Lin**, and Allan Hanbury. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. arXiv:2104.06967, April 2021. [Later appears as (C181)]

(E99) Xueguang Ma, Kai Sun, Ronak Pradeep, and **Jimmy Lin**. A Replication Study of Dense Passage Retriever. arXiv:2104.05740, April 2021. [Later appears as (C192)]

(E98) Rodrigo Nogueira, Zhiying Jiang, and **Jimmy Lin**. Investigating the Limitations of the Transformers with Simple Arithmetic Tasks. arXiv:2102.13019, February 2021. [Later appears as (W63)]

(E97) **Jimmy Lin**, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard. arXiv:2102.12887, February 2021. [Later appears as (C178)]

(E96) **Jimmy Lin**, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations. arXiv:2102.10073, February 2021. [Later appears as (C177)]

(E95) Ronak Pradeep, Rodrigo Nogueira, and **Jimmy Lin**. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667, January 2021.

(E94) Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and **Jimmy Lin**. Scientific Claim Verification with VertT5erini. arXiv:2010.11930, October 2020.

(E93) Sheng-Chieh Lin, Jheng-Hong Yang, and **Jimmy Lin**. Distilling Dense Representations for Ranking using Tightly-Coupled Teachers. arXiv:2010.11386, October 2020.

(E92) Minghan Li, He Bai, Luchen Tan, Kun Xiong, Ming Li, and **Jimmy Lin**. Latte-Mix: Measuring Sentence Semantic Similarity with Latent Categorical Mixtures. arXiv:2010.11351, October 2020.

(E91) Martin Gauch, Frederik Kratzert, Daniel Klotz, Grey Nearing, **Jimmy Lin**, and Sepp Hochreiter. Rainfall-Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network. arXiv:2010.07921, October 2020.

(E90) **Jimmy Lin**, Rodrigo Nogueira, and Andrew Yates. Pretrained Transformers for Text Ranking: BERT and Beyond. arXiv:2010.06467, October 2020. [Working draft of (B4)]

(E89) Raphael Tang, Jaejun Lee, Afsaneh Razi, Julia Cambre, Ian Bicking, Jofish Kaye, and **Jimmy Lin**. Howl: A Deployed, Open-Source Wake Word Detection System. arXiv:2008.09606, August 2020.

(E88) Mohan Zhang, Luchen Tan, Zhengkai Tu, Zihang Fu, Kun Xiong, Ming Li, and **Jimmy Lin**. To Paraphrase or Not To Paraphrase: User-Controllable Selective Paraphrase Generation. arXiv:2008.09290, August 2020.

(E87) Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, and **Jimmy Lin**. Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset. arXiv:2007.07846, July 2020.

(E86) Martin Gauch and **Jimmy Lin**. A Data Scientist's Guide to Streamflow Prediction. arXiv:2006.12975, June 2020.

(E85) Mashrekur Rahman, Jonathan M. Frame, **Jimmy Lin**, and Grey Nearing. Hidden Stories: Topic Modeling in Hydrology Literature. EarthArXiv, doi:10.31223/osf.io/2sy7a, May 2020.

(E84) Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and **Jimmy Lin**. Query Reformulation using Query History for Passage Retrieval in Conversational Search. arXiv:2005.02230, May 2020.

(E83) He Bai, Peng Shi, **Jimmy Lin**, Luchen Tan, Kun Xiong, Wen Gao, and Ming Li. SegaBERT: Pre-training of Segment-aware BERT for Language Understanding. arXiv:2004.14996, April 2020.

(E82) Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and **Jimmy Lin**. Showing Your Work Doesn't Always Work. arXiv:2004.13705, April 2020.

(E81) Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and **Jimmy Lin**. DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference. arXiv:2004.12993, April 2020.

(E80) Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and **Jimmy Lin**. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. arXiv:2004.11339, April 2020.

(E79) Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and **Jimmy Lin**. Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned. arXiv:2004.05125, April 2020.

(E78) He Bai, Peng Shi, **Jimmy Lin**, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu, and Ming Li. Semantics of the Unwritten. arXiv:2004.02251, April 2020.

(E77) Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and **Jimmy Lin**. Conversational Question Reformulation via Sequence-to-Sequence Architectures and Pretrained Language Models. arXiv:2004.01909, April 2020.

(E76) Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and **Jimmy Lin**. TTTTTackling WinoGrande Schemas. arXiv:2003.08380, March 2020.

(E75) **Jimmy Lin**, Joel Mackenzie, Chris Kamphuis, Craig Macdonald, Antonio Mallia, Michał Siedlaczek, Andrew Trotman, and Arjen de Vries. Supporting Interoperability Between Open-Source Search Engines with the Common Index File Format. arXiv:2003.08276, March 2020.

(E74) Rodrigo Nogueira, Zhiying Jiang, and **Jimmy Lin**. Document Ranking with a Pretrained Sequence-to-Sequence Model. arXiv:2003.06713, March 2020.

(E73) Ruixue Zhang, Wei Yang, Luyun Lin, Zhengkai Tu, Yuqing Xie, Zihang Fu, Yuhao Xie, Luchen Tan, Kun Xiong, and **Jimmy Lin**. Rapid Adaptation of BERT for Information Extraction on Domain-Specific Business Documents. arXiv:2002.01861, February 2020.

(E72) **Jimmy Lin**. A Prototype of Serverless Lucene. arXiv:2002.01447, February 2020.

(E71) Rodrigo Nogueira, Zhiying Jiang, Kyunghyun Cho, and **Jimmy Lin**. Navigation-Based Candidate Expansion and Pretrained Language Models for Citation Recommendation. arXiv:2001.08687, January 2020.

(E70) Nick Ruest, **Jimmy Lin**, Ian Milligan, and Samantha Fritz. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. arXiv:2001.05399, January 2020.

(E69) Rodrigo Nogueira and **Jimmy Lin**. From doc2query to docTTTTTquery. December 2019.

(E68) Martin Gauch, Juliane Mai, and **Jimmy Lin**. The Proper Care and Feeding of CAMELS: How Limited Training Data Affects Streamflow Prediction. arXiv:1911.07249, November 2019.

(E67) Achyudh Ram, Ji Xin, Meiyappan Nagappan, Yaoliang Yu, Rocío Cabrera Lozoya, Antonino Sabetta, and **Jimmy Lin**. Exploiting Token and Path-based Representations of Code for Identifying Security-Relevant Commits. arXiv:1911.07620, November 2019.

(E66) Linqing Liu, Huan Wang, **Jimmy Lin**, Richard Socher, and Caiming Xiong. Attentive Student Meets Multi-Task Teacher: Improved Knowledge Distillation for Pretrained Models. arXiv:1911.03588, November 2019.

(E65) Jaejun Lee, Raphael Tang, and **Jimmy Lin**. What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. arXiv:1911.03090, November 2019.

(E64) Peng Shi and **Jimmy Lin**. Cross-Lingual Relevance Transfer for Document Retrieval. arXiv:1911.02989, November 2019.

(E63) Yinan Zhang, Raphael Tang, and **Jimmy Lin**. Explicit Pairwise Word Interaction Modeling Improves Pretrained Transformers for English Semantic Similarity Tasks. arXiv:1911.02847, November 2019.

(E62) Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and **Jimmy Lin**. Multi-Stage Document Ranking with BERT. arXiv:1910.14424, October 2019.

(E61) **Jimmy Lin**, Lori Paniak, and Gordon Boerke. The Performance Envelope of Inverted Indexing on Modern Hardware. arXiv:1910.11028, October 2019.

(E60) Tommaso Teofili and **Jimmy Lin**. Lucene for Approximate Nearest-Neighbors Search on Arbitrary Dense Vectors. arXiv:1910.10208, October 2019.

(E59) Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, **Jimmy Lin**, and Xu Sun. Aligning Cross-Lingual Entities with Multi-Aspect Information. arXiv:1910.06575, October 2019.

(E58) Hamidreza Shahidi, Ming Li, and **Jimmy Lin**. Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data. arXiv:1909.10158, September 2019.

(E57) Wei Yang, Kuang Lu, Peilin Yang, and **Jimmy Lin**. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. arXiv:1904.09171, April 2019.

(E56) **Jimmy Lin**. The Simplest Thing That Can Possibly Work: Pseudo-Relevance Feedback Using Text Classification. arXiv:1904.08861, April 2019.

(E55) Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and **Jimmy Lin**. DocBERT: BERT for Document Classification. arXiv:1904.08398, April 2019 (v1), updated August 2019 (v2, v3).

(E54) Rodrigo Nogueira, Wei Yang, **Jimmy Lin**, and Kyunghyun Cho. Document Expansion by Query Prediction. arXiv:1904.08375, April 2019.

(E53) Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and **Jimmy Lin**. Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering. arXiv:1904.06652, April 2019.

(E52) Peng Shi and **Jimmy Lin**. Simple BERT Models for Relation Extraction and Semantic Role Labeling. arXiv:1904.05255, April 2019.

(E51) Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and **Jimmy Lin**. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. arXiv:1903.12136, March 2019.

(E50) Wei Yang, Haotian Zhang, and **Jimmy Lin**. Simple Applications of BERT for Ad Hoc Document Retrieval. arXiv:1903.10972, March 2019.

(E49) Michael Azmy, Peng Shi, **Jimmy Lin**, and Ihab F. Ilyas. Matching Entities Across Different Knowledge Graphs with Graph Embeddings. arXiv:1903.06607, March 2019.

(E48) Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and **Jimmy Lin**. End-to-End Open-Domain Question Answering with BERTserini. arXiv:1902.01718, February 2019.

(E47) Raphael Tang, Gefei Yang, Hong Wei, Yajie Mao, Ferhan Ture, and **Jimmy Lin**. Streaming Voice Query Recognition using Causal Convolutional Recurrent Neural Networks. arXiv:1812.07754, December 2018.

(E46) Peng Shi, Jinfeng Rao, and **Jimmy Lin**. Simple Attention-Based Representation Learning for Ranking Short Social Media Posts. arXiv:1811.01013, November 2018.

(E45) Raphael Tang and **Jimmy Lin**. Progress and Tradeoffs in Neural Language Models. arXiv:1811.00942, November 2018.

(E44) Jaejun Lee, Raphael Tang, and **Jimmy Lin**. JavaScript Convolutional Neural Networks for Keyword Spotting in the Browser: An Experimental Analysis. arXiv:1810.12859, October 2018.

(E43) Raphael Tang and **Jimmy Lin**. Adaptive Pruning of Neural Language Models for Mobile Devices. arXiv:1809.10282, September 2018.

(E42) **Jimmy Lin** and Peilin Yang. Repeatability Corner Cases in Document Ranking: The Impact of Score Ties. arXiv:1807.05798, May 2018.

(E41) Jinfeng Rao, Wei Yang, Yuhao Zhang, Ferhan Ture, and **Jimmy Lin**. Multi-Perspective Relevance Matching with Hierarchical ConvNets for Social Media Search. arXiv:1805.08159, May 2018.

(E40) Kareem El Gebaly and **Jimmy Lin**. In-Browser Split-Execution Support for Interactive Analytics in the Cloud. arXiv:1804.08822, April 2018.

(E39) Youngbin Kim and **Jimmy Lin**. Serverless Data Analytics with Flint. arXiv:1803.06354, March 2018. [Later appears as (W47)]

(E38) Salman Mohammed, Peng Shi, and **Jimmy Lin**. Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks. arXiv:1712.01969, December 2017. [Later appears as (C120)]

(E37) Raphael Tang, Weijie Wang, Zhucheng Tu, and **Jimmy Lin**. An Experimental Analysis of the Power Consumption of Convolutional Neural Networks for Keyword Spotting. arXiv:1711.00333, October 2017. [Later appears as (C117)]

(E36) Raphael Tang and **Jimmy Lin**. Deep Residual Learning for Small-Footprint Keyword Spotting. arXiv:1710.10361, October 2017. [Later appears as (C116)]

(E35) Raphael Tang and **Jimmy Lin**. Honk: A PyTorch Reimplementation of Convolutional Neural Networks for Keyword Spotting. arXiv:1710.06554, October 2017.

(E34) Siddhartha Sahu, Amine Mhedhbi, Semih Salihoglu, Jimmy Lin, M. Tamer Özsu. The Ubiquity of Large Graphs and Surprising Challenges of Graph Processing: A User Survey. arXiv:1709.03188, September 2017. [Later appears as (J41)]

(E33) Jinfeng Rao, Ferhan Ture, Hua He, Oliver Jojic, **Jimmy Lin**. Talking to Your TV: Context-Aware Voice Search with Hierarchical Recurrent Neural Networks. arXiv:1705.04892, April 2017. [Later appears as (C113)]

(E32) Salman Mohammed, Nimesh Ghelani, and **Jimmy Lin**. Distant Supervision for Topic Classification of Tweets in Curated Streams. arXiv:1704.06726, April 2017.

(E31) Joel Mackenzie, J. Shane Culpepper, Roi Blanco, Matt Crane, Charles L. A. Clarke, and **Jimmy Lin**. Efficient and Effective Tail Latency Minimization in Multi-Stage Retrieval Systems. arXiv:1704.03970, April 2017. [Later appears as (C115)]

(E30) $^{\alpha}$Charles L. A. Clarke, Gordon V. Cormack, **Jimmy Lin**, and Adam Roegiest. Ten Blue Links on Mars. arXiv:1610.06468, October 2016. [Later appears as (C99)]

(E29) J. Shane Culpepper, Charles L. A. Clarke, **Jimmy Lin**. Dynamic Trade-Off Prediction in Multi-Stage Retrieval Systems. arXiv:1610.02502, October 2016. [Later appears as (C95)]

(E28) Luchen Tan, **Jimmy Lin**, Adam Roegiest, and Charles L. A. Clarke. The Effects of Latency Penalties in Evaluating Push Notification Systems. arXiv:1606.03066, May 2016.

(E27) Kareem El Gebaly and **Jimmy Lin**. Afterburner: The Case for In-Browser Analytics. arXiv:1605.04035, May 2016. [Later appears as (C100)]

(E26) Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, **Jimmy Lin**, Simon Mercer, Martin Potthast. Evaluation-as-a-Service: Overview and Outlook. arXiv:1512.07454, December 2015. [Later appears as (J43)]

(E25) Cody Buntain, **Jimmy Lin**, and Jennifer Golbeck. Learning to Discover Key Moments in Social Media Streams. arXiv:1508.00488, August 2015. (10 pages) [Later appears as (C76)]

(E24) **Jimmy Lin**. On the Feasibility and Implications of Self-Contained Search Engines in the Browser. arXiv:1410.4500, October 2014. (6 pages) [Later appears as (C74)]

(E23) Sarah Weissman, Samet Ayhan, Joshua Bradley, and **Jimmy Lin**. Identifying Duplicate and Contradictory Information in Wikipedia. arXiv:1406.1143, June 2014. (4 pages) [Later appears as (C70)]

(E22) Abdul Quamar, Amol Deshpande, and **Jimmy Lin**. NScale: Neighborhood-centric Large-Scale Graph Analytics in the Cloud. arXiv:1405.1499, May 2014. (12 pages) [Later appears as (J36)]

(E21) Nima Asadi and **Jimmy Lin**. Fast, Incremental Inverted Indexing in Main Memory for Web-Scale Collections. arXiv:1305.0699, May 2013. (10 pages) [Later appears as (W29)]

(E20) **Jimmy Lin**. Monoidify! Monoids as a Design Principle for Efficient MapReduce Algorithms. arXiv:1304.7544, April 2013. (3 pages)

(E19) Nima Asadi, **Jimmy Lin**, and Michael Busch. Dynamic Memory Allocation Policies for Postings in Real-Time Twitter Search. arXiv:1302.5302, February 2013. (10 pages) [Later appears as (C60)]

(E18) Nima Asadi, **Jimmy Lin**, and Arjen P. de Vries. Runtime Optimizations for Prediction with Tree-Based Models. arXiv:1212.2287, December 2012. (9 pages) [Later appears as (J32)]

(E17) Gilad Mishne, Jeff Dalton, Zhenghua Li, Aneesh Sharma, and **Jimmy Lin**. Fast Data in the Era of Big Data: Twitter's Real-Time Related Query Suggestion Architecture. arXiv:1210.7350, October 2012. (11 pages) [Later appears as (C55)]

(E16) **Jimmy Lin**. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! arXiv:1209.2191, September 2012. (8 pages) [Later appears as (N4)]

(E15) **Jimmy Lin** and Gilad Mishne. A Study of "Churn" in Tweets and Real-Time Search Queries (Extended Version). arXiv:1205.6855, May 2012. (8 pages) [Extended version of (C47)]

(E14) Tamer Elsayed, Ferhan Ture, and **Jimmy Lin**. Brute-Force Approaches to Batch Retrieval: Scalable Indexing with MapReduce, or Why Bother? Technical Report HCIL-2010-23, University of Maryland, College Park, October 2010. (20 pages)

(E13) Di-Wei Huang and **Jimmy Lin**. Scaling Populations of a Genetic Algorithm for Job Shop Scheduling Problems using MapReduce. Technical Report HCIL-2010-14, University of Maryland, College Park, June 2010. (6 pages) [Later appears as (C37)]

(E12) Alexander J. Quinn, Benjamin B. Bederson, Tom Yeh, and **Jimmy Lin**. CrowdFlow: Integrating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility. Technical Report HCIL-2010-09, University of Maryland, College Park, May 2010. (8 pages)

(E11) **Jimmy Lin**, Anand Bahety, Shravya Konda, and Samantha Mahindrakar. Low-Latency, High-Throughput Access to Static Global Resources within the Hadoop Framework. Technical Report HCIL-2009-01, University of Maryland, College Park, January 2009. (15 pages)

(E10) **Jimmy Lin**. Scalable Language Processing Algorithms for the Masses: A Case Study in Computing Word Co-occurrence Matrices with MapReduce. Technical Report HCIL-2008-28, University of Maryland, College Park, June 2008. (9 pages) [Later appears as (C31)]

(E9) $^{\alpha}$**Jimmy Lin** and Mark D. Smucker. How Do Users Find Things with PubMed? Towards Automatic Utility Evaluation with User Simulations. Technical Report LAMP-TR-148/HCIL-2008-07, University of Maryland, College Park, February 2008. (17 pages) [Later appears as (C30)]

(E8) Timothy Hawes, **Jimmy Lin**, and Philip Resnik. Elements of a Computational Model for Multi-Party Discourse: The Turn-Taking Behavior of Supreme Court Justices. Technical Report LAMP-TR-147/HCIL-2008-02, University of Maryland, College Park, January 2008. (12 pages) [Later appears as (J21)]

(E7) **Jimmy Lin**. PageRank without Hyperlinks: Reranking with Related Document Networks. Technical Report LAMP-TR-146/HCIL-2008-01, University of Maryland, College Park, January 2008. (12 pages) [Later appears as (J11)]

(E6) **Jimmy Lin**, Michael DiCuccio, Vahan Grigoryan, and W. John Wilbur. Exploring the Effectiveness of Related Article Search in PubMed. Technical Report LAMP-TR-145/CS-TR-4877/UMIACS-TR-2007-36/HCIL-2007-10, University of Maryland, College Park, July 2007. (11 pages) [Later appears as (J13)]

(E5) **Jimmy Lin.** The Role of Information Retrieval in Answering Relationship Questions. Technical Report LAMP-TR-130/CS-TR-4787/UMIACS-TR-2006-11, University of Maryland, College Park, February 2006. (11 pages) [Later appears as (C21)]

(E4) **Jimmy Lin** and Dina Demner-Fushman. Will Pyramids Built of Nuggets Topple Over? Technical Report LAMP-TR-127/CS-TR-4771/UMIACS-TR-2005-71, University of Maryland, College Park, December 2005. (10 pages) [Later appears as (C17)]

(E3) Grazia Russo-Lassner, **Jimmy Lin**, and Philip Resnik. A Paraphrase-Based Approach to Machine Translation Evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, August 2005. (11 pages)

(E2) **Jimmy Lin** and Dina Demner-Fushman. Automatically Evaluating Answers to Definition Questions. Technical Report LAMP-TR-119/CS-TR-4695/UMIACS-TR-2005-04, University of Maryland, College Park, February 2005. (12 pages) [Later appears as (C16)]

(E1) **Jimmy Lin**. Evaluation of Resources for Question Answering Evaluation. Technical Report LAMP-TR-118/CS-TR-4693/UMIACS-TR-2005-03, University of Maryland, College Park, February 2005. (12 pages) [Later appears as (C15)]

3 Patents

(P1) Gilad Mishne and Jimmy Lin (Inventors); Twitter Inc. (Assignee). Search Relevance Using Messages of a Messaging Platform. US Patent 9,836,461. December 5, 2017.

4 Open-Source Software

This section lists open-source software I have developed or contributed to, including implementations by students and postdocs.

(S31) Chatty Goose: an open-source Python conversational search framework that provides strong, reproducible reranking pipelines built on recent advances in neural models.

https://github.com/castorini/chatty-goose
Initial release: July 2021
Described in (C176).

(S30) Howl: wake word detection modeling for Firefox Voice and other open datasets.

http://howl.ai/
Initial release: August 2020
Described in (W56).

(S29) Pyserini: a Python toolkit for reproducible information retrieval research with sparse and dense representations.

http://pyserini.io/
Initial release: November 2019
Described in (C161), (C177), (E96).

(S28) Honkling: keyword spotting using TensorflowJS in the browser.

http://honkling.ai/
Initial release: May 2018
Described in (C142), (E44).

(S27) Honk: PyTorch implementations of neural network models for keyword spotting.

http://honk.ai/
Initial release: September 2017
Provides implementation for (C116), (C117), (E35), (E36), (E37).

(S26) The Archives Unleashed Toolkit: an open-source toolkit for analyzing web archives (successor to Warcbase).

http://archivesunleashed.org/
Initial release: July 2017
Described in (C162).

(S25) BuboQA: simple question answering over knowledge graphs.

http://buboqa.io/
Initial release: June 2017
Provides implementation for (C120), (E38).

(S24) Castor: PyTorch deep learning models for text processing.

http://castor.ai/
Initial release: March 2017
Provides implementation for (C104), (C105), (W44), (W45).

(S23) GraphJet: a real-time graph processing library.

http://graphjet.io/
Initial release: July 2016
Provides implementation for (J38).

(S22) Afterburner: interactive in-browser SQL analytics.

http://afterburnerdb.io/
Initial release: May 2016
Provides implementation for (C100), (W46), (E27), (E40).

(S21) Bespin: reference implementations of data-intensive algorithms in MapReduce and Spark.

http://bespin.io/
Initial release: November 2015
Used in data-intensive distributed processing course at the University of Waterloo.

(S20) Anserini: an open-source information retrieval toolkit built on Lucene.

http://anserini.io/
Initial release: October 2015
Described in (J42), (C102), (C137).

(S19) TS4: Tweet Streaming Selective Search with Spark.

https://github.com/ylwang99/TS4
Initial release: September 2015
Provides implementation for (C98).

(S18) IR-Reproducibility: the open-source information retrieval reproducibility challenge.

https://github.com/lintool/IR-Reproducibility
Initial release: May 2015
Code for efforts described in (C78), (C153), (W34), (N8).

(S17) C-bfscan: document retrieval using brute force scans.

https://github.com/ylwang99/c-bfscan
Initial release: November 2014
Provides implementation for (C73).

(S16) JASS: anytime ranking for impact-ordered indexes.

https://github.com/lintool/JASS
Initial release: November 2014
Provides implementation for (J39), (C75), (C96), (C97).

(S15) Hone: a MapReduce execution engine for multi-core, shared-memory architectures.

https://github.com/lintool/hadoop1
Initial release: September 2013
Provides implementation for (J27), (C61).

(S14) Warcbase: an open-source platform for managing web archives built on Hadoop and HBase.

http://warcbase.org/
Initial release: July 2013
Provides implementation for (J40), (A25), (A28), (W30), (W33).

(S13) ClueWeb tools: tools for manipulating the ClueWeb crawls.

http://clueweb.org/
Initial release: July 2013
Provides implementation for (C66), (A24).

(S12) WikiClean: a Java Wikipedia markup to plain text converter.

http://wikiclean.org/
Initial release: June 2013
Used for data cleaning in (C70).

(S11) OptTrees: architecture-conscious implementation of regression trees.

https://github.com/lintool/OptTrees
Initial release: April 2013
Provides implementation for (J32), (C52), (E18).

(S10) Wikiduper: near-duplicate sentence detection on Wikipedia.

https://github.com/seweissman/wikiduper
Initial release: April 2013
Provides implementation for (C70).

(S9) Zambezi: a multi-stage, in-memory search engine for streaming documents.

https://github.com/lintool/Zambezi
Initial release: October 2012
Provides implementation for (J26), (W29).

(S8) Cassovary: a simple big graph processing library for the JVM.

http://cassovary.io/
Initial release: February 2012
Provides implementation for (C53).

(S7) Mr.LDA: flexible, scalable, multilingual topic modeling using variational inference in MapReduce.

http://mrlda.cc/
Initial release: January 2012

(S6) Twitter tools: tools to support the TREC Microblog tracks.

http://twittertools.cc/
Initial release: May 2011
Provides implementation of tools and infrastructure described in (C64), (C65), (N5), (R18), (R19), (R20), (R22), (R23), (R25), (R26), (R27).

(S5) Ivory: a Hadoop toolkit for web-scale information retrieval research.

http://ivory.cc/
Initial release: August 2009
Provides implementations for (J28), (J33), (C36), (C37), (C39), (C40), (C49), (C50), (C51), (A21).

(S4) Twitter Commons: Twitter common libraries for Python and the Java Virtual Machine.

http://twitter.github.io/commons/
Initial release: June 2009
Contributed to open-source common libraries for the JVM used by software engineers at Twitter. Library is defunct but see above link to documentation in the Internet Archive's Wayback Machine.

(S3) Cloud[9]: a Hadoop toolkit for working with big data.

http://cloud9lib.org/
Initial release: October 2007
Provides implementations described in (B3) and used as a teaching tool.

(S2) Aranea: a system for redundancy-based factoid question answering.

http://cs.uwaterloo.ca/~jimmylin/resources.html
Initial release: June 2005
The system described in (J6); reimplementation of (C4), (A3), and (R1).

(S1) Pourpre: a program for automatically evaluating answers to complex questions.

http://cs.uwaterloo.ca/~jimmylin/resources.html
Initial release: May 2005
Provides implementation of the evaluation techniques described in (J3) and (C16).

5 Prizes and Awards

- ACM Fellow (Class of 2022).
- 2021 Outstanding Performance Award.
- 2018 Outstanding Performance Award: "Starting May 2005, the University of Waterloo established an Outstanding Performance Fund to reward faculty members for outstanding contribution in teaching and scholarship."
- Publication (C125) describes research that contributed to an Emmy Award for Technology and Engineering presented to Comcast, recognizing the Xfinity X1 Voice Remote and the platform that powers it for "Contextual Voice Navigation for Discovering and Interacting with TV Content".
- Anserini (S20) was recognized as the "Best Open-Source Project" at Search Solutions 2018, sponsored by the British Computer Society Information Retrieval Specialist Group (BCS-IRSG).
- Winner of the 2014 University System of Maryland (USM) Regents' Faculty Award for Research, Scholarship, or Creative Activity: "The awards are the highest honor presented by the board to exemplary faculty members."
- University of Maryland Research Leader (2007—2010, 2013), for exceptional external sponsored research funding.
- Recognized as Outstanding Area Chair at EMNLP 2020.
- Publication (J51): 2023 Award for Best Case Study, Journal of Hydrologic Engineering.
- Publication (C204): Best paper from ARQMath-3 @ CLEF 2022.
- Publication (C140): Best Short Paper Honorable Mention at SIGIR 2019.
- Publication (C129): Best Reproducibility Paper — Honourable Mention at ECIR 2019.
- Publication (J41): Best Paper at VLDB 2018.
- Publication (C110): Best Short Paper at ICTIR 2017.
- Publication (C95): Best Paper at ADCS 2016.
- Publication (C91): Best Short Paper at ICTIR 2016.
- Publication (C9): Best Student Paper at SIGIR 2003.

6 Presentations and Publicity

6.1 Keynotes and Invited Talks

- "Efficiency Tradeoffs in the Design of Neural Search Systems"
  Keynote at the NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing, Online (12/2/2022)
- "A Conceptual Framework for a Representational Approach to Information Retrieval"
  Distinguished Lecture, Pinterest Labs Tech Talk Series, Online (4/5/2022)
- "The Attack of the Muppets: Data Management in the Era of Pretrained Transformers"
  Keynote at the 37th IEEE International Conference on Data Engineering (ICDE 2021), Online (4/21/2021)
- "Domain Adaptation to Scientific Texts and the Limits of Scale?"
  Keynote at the 11th International Workshop on Bibliometric-enhanced Information Retrieval, Online (4/1/2021)
- "Natural Language Processing and Information Retrieval: Together at Last"
  Invited Talk at the SIGIR 2020 Summer School, Online (7/25/2020)
- "Computing without Servers, V8, Rocket Ships, and Other Batsh*t Crazy Ideas in Data Systems"
  Keynote at the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy (8/29/2018)
- "Lessons Learned From Building Big Data Systems at Twitter"
  Keynote at the Scottish Informatics & Computer Science Alliance 2015 PhD Conference, University of Glasgow, Scotland (6/24/2015)
- "Scaling Big Data Mining Infrastructure: The Twitter Experience"
  Keynote at KDD 2014 Workshop on Big Data, Streams and Heterogeneous Source Mining, New York, New York (8/24/2014)
- "The Evolution of WTF: Follower Recommendation Services at Twitter"
  Invited talk at SIGIR Symposium on IR in Practice 2014 (SIRIP 2014), Gold Coast, Australia (7/7/2014)
- "Tackling Complex Problems with NLP, Big Data, and Visualization"
  Keynote at ACL 2014 Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland (6/27/2014)
- "Real-Time Search Infrastructure at Twitter"
  Invited talk at the ECIR 2013 Industry Day, Moscow, Russia (3/27/2013)
- "Graph Processing at Twitter: User Recommendations and Personalization"
  Keynote at the VLDB 2012 Workshop on Personalized Access, Profile Management, and Context Awareness in Databases, Istanbul, Turkey (8/31/2012)
- "Infrastructure and Challenges for Real-Time Discovery at Twitter"
  Keynote at the ICWSM 2012 Workshop on Real-Time Analysis and Mining of Social Streams, Dublin, Ireland (6/4/2012)
- "Flexibility without Anarchy: Analytics Infrastructure at Twitter"
  Keynote at the SIGMOD 2012 Workshop on Scalable Workflow Enactment Engines and Technologies (SWEET'12), Scottsdale, Arizona (5/20/2012)
- "Graph Processing at Twitter"
  Keynote at the SIGMOD 2012 Workshop on Databases and Social Networks (DBSocial 2012), Scottsdale, Arizona (5/20/2012)
- "Twitter and Data Science"
  Industrial Keynote at the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, Spain (7/19/2011)

- "Large-Scale Data Processing with MapReduce"
  Keynote lecture at the SIKS/BiGGrid Advanced Tutorial on Big Data, University of Twente, The Netherlands (11/30/2011)

6.2 Seminars and Other External Talks

- "ChatGPT: What executives really need to know", WatSPEED Webinar (2/16/2023)
- "A Conceptual Framework for a Representational Approach to Information Retrieval"
  Presentations with similar titles and content given in the following venues:
  - Vectara, Online (12/8/2022)
  - Manulife, Online (10/11/2022)
  - Faire, Online (10/7/2022)
  - ICML 2022 Workshop on Knowledge Retrieval and Language Models, Online (7/22/2022)
  - Forum for Information Retrieval Evaluation (FIRE), Online (12/13/2021)
  - The University of Queensland, Online (4/7/2022)
  - IIT Kharagpur, Online (4/20/2022)
- "Domain Adaptation to Scientific Texts and the Limits of Scale?"
  University College London, Online (10/03/2021)
- "Pretrained Transformers for Text Ranking: BERT and Beyond"
  RMIT University, Online (10/28/2020)
- "Searching the COVID-19 Open Research Dataset (CORD-19) with the Neural Covidex"
  Data Science @ Regensburg, Online (5/27/2020)
- "Rigor in Deep Learning for NLP and IR"
  Presentations with similar titles and content given in the following venues:
  - Max Planck Institute for Informatics, Saarbrücken, Germany (4/12/2019)
  - Academia Sinica, Taipei, Taiwan (8/7/2019)
  - National Taiwan University, Taipei, Taiwan (8/16/2019)
- "The Academic: There and Back Again."
  Presentations with similar titles and content given in the following venues:
  - Thomson Reuters, Toronto (11/6/2017)
  - Carnegie Mellon University (10/7/2016)
  - Georgetown University (11/14/2016)
  - Harvard University (11/28/2016)
  - MIT (11/29/2016)
- "Dogma and Heresy in Information Retrieval"
  Cheriton Research Symposium, David R. Cheriton School of Computer Science, University of Waterloo, Canada (9/15/2015)
- "Building Effective and Efficient Information Retrieval Systems"
  Presentations with similar titles and content given in the following venues:
  - University of Glasgow, Scotland (6/26/2015)
  - University of Edinburgh, Scotland (6/22/2015)
- "Warcbase: Building a Scalable Web Archiving Platform on Hadoop and HBase"
  Presentations with similar titles and content given in the following venues:
  - Archival Education and Research Institute (AERI), College Park, Maryland (7/14/2015)

  - Columbia University (6/4/2015)
  - HBaseCon 2015, San Francisco, California (5/7/2015)
  - International Internet Preservation Consortium (IIPC) 2015 General Assembly, Stanford University, California (4/28/2015)

- "Effective and Efficient Information Access Systems for Making Sense of Big Data"
  Presentations with similar titles and content given in the following venues:
  - NYU Poly (3/11/2015)
  - University of Waterloo (2/12/2015)

- "The Evolution of WTF: Follower Recommendation Services at Twitter"
  Presentations with similar titles and content given in the following venues:
  - Washington Post, Washington, D.C. (11/30/2018)
  - University of Waterloo (2/13/2015)
  - Technische Universiteit Delft, Delft, The Netherlands (12/9/2014)
  - NYU Poly (7/30/2014)

- "Infrastructure for Supporting Exploration and Discovery in Web Archives"
  Presentations with similar titles and content given in the following venues:
  - Society of American Archivists, Web Archiving Roundtable, Washington, D.C. (8/13/2014)
  - Digital Preservation 2014, Washington, D.C. (7/23/2014)
  - Working with Internet Archives for Research (WIRE) Workshop, Boston, Massachusetts (6/18/2014)

- "Making Sense of Big Data"
  Presentations with similar titles and content given in the following venues:
  - University of Michigan, School of Information (10/8/2014)
  - University of California, Berkeley, Information School (2/24/2014)

- "HBase and Warcbase"
  SCAPE Hackathon, Vienna (12/3/2013)
- "Real-Time Search a Twitter"
  Presentations with similar titles and content given in the following venues:
  - Gavagai AB, Stockholm, Sweden (8/9/2013)
  - Universiteit van Amsterdam, Amsterdam, The Netherlands (8/6/2013)
  - Technische Universität Berlin, Berlin, Germany (7/25/2013)
  - Hasso-Plattner-Institut, University of Potsdam, Potsdam, Germany (7/23/2013)
  - Technische Universiteit Delft, Delft, The Netherlands (6/18/2013)

- "Overview of Graph Processing at Twitter"
  Presentations with similar titles and content given in the following venues:
  - Recommender Systems Meetup, Amsterdam, The Netherlands (8/19/2013)
  - plista GmbH, Berlin, Germany (7/24/2013)
  - The Netherlands Hadoop Users Group (NL HUG), Amsterdam, The Netherlands (3/19/2013)

- "Scaling Big Data Mining Infrastructure: The Twitter Experience"
  Presentations with similar titles and content given in the following venues:
  - Sydney Big Data Meetup, Sydney, Australia (7/4/2014)
  - Data Science DC Meetup, Washington, D.C. (10/16/2013)
  - Belgian Big Data Meetup, Ghent, Belgium (7/16/2013)

  - 2013 Hadoop Summit Europe, Amsterdam, The Netherlands (3/21/2013)
  - LivingSocial Inc., Washington, D.C. (3/12/2013)
  - NYU Poly, New York (2/25/2013)

- "Training a Smarter Pig: Large-Scale Machine Learning at Twitter"
  2012 Hadoop Summit, San Jose, California (6/13/2012)
- "Overview of the Analytics Stack at Twitter"
  University of California, Irvine (5/25/2012)
- "Twitter: Discovering the Global Conversation"
  University of California, Irvine (5/25/2012)
- "Big Data Genomics"
  Molecular Med Tri-Con 2013, San Francisco, California (2/19/2012)
- "Hadoop/MapReduce as a Platform for Data-Intensive Computing"
  Presentations with similar titles and content given in the following venues:
  - Lawrence Livermore National Laboratory, Livermore, California (1/16/2013)
  - Laboratory for Telecommunications Sciences, College Park, Maryland (2/9/2012)
  - Universiteit van Amsterdam, The Netherlands (12/02/2011)
  - Georgetown University (9/23/2011)
  - Cyberinfrastructure Days 2010, University of Michigan (11/3/2010)
- "Reflections on the Role of Industry and Academia in Social Media Research"
  National Science Foundation, Arlington, Virginia (9/1/2011)
- "Research in Cloud Computing"
  As part of "The Cloud of Things—The Next Phase of Computing", a Congressional Luncheon Briefing in conjunction with the Congressional Research & Development Caucus (9/2011)
- "Twitter: What's Happening?"
  University of Michigan (11/4/2010)
- "Design Patterns for MapReduce"
  Berkeley RADLab (10/6/2010)
- "Embracing the Data Deluge: Data-Intensive Computing for the Masses"
  Microsoft Faculty Summit 2010, Seattle, Washington (7/13/2010)
- "Clouds, Maryland, and NARA: Beginning a Dialogue"
  National Archives and Records Administration, College Park, Maryland (3/31/2010)
- "Design Patterns for Efficient Graph Algorithms in MapReduce"
  2010 Hadoop Summit, Santa Clara, California (6/29/2010)
- "Why Cloud Computing? The Academic Perspective"
  UMD IT Innovation Summit, College Park, Maryland (1/20/2010)
- "Research and Education in the Clouds: Experiences at the University of Maryland"
  Google Tech Talk, Mountain View, California (5/6/2009)
- "Connecting People with Relevant Information: Knowledge-Driven and Data-Driven Approaches to Information Retrieval"
  National Science Foundation, Arlington, Virginia (4/15/2009)
- "Genomic Sequence Analysis with Cloud Computing"
  Presentations with similar titles and content given in the following venues:
  - 2009 Hadoop Summit, Santa Clara, California (6/10/2009)
  - Genome Informatics Alliance Meeting, Healdsburg, California (3/16/2009)
- "Scalable Text Processing with MapReduce"
  Presentations with similar titles and content given in the following venues:

  - Virginia Tech, Northern Virginia Center (10/16/2009)
  - 2009 CLuE PI Meeting, Mountain View, California (10/5/2009)
  - Institute for Defense Analyses, Center for Computing Sciences, Bowie, Maryland (6/21/2009)
  - Johns Hopkins University (6/17/2009)
  - Space Telescope Science Institute, Baltimore, Maryland (11/20/2008)
  - IBM Silicon Valley Laboratory (7/22/2008)
  - IBM Almaden Research Center (7/21/2008)
  - The 9th IEEE/NATEA Conference on New Frontiers in Computing (7/19/2008)
  - IBM T. J. Watson Research Center (7/14/2008)
  - Johns Hopkins University, Human Language Technology Center of Excellence (6/24/2008)
  - Laboratory for Telecommunications Sciences, College Park, Maryland (6/12/2008)
  - The 25th Annual Human-Computer Interaction Lab Symposium (5/29/2008)
  - Carnegie Mellon University, Language Technologies Institute (5/8/2008)
  - National Science Foundation (4/29/2008)
  - Yahoo! Hadoop Summit (3/25/2008)
  - Georgia Institute of Technology, College of Computing (2/21/2008)

- "A Hammer and a Screwdriver for Text Analysis"
  Emory University, School of Law (2/22/2008)
- "A Framework for Conceptual Retrieval and an Application in Clinical Question Answering"
  Presentations with similar titles and content given in the following venues:
  - Emory University, Department of Math and Computer Science (2/21/2008)
  - UCLA, Institute for Pure and Applied Mathematics (10/4/2007)
  - IBM T.J. Watson Research Center (8/15/2007)
  - Microsoft Research (3/30/2007)
  - University of Washington, The iSchool (3/29/2007)
  - Rutgers University, School of Communication, Information and Library Studies (2/22/2007)
  - Johns Hopkins University, Division of Health Sciences Informatics (5/19/2006)
- "The Digital Docket: Interfaces to Support the Scholarly Exploration of Text Collections"
  The 24th Annual Human-Computer Interaction Lab Symposium (5/31/2007)
- "Two Approaches to Web-Based Question Answering"
  America Online (6/9/2005)
- "The Event Structure of English and Mandarin"
  University of Maryland, Department of Linguistics (2/11/2004)
- "Intuitive Information Access: Answering Questions using the Web, Databases, and Video"
  University of Maryland, Computational Linguistics and Information Processing Lab (2/10/2004)
- "Event Structure in Mandarin"
  University of Pennsylvania, Department of Computer & Information Sciences (12/1/2003)
- "Towards Intelligent Information Access Systems"
  Information Sciences Institute, University of Southern California (11/1/2002)

6.3 Internal Talks

- "dstlr: Scalable Knowledge Graph Construction from Unstructured Text — Project Update"
  Waterloo–Huawei Joint Innovation Lab 2022 Workshop (6/1/2022)
- "dstlr: Scalable Knowledge Graph Construction from Unstructured Text — Project Update"
  Waterloo–Huawei Joint Innovation Lab 2021 Workshop (6/7/2021)

- "dstlr: Scalable Knowledge Graph Construction from Unstructured Text — Project Update"
  Waterloo–Huawei Joint Innovation Lab 2020 Workshop (6/10/2020)
- "dstlr: Scalable Knowledge Graph Construction from Unstructured Text"
  Waterloo–Huawei Joint Innovation Lab 2019 Workshop (5/30/2019)
- "Computational Journalism Research using Twitter in Maryland's CLIP Lab"
  University of Maryland, Center for the Advanced Study of Communities and Information (2/24/2015)
- "The Evolution of WTF: Follower Recommendation Services at Twitter"
  University of Maryland, Computational Linguistics and Information Processing Lab (12/12/2014)
- "Is Searching Full Text More Effective Than Searching Abstracts?"
  National Library of Medicine, National Center for Biotechnology Information (12/16/2008)
- "Scalable Text Processing with MapReduce"
  Presentations with similar titles and content given in the following venues:
  - National Library of Medicine, Lister Hill Center (6/13/2008)
  - National Library of Medicine, National Center for Biotechnology Information (10/30/2008)
  - University of Maryland, The iSchool (10/22/2008)
  - University of Maryland, Computational Linguistics and Information Processing Lab (10/17/2007)
- "Understanding PubMed Users for Enhanced Text Retrieval"
  National Library of Medicine, National Center for Biotechnology Information (3/11/2008)
- "The Digital Docket: Information Retrieval Meets Political Science"
  Maryland Institute for Technology in the Humanities (MiTH) Digital Dialogue Series (12/5/2006)
- "Supporting Knowledge Discovery within the MEDLINE Database: A Progress Report"
  National Library of Medicine, National Center for Biotechnology Information (8/23/2006)
- "Moving Towards Question Answering for Evidence-Based Medicine"
  Presentations with similar titles and content given in the following venues:
  - University of Maryland, The iSchool (9/22/2005)
  - National Library of Medicine, Lister Hill Center (7/25/2005)
- "Question Answering in the Broader Context of Information Access"
  University of Maryland, Computational Linguistics and Information Processing Lab (9/15/2004)

6.4 Panels

- Panelist, Careers in NLP Panel, a panel in the Industry Track of NAACL HLT 2021, June 2021, Online. (Invited)
- Panel Moderator, GPU Panel for High-Throughput Computing, UMD/NVIDIA GPU Summit, October 2014, College Park, Maryland. (Invited)
- Panelist, Workshop on NLP and Social Dynamics at ACL 2014, June 2014, Baltimore, Maryland. (Invited)
- Panelist, "How much data is enough?" at the Third Annual Robert H. Smith School of Business and IBM Business Analytics Workshop: Leveraging the Power of Analytics over the Internet, April 2013, College Park, Maryland. (Invited)
- Panelist, "Is Social Media Making News Generation and Consumption Better?" at the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), July 2011, Barcelona, Spain. (Invited)
- Organizer, "Investing in the Cloud: A Breakfast Conversation", an event bringing together roughly two dozen local venture capitalists to discuss the implications of cloud computing on entrepreneurship, May 13, 2009.

- Chair, "Is There a Cloud in Your Future? Applications of 'Cloud Computing' to Web-scale Problems", panel at iConference 2008, February 2008, Los Angeles, California. (Refereed)
- Panelist, Workshop on Pragmatics of Question Answering at HLT/NAACL 2004, May 2004, Boston, Massachusetts. (Invited)
- Panelist, The 2003 AAAI Spring Symposium on New Directions in Question Answering, March 2003, Palo Alto, California. (Invited)

6.5 Media Mentions

- February 2023. Interviewed for Waterloo Region Record article, "Is AI a threat to academic integrity?".
- November 2021, University of Waterloo news release, "New AI brings the power of natural language processing to African languages". Additional mentions in the following outlets:
  - Communications of the ACM News
  - ACM Blog
  - The Institution of Engineering and Technology
  - Marktechpost
  - AIhub
  - Technology.org
- September 2021, University of Waterloo news release, "Computer scientists create new search systems to limit COVID-19 misinformation". Additional mentions in the following outlets:
  - The Economic Times
  - CityNews Kitchner
  - Communications of the ACM News
  - The Hack Posts
- June 2021. Interviewed for article "Province mulls ban on 'significant' AI decisions without Ontarians' consent" by Queen's Park Today.
- April 2021. Research featured in the Waterloo Region Record article, "Waterloo researchers use technology to accurately predict flooding".
- May 2020. Research featured in the Waterloo Region Record article, "Staying on top of the research: University of Waterloo researchers create search engine to sort through all the new information on COVID-19".
- September 2018. "Getting your TV to better understand you", press for (C124), production deployment of neural network models to improve voice query understanding on Comcast's "voice remote":
  - Communications of the ACM News
  - IEEE GlobalSpec Engineering 360
  - Phys.org
  - (e)Science News
- August 2015. Research on web archives mentioned in Canadian Broadcasting Corporation (CBC) news article "Waterloo professor restores deleted political platforms, promises".
- June 2014. Interviewed for article "Now You See It", feature in Odyssey Magazine, published by the College of Computer, Mathematical, and Natural Sciences at the University of Maryland.
- September 2, 2011. Interviewed for an article about cloud computing by the Baltimore Business Journal.
- June 21, 2011. Interviewed by MSN Money for article "How to invest in the cloud".

- May 25, 2011. Interviewed by the Chronicle of High Education for article "This Is What a Sabbatical at Twitter Looks Like".
- February 18, 2011. Interviewed for Technology Review article "HP Sees Cloudy Skies Ahead".
- October 12, 2009. Interviewed for the New York Times article "Training to Climb an Everest of Digital Data"; article appeared in front page of the Business Section, with picture.
- April 1, 2009. Interviewed for the Washington Post article on cloud computing "Tech Firms Seek to Get Agencies on Board With Cloud Computing".
- October, 2007. Google/IBM Academic Cloud Computing Initiative (ACCI). The project aims to help faculty and students address the challenges of "web-scale" computing by making computer clusters available for research and education. Six universities were involved at the outset: Berkeley, Carnegie Mellon, MIT, Stanford, University of Maryland, and University of Washington. I am the lead faculty at Maryland for this project, which was mentioned in the following media outlets († denotes cases where I was interviewed):
  - Press Release, Google (October 8, 2007)
  - Press Release, IBM (October 8, 2007)
  - Press Release, University of Maryland (October 8, 2007)
  - Wall Street Journal (October 8, 2007)
  - New York Times (October 8, 2007)
  - BBC News (October 9, 2007)
  - †Washington Post (October 9, 2007)
  - †Design News (October 12, 2007)
  - †CTV, local news station for Prince George's County (October 12, 2007)
  - †Information Week (October 13, 2007)
  - †Technology Review (December 3, 2007)
  - BusinessWeek (December 13, 2007)
  - †Baltimore Business Journal (January 25, 2008)

7 Funding: University of Waterloo

(N18) Huawei Unrestricted Gift

Role: PI
Sponsor: Huawei
Duration: 5/2022
Award amount: $160,000 CAD

(N17) Conversational Modeling and Dialogue Generation for Consumer Understanding with Neural Techniques

Role: PI
Sponsor: MITACS (Elevate Program)
Duration: 8/2022—8/2024
Award amount: $160,000 CAD

(N16) Multimodal Representational Learning

Role: PI
Sponsor: Naver Labs Europe
Duration: 3/2022—3/2023
Award amount: $53,000 CAD

(N15) Google Unrestricted Gift

Role: PI
Sponsor: Google
Duration: 8/2021
Award amount: $30,000 USD

(N14) Pretrained Transformers for Effective and Efficient Information Access: BERT and Beyond, NSERC Discovery Grant (RGPIN-2021-02490)

Role: applicant
Sponsor: Natural Sciences and Engineering Research Council of Canada
Duration: 4/2021—3/2026
Award amount: $320,000 CAD

(N13) Global Water Futures

Role: core lead, computer science and data management
Sponsor: Canada First Research Excellence Fund
Duration: 1/2018—12/2020 (Phase 1), 1/2021—12/2023 (Phase 2)
Allocated share: $370,592 CAD (Phase 1), $444,320 (Phase 2)

(N12) The Archives Unleashed Project – Phase II

Role: co-PI (PI: Ian Milligan, co-PI: Nick Ruest, Jefferson Bailey)
Sponsor: Andrew W. Mellon Foundation
Duration: 6/2020—5/2023
Award amount: $800,000 USD

(N11) MANNERSS: Multilingual Adaptation for NER in the Social Sciences research project, Grant

Role: co-PI (PI: Ian Milligan)
Sponsor: Microsoft
Duration: 6/2020—12/2021
Award amount: $25,000 CAD

(N10) Hybrid Neural-Physical Models for Streamflow Prediction, Grant

Role: PI
Sponsor: Microsoft
Duration: 6/2020—12/2021
Award amount: $25,000 CAD

(N9) Gift to support the Archives Unleashed Project

Role: co-recipient (with Ian Milligan)
Sponsor: Start Smart Labs
Duration: 8/2018—1/2020
Award amount: $90,000 USD

(N8) RDF Data Unification: Breaking the Data Silos in DSTLR

Role: PI
Sponsor: Waterloo–Huawei Joint Innovation Lab
Duration: 11/2018—10/2021, no-cost extension to 10/2022
Award amount: $303,112 CAD

(N7) Postdoctoral Fellowship in Software Curation for the Sciences and Social Sciences

Role: co-supervisor (with Alison Hitchens)
Sponsor: Council on Library and Information Resources
Duration: 7/2018—6/2020
Award amount: $56,000 USD

(N6) Open-Domain Contextual Conversation Generation, Mitacs Accelerate Internship

Role: supervisor
Sponsor: Mitacs
Duration: 5/2018—11/2018
Award amount: $17,000 CAD

(N5) The Archives Unleashed Project

Role: co-PI (PI: Ian Milligan, co-PI: Nick Ruest)
Sponsor: Andrew W. Mellon Foundation
Duration: 6/2017—5/2020
Award amount: $462,000 USD

(N4) Modeling Time, Space, and Networks for Effective and Efficient Information Retrieval, NSERC Discovery Accelerator Supplement

Role: applicant
Sponsor: Natural Sciences and Engineering Research Council of Canada
Duration: 4/2016—3/2019
Award amount: $120,000 CAD

(N3) Modeling Time, Space, and Networks for Effective and Efficient Information Retrieval, NSERC Discovery Grant (RGPIN-2016-04138)

Role: applicant
Sponsor: Natural Sciences and Engineering Research Council of Canada
Duration: 4/2016—3/2021
Award amount: $250,000 CAD

(N2) AWS Cloud Credits for Research

Role: recipient
Sponsor: Amazon
Award date: 3/2016
Award amount: $5,000 USD in Amazon Web Services credits for supporting research

(N1) Unlock your Web Archives Hackathon, SSHRC Connection Grant

Role: co-applicant (applicant: Ian Milligan, co-applicants: Matthew Weber, Nathalie Casemajor Loustau, Nicholas Worby)
Sponsor: Social Sciences and Humanities Research Council of Canada
Duration: 10/2015—9/2016
Award amount: $23,715 CAD

8 Funding: University of Maryland

(M33) III: Small: Safely Searching Among Sensitive Content (IIS-1618695)

Role: co-PI (PI: Douglas W. Oard, co-PI: Katie Shilton)
Sponsor: National Science Foundation
Duration: 9/2016—8/2019
Award amount: $514,716 USD

(M32) LTS DO 0020: Prototyping Search and Analytics for the Web in a Box (Contract #H98230-13-D-0056)

Role: PI
Sponsor: Laboratory for Telecommunication Sciences
Duration: 3/2015—3/2016
Award amount: $123,832 USD

(M31) II-EN: Hadoop NextGen Infrastructure for Heterogeneous Approaches to Data-Intensive Computing (CNS-1405688)

Role: PI
Sponsor: National Science Foundation
Duration: 8/2014—7/2017
Award amount: $499,852 USD

(M30) Student Travel Support for the 2014 IEEE International Conference on Big Data (IIS-1444666)

Role: PI
Sponsor: National Science Foundation
Duration: 6/2014—5/2015
Award amount: $20,000 USD

(M29) Warcbase: Building a Scalable Web Archiving Platform on HBase and Hadoop

Role: PI
Sponsor: Andrew W. Mellon Foundation (via Columbia University)
Duration: 04/2014–11/2015
Award amount: $51,925 USD

(M28) AWS in Education Research Grant Award

Role: recipient
Sponsor: Amazon
Award date: 1/2014
Award amount: $15,000 USD in Amazon Web Services credits for supporting research

(M27) Y! STAR (Yahoo! Servers To Academic Researchers)

Role: recipient
Sponsor: Yahoo!
Award date: 10/2013
Award amount: donation of 50 servers, valued at $17,000 USD

(M26) Curate Cloud: Building Digital Curation Excellence through Professional Education, Cloud Computing and Community Outreach

Role: PI
Sponsor: Institute of Museum and Library Services
Duration: 6/2013—5/2016
Award amount: $299,730 USD

(M25) III: Small: Providing Relevant and Timely Results: Real-Time Search Architectures and Relevance Algorithms (IIS-1218043)

Role: PI
Sponsor: National Science Foundation
Duration: 9/2012—9/2015
Award amount: $499,960 USD

(M24) DELPHI: Distributed Empirical Language Processing for Human Interaction, DARPA BOLT (Broad Operational Language Translation)

Role: co-PI (PI: Douglas W. Oard, co-PIs: Hal Daumé, Philip Resnik)
Sponsor: DARPA, via IBM
Duration: 10/2011—3/2015
Award amount: $1,966,776 USD

(M23) EAGER: Learning to Efficiently Rank with Cascades (IIS-1144034)

Role: PI
Sponsor: National Science Foundation
Duration: 9/2011—8/2013 (no cost-extension until 8/2014)
Award amount: $150,000 USD

(M22) DC: Small: Cross-Language Bayesian Models for Web-Scale Text Analysis Using MapReduce (CCF-1018625)

Role: PI (co-PIs: Jordan Boyd-Graber, Philip Resnik)
Sponsor: National Science Foundation
Duration: 9/2010—8/2013 (no cost-extension until 8/2014)
Award amount: $449,953 USD

(M21) III: Small: Ivory—A Hadoop Toolkit for Distributed Text Retrieval (IIS-0916043)

Role: PI
Sponsor: National Science Foundation
Duration: 10/2009—9/2012 (no cost-extension until 9/2013)
Award amount: $499,662 USD

(M20) IC Postdoctoral Research Fellowship

Role: PI (co-PI: Douglas W. Oard)
Sponsor: Intelligence Community
Duration: 6/2009—5/2011
Award amount: $240,000 USD

(M19) IBM Faculty Award

Role: recipient
Sponsor: IBM
Award date: 10/2009
Award amount: $25,000 USD

(M18) 2009 PI Meeting for the Cluster Exploratory Program (IIS-0957954)

Role: PI
Sponsor: National Science Foundation
Duration: 10/2009—9/2010
Award amount: $49,991 USD

(M17) Putting the Clouds in Context: Statistical Machine Translation with MapReduce (IIS-0836560)

Role: PI (co-PI: Philip Resnik)
Sponsor: National Science Foundation
Duration: 8/2008—7/2010
Award amount: $200,000 USD
Additional resources: access to a 500-machine Hadoop cluster to support research

(M16) Google/IBM Academic Cloud Computing Initiative

Role: PI
Sponsor: Google and IBM
Duration: 8/2007—7/2010
Award amount: access to a 20-machine computer cluster from 8/2007 to 7/2008; from 8/2008, project merged with (M17)

(M15) III-COR: iOPENER—A Flexible Framework to Support Rapid Learning in Unfamiliar Research Domains (IIS-0705832)

Role: co-PI (PI: Bonnie Dorr, co-PIs: Judith L. Klavans, Ben Shneiderman)
Sponsor: National Science Foundation

Cloudera Exhibit 1006 - Page 69 of 86

Duration: 9/2007—9/2010
Award amount: $720,000 USD

(M14) DHB: A Computational Approach to Understanding the Dynamics of the Judicial System (BCS-0624067)

Role: co-PI (PI: Wayne McIntosh)
Sponsor: National Science Foundation
Duration: 9/2006—9/2009
Award amount: $749,996 USD

(M13) Rosetta: An Analyst Copilot, DARPA GALE (Global Autonomous Language Exploitation)

Role: co-PI (PI: Douglas W. Oard, co-PIs: Bonnie Dorr, Philip Resnik, Amy Weinberg)
Sponsor: DARPA (via sub-award from IBM)
Duration: 9/2005—8/2011
Award amount: $3,093,920 USD

(M12) Amazon Web Services for Cloud Computing Course

Role: recipient
Sponsor: Amazon
Award date: 9/2008
Award amount: $2,500 USD in Amazon Web Services credits for supporting teaching

(M11) Amazon Web Services for Cloud Computing Research

Role: recipient
Sponsor: Amazon
Award date: 6/2008
Award amount: $2,100 USD in Amazon Web Services credits for supporting research

(M10) National Center for Biotechnology Information (NCBI) Faculty Fellowship

Role: PI
Sponsor: National Library of Medicine, National Institutes of Health
Duration: 6/2008—5/2009
Award amount: $88,947 USD

(M9) Computational Linguistics for Metadata Building (CLiMB-2)

Role: senior personnel (PI: Judith L. Klavans)
Sponsor: Andrew W. Mellon Foundation
Duration: 10/2005—7/2008
Award amount: $841,000 USD

(M8) National Center for Biotechnology Information (NCBI) Faculty Fellowship

Role: PI
Sponsor: National Library of Medicine, National Institutes of Health
Duration: 6/2007—5/2008
Award amount: $61,950 USD

(M7) AOL Fellowship for G. Craig Murray

Role: PI
Sponsor: America Online, Inc.
Duration: 8/2006—11/2006
Award amount: $40,000 USD

(M6) Exploration of Sentence-Level Relevance in MEDLINE Abstracts

Role: PI
Sponsor: National Library of Medicine, National Institutes of Health

Duration: 6/2006—12/2006
Award amount: $17,000 USD

(M5) How to Answer Procedural Questions?

Role: PI
Sponsor: America Online, Inc.
Duration: 6/2006—9/2006
Award amount: $15,475 USD

(M4) AOL Fellowship for G. Craig Murray

Role: PI
Sponsor: America Online, Inc.
Duration: 8/2005—8/2006
Award amount: $40,000 USD

(M3) Joint Institute for Knowledge Discovery

Role: senior personnel (PI: V.S. Subramanian)
Sponsor: Department of Defense
Duration: 3/2005—8/2007
Award amount: $1,840,000 USD

(M2) Research in Clinical Question Answering

Role: PI
Sponsor: National Library of Medicine, National Institutes of Health
Duration: 6/2005—9/2005
Award amount: $24,192 USD

(M1) Evaluating Advanced Question Answering Systems, AQUAINT (Advanced Question Answering for Intelligence)

Role: PI
Sponsor: ARDA (via sub-award from MIT)
Duration: 11/2004—7/2005
Award amount: $50,000 USD

9 Teaching: University of Waterloo

9.1 Courses Taught

- CS 848 The Art and Science of Empirical Computer Science: Fall 2022 (Enrollment: 22)
- CS 451/651 Data-Intensive Distributed Computing: Fall 2018 (Enrollment: 78)
- CS 848 Advanced Topics in Databases: Spring 2018 (Enrollment: 4)
- CS 451/651/431/631 Data-Intensive Distributed Computing: Winter 2018 (Enrollment: 91)
- CS 489/698 Big Data Infrastructure: Winter 2017 (Enrollment: 86)
- CS 499R Readings in Computer Science: Fall 2016 (Enrollment: 4)
- CS 898 Advanced Special Topics in Computer Science: Fall 2016 (Enrollment: 4)
- CS 489/698 Big Data Infrastructure: Winter 2016 (Enrollment: 68)

10 Teaching: University of Maryland

10.1 Courses Taught

- INST 767: Big Data Infrastructure
  - Spring 2015 (Enrollment: 18)
- INFM 603: Information Technology and Organizational Context
  - Fall 2014 (Enrollment: 28)
  - Spring 2014 (Enrollment: 17)
  - Fall 2013 (Enrollment: 30)
  - Fall 2012 (Enrollment: 20)
- INFM 718G/CMSC 838G: Web-Scale Information Processing Applications
  - Spring 2013 (Enrollment: 33)
  - Spring 2010 (Enrollment: 23)
- LBSC 888: Doctoral Seminar
  - Fall 2009 (Enrollment: 13)
- CMSC 723: Computational Linguistics I
  - Fall 2009 (Enrollment: 22)
- LBSC 878A: Web-Scale Information Processing Applications
  - Fall 2008 (Enrollment: 20)
  - Spring 2008 (Enrollment: 9)
- LBSC 690: Information Technology
  - Fall 2008 (Enrollment: 30)
  - Fall 2007 (Enrollment: 28)
  - Spring 2007 (Enrollment: 25)
  - Fall 2006 (Enrollment: 25/18; two sections)
  - Fall 2005 (Enrollment: 25/17; two sections)
  - Fall 2004 (Enrollment: 20/24; two sections)
- LBSC 796/INFM 718R: Information Retrieval Systems
  - Spring 2006 (Enrollment: 9)
  - Spring 2005 (Enrollment: 15)
- INFM 700: Information Architecture
  - Spring 2008 (Enrollment: 22)
- LBSC 709: Independent Study
  - Summer 2013 (Sarah Weissman)
  - Fall 2012 (Mohamed Mohideen)
  - Fall 2011 (Dongming Zhang)
  - Fall 2010 (Michael McGrath)

Cloudera Exhibit 1006 - Page 72 of 86

  - Spring 2007 (Greg Jablonski)
  - Spring 2005 (Xiaoli Huang)

- CMSC 390: Honors Paper

  - Spring 2008 (Enrollment: 2)

10.2 Course or Curriculum Development

- Development of courses that had not existed previously:

  - INST 767: Big Data Infrastructure (Spring 2015)
  - LBSC 878A: Web-Scale Information Processing Applications (Spring 2008)
  - INFM 700: Information Architecture (Spring 2008)

- Teaching for the first time a course that had previously existed, with significant revision of content:

  - INFM 603: Information Technology and Organizational Context (Fall 2012)
  - LBSC 888: Doctoral Seminar (Fall 2009)
  - CMSC 723: Computational Linguistics I (Fall 2009)
  - LBSC 796/INFM 718R: Information Retrieval Systems (Spring 2005)
  - LBSC 690: Information Technology (Fall 2004)

- Other significant course revisions:

  - INFM 718G/CMSC 838G: Web-Scale Information Processing Applications (Spring 2010)
  - LBSC 878A: Web-Scale Information Processing Applications (Fall 2008)
  - LBSC 690: Information Technology (Fall 2008)

11 Advising: University of Waterloo

11.1 Undergraduate

11.1.1 Advisor, Independent Research Projects

- Winter 2022: Arthur Chen, Haonan Chen, Lingwei Gu, Yilin (Larry) Li, Manveer Singh Tamber, Mikhail Tsirlin, Yuetong Wang, Matthew Y. R. Yang
- Fall 2021: Mayank Anand, Yilin (Larry) Li, Nima Sadri, Yuetong Wang
- Spring 2021: Mayank Anand, Arthur Chen, Shane Ding, Yilin (Larry) Li, Estella Liu, Sailesh Nankani, Alvis Wong, Jiarui Zhang
- Winter 2021: Yuxuan Ji, Kai Sun
- Fall 2020: Hang Cui, Shane Ding, Qing Guo, Chengcheng (Stephanie) Hu, Yuqi Liu, Yicheng (Kevin) Xu, Ruizhou (Richard) Xu, Edwin Zhang, Lijia (Lizzy) Zhang, Yuxin (Vicky) Zhu
- Spring 2020: Shane Ding, Xiao (Johnson) Han, Chengcheng (Stephanie) Hu, Alvis Wong, Edwin Zhang
- Winter 2020: Tobi Adewoye, James Bai, Nikhil Gupta, Xiao (Johnson) Han, Edwin Zhang
- Fall 2019: Kelvin Jiang, Weihua Li, Apoorv Pandey, Ronak Pradeep
- Spring 2019: Hedi (Edward) Lu, Sheng Jin (Emily) Wang, Ze Zhong (Charles) Wu
- Winter 2019: Haonan Li, Boying (Ashley) Liu, Mengfei Liu, Xinyu Liu, Ziwei Zhao
- Fall 2018: Xiaoran (Gloria) Huang, Lidiya Murakhovs'ka, Heli Wang
- Spring 2018: Joejun Lee, Keer Liu, Mengfei Liu, Xinyu Liu, Danlian Wang, Whuan (Joseph) Zhou

Cloudera Exhibit 1006 - Page 73 of 86

- Winter 2018: Yiyun (Isa) Liang, Chuyang (Charles) Qi, Heli Wang, Andrew Wesson, Arlene (Yu) Zhang, Yinan Zhang
- Fall 2017: Liyan Guan, Billy Zhengxu Jin, Mengping Li, Mengfei Liu, Mingyue (Emily) Ma, Raphael Tang, Weijie (David) Wang, Alice Zhou
- Spring 2017: Arthur Guo, Laetitia Huang, Junchen Zhang
- Winter 2017: Shirley Du, Borui Lin, Michael Rose, Zhucheng (Michael) Tu, Patrick White, Alice Zhou, Shine Wang, Zack Wang
- Fall 2016: Youngbin Kim, Jonathan Ma, Ronak Patel, Pranjal Shah
- Spring 2016: Youngbin Kim, Xin Qian, Michael Rose, Zhucheng (Michael) Tu, Patrick White, Alice Zhou
- Winter 2016: Dayang Shi, Xin Qian
- Fall 2015: Alice Zhou, Xin Qian

11.1.2 Advisor, Software Engineering Final-Year Design Projects

- 2016: Zhucheng (Michael) Tu, Michael Rose, and Patrick White

11.2 Master's

11.2.1 Advisor/Co-Advisor

- Mofetoluwa Adeyemi
- Akintunde Oladipo
- Aileen Lin
- Odunayo Ogundepo
- Kelechi Ogueji

  Completed August 2022
  Thesis: AfriBERTa: Towards Viable Multilingual Language Models for Low-resource Languages
- Yuqi Liu

  Completed August 2022
  Thesis: Simple Yet Effective Pseudo Relevance Feedback with Rocchio's Technique and Text Classification
- Chengcheng (Stephanie) Hu

  Completed April 2022
  Thesis: Learning to Rank in the Age of Muppets
- Anup Anand Deshmukh (with Ming Li)

  Completed September 2021
  Thesis: Unsupervised Syntactic Structure Induction in Natural Language Processing
- Xinyu (Mavis) Liu

  Completed April 2021
  Thesis: Total Relation Recall: High-Recall Relation Extraction
- Ashutosh Adhikari (with Pascal Poupart)

  Completed December 2020
  Thesis: Generalization on Text-based Games using Structured Belief Representations
- Hsiu-Wei Yang

  Completed December 2020
  Thesis: Cross-Lingual Entity Matching for Knowledge Graphs

- Linqing Liu

  Completed September 2020
  Thesis: Towards Effective Utilization of Pretrained Language Models — Knowledge Distillation from BERT

- Achyudh Ram (with Mei Nagappan)

  Completed July 2020
  Thesis: Exploiting Token and Path-based Representations of Code for Identifying Security-Relevant Commits

- Ryan Clancy

  Completed April 2020
  Thesis: dstlr: Scalable Knowledge Graph Construction from Text Collections

- Martin Gauch

  Completed April 2020
  Thesis: Machine Learning for Streamflow Prediction

- Jaejun (Brandon) Lee

  Completed December 2019
  Thesis: Honkling: In-Browser Personalization for Ubiquitous Keyword Spotting

- Zeynep Akkalyoncu Yilmaz

  Completed December 2019
  Thesis: Cross-Domain Sentence Modeling for Relevance Transfer with BERT

- Hamidreza Shahidi (w/ Ming Li)

  Completed August 2019
  Thesis: Neural Text Generation from Structured and Unstructured Data

- Wei (Victor) Yang

  Completed April 2019
  Thesis: End-to-End Neural Information Retrieval

- Youngbin Kim

  Completed August 2018
  Thesis: Serverless Data Analytics with Flint

- Royal Sequeira

  Completed August 2018
  Thesis: Simple Convolutional Neural Networks with Linguistically-Annotated Input for Answer Selection in Question Answering

- Zhucheng (Michael) Tu

  Completed May 2018
  Thesis: An Experimental Analysis of Multi-Perspective Convolutional Neural Networks

- Salman Mohammed

  Completed December 2017
  Thesis: The Limited Effectiveness of Neural Networks for Simple Question Answering on Knowledge Graphs

11.2.2 Member, Thesis Committees

- Christopher Bobotsis: Winter 2023
- Runcheng Liu: Fall 2022
- Chang Liu: Fall 2022

- Venkatraman Arumugam: Spring 2022
- Zaid Hassan Chaudhry: Spring 2022
- Ethan Ward: Fall 2021
- Udhav Sethi: Spring 2021
- Mohammadali Niknamian: Winter 2021
- Omar Attia: Fall 2020
- Xin Lian: Fall 2019
- Michael Farag: Spring 2019
- Junnan Chen: Spring 2018
- Deepak Rishi: Fall 2017

11.3 Ph.D.

11.3.1 Advisor/Co-Advisor

All students in Computer Science unless otherwise noted.

- Qiutong (Glen) Yu (with Bryan Tolson) [Civil and Environmental Engineering]
- Nandan Thakur
- Xueguang Ma
- Yuqing (Sherry) Xie (with Ming Li)
- Xinyu (Crystina) Zhang
- Ronak Pradeep
- Wei Zhong
- Minghan Li
- Jheng-Hong (Matt) Yang
- Sheng-Chieh (Jack) Lin
- Zhiying Jiang
- Peng Shi

  Completed January 2023
  Thesis: Facilitating Information Access for Heterogeneous Data Across Many Languages
- Ji Xin (with Yaoliang Yu)

  Completed January 2023
  Thesis: Efficient Inference of Transformers in Natural Language Processing: Early Exiting and Beyond
- Kareem El Gebaly (with Ashraf Aboulnaga and Lukasz Golab)

  Completed May 2018
  Thesis: Analytics for Everyone

11.3.2 Member, Thesis Committees

- Khaled Ammar: completed February 2023
- He Bai: completed January 2023
- Hung Viet Pham: completed August 2022
- Georgios Michalopoulos: completed July 2022
- Ilia Sucholutsky: completed June 2021

Cloudera Exhibit 1006 - Page 76 of 86

- Mina Farid: completed May 2020
- Christian Gorenflo: completed May 2020
- Adam Roegiest: completed March 2017
- Luchen Tan: completed January 2017
- Gaurav Baruah: completed August 2016
- Gregorz Drzadzewski: completed December 2015

11.3.3 Member, Comprehensive-II (Depth) Examination Committees

- Partha Chakraborty: Fall 2022
- He Bai: Winter 2022
- Georgios Michalopoulos: Fall 2021
- Ankit Vadehra: Spring 2021
- Hung Pham: Fall 2020
- Atif Khan: Fall 2019
- Christian Gorenflo: Spring 2018
- Adam Roegiest: Winter 2016
- Khaled Ammar: Fall 2016
- Gaurav Baruah: Fall 2015
- Luchen Tan: Fall 2015

11.3.4 Member, External Committees

- Behrooz Mansouri (Rochester Institute of Technology, Rochester, New York): completed July 2022
- Christian Hansen (University of Copenhagen, Denmark): completed August 2021
- Casper Hansen (University of Copenhagen, Denmark): completed August 2021
- Sean MacAvaney (Georgetown University, Washington, D.C.): completed March 2021
- Abdalghani Abujabal (Max-Planck Institute for Informatics, Saarbrücken, Germany): completed April 2019
- Arman Cohan (Georgetown University, Washington, D.C.): completed March 2018
- Salvatore Trani (Computer Science, University of Pisa, Italy): completed January 2017

11.4 Post-Doctoral

11.4.1 Advisor/Co-Advisor

- Rodrigo Nogueira: January 2020—March 2022
- Qian Zhang: August 2018—July 2020 (co-advised with Alison Hitchens)
- Ryan Deschamps: May 2017—April 2019 (co-advised with Ian Milligan)
- Luchen Tan: September 2017—June 2018
- Gaurav Baruah: October 2016—July 2017
- Matt Crane: March 2016—February 2018

12 Advising: University of Maryland

12.1 Undergraduate

12.1.1 Advisor, Independent Research Projects

- Ryan Wolniak (Computer Science): Fall 2014
- Aaron Cordova (Computer Science): Spring 2008
- Alan Jackoway (Computer Science): Spring 2008
- Alex Mont (Computer Science): Spring 2008

12.2 Master's

12.2.1 Advisor

- Tim Hawes (Linguistics)

  Completed Winter 2009
  Thesis: Computational Analysis of the Conversational Dynamics of the United States Supreme Court

12.2.2 Member, Thesis Committees

- Rashmi Sankepally (The iSchool): Fall 2014
- Zahra Ashktorab (The iSchool): Spring 2013

12.2.3 Advisor, Independent Research Projects

- Zhengzheng Xu (Computer Science): Fall 2013—Spring 2014
- Sarah Weissman (The iSchool): Summer 2013—Fall 2013
- Mohamed Mohideen (The iSchool): Fall 2012
- Michael McGrath (The iSchool): Summer 2010—Fall 2010
- Mike Lieberman (Computer Science): Spring 2009
- Anand Bahety (Computer Science): Fall 2008
- Shravya Konda (Computer Science): Fall 2008
- Samantha Mahindrakar (The iSchool): Fall 2008
- Greg Jablonski (The iSchool): Spring 2007, Spring 2008
- Lynne Plettenberg (The iSchool): Summer 2005

12.3 Ph.D.

12.3.1 Advisor/Co-Advisor

- Jinfeng Rao (Computer Science)

  Completed June 2018
  Dissertation: Temporal Context Modeling for Text Streams
- Hua He (Computer Science)

  Completed February 2018
  Dissertation: Architecture, Models, and Algorithms for Textual Similarity

Cloudera Exhibit 1006 - Page 78 of 86

- Yulu Wang (Computer Science)

  Completed January 2018
  Dissertation: Selective Search Architectures and Brute Force Scan Techniques for Summarizing Social Media Posts

  First permanent position: Software Engineer, Google (Mountain View, California)

- Ashwin Kayyorr (Computer Science), with Amol Deshpande

  Completed Winter 2014
  Dissertation: Minimizing Resource Consumption through Workload Consolidation in Large-Scale Distributed Data Platforms

  First permanent position: Research Scientist, Symantec Research Labs (Culver City, California)

- Ferhan Ture (Computer Science)

  Completed Summer 2013
  Dissertation: Searching to Translate and Translating to Search: When Information Retrieval Meets Machine Translation

  First permanent position: Research Scientist, BBN (Columbia, Maryland)

- Nima Asadi (Computer Science)

  Completed Summer 2013
  Dissertation: Multi-Stage Search Architectures for Streaming Documents

  First permanent position: Software Engineer, Google (Mountain View, California)

- Lidan Wang (Computer Science)

  Completed Summer 2012
  Dissertation: Learning to Efficiently Rank

  First permanent position: Applied Researcher, Microsoft Bing (Seattle, Washington)

- David Zajic (Computer Science), with Bonnie Dorr

  Completed Spring 2007
  Dissertation: Multiple Alternative Sentence Compression as a Tool for Automatic Summarization Tasks

  First permanent position: Assistant Research Scientist, Center for the Advanced Study of Language, University of Maryland (College Park, Maryland)

- Dina Demner-Fushman (Computer Science), with Douglas Oard

  Completed Winter 2006
  Dissertation: Complex Question Answering Based on Semantic Domain Model of Clinical Medicine

  First permanent position: Staff Scientist, National Library of Medicine (Bethesda, Maryland)

12.3.2 Member, Dissertation Committees

- Cody Buntain (Computer Science): completed Summer 2016
- Tan Xu (The iSchool): completed Fall 2015
- Udayan Khurana (Computer Science): completed Spring 2015
- Ejaz Ahmed (Computer Science): completed Spring 2015
- Sameh Khamis (Computer Science): completed Spring 2015
- Ke Zhai (Computer Science): completed Winter 2014
- Christopher LaTerza (Linguistics): completed Summer 2014
- Vladimir Eidelman (Computer Science): completed Winter 2013
- Zheng Wei (Electrical and Computer Engineering): completed Summer 2012
- Benjamin Smith (Psychology): completed Fall 2011

- Denis Filimonov (Computer Science): completed Fall 2011
- Hua Wei (Geography): completed Summer 2011
- Sang Chul Song (Electrical and Computer Engineering): completed Summer 2010
- Michael Schatz (Computer Science): completed Spring 2010
- Matt Snover (Computer Science): completed Spring 2010
- Tamer Elsayed (Computer Science): completed Summer 2009
- Woei-jyh Adam Lee (Computer Science): completed Fall 2008
- Yejun Wu (The iSchool): completed Spring 2008

12.3.3 External Committees

- Ke Tao (Computer Science, Delft University of Technology, the Netherlands): completed December 2014

12.4 Post-Doctoral

12.4.1 Advisor/Co-Advisor

- Jiaul Paik: Spring 2015—Spring 2016
- Earl Wagner: Fall 2010—Fall 2012 (co-advised with Douglas Oard)
- Tamer Elsayed: Fall 2009—Fall 2010

13 Service: Professional

13.1 Editorial Boards

- Action Editor, Transactions of the Association for Computational Linguistics: 2019—present
- Member, Standing Review Committee, Transactions of the Association for Computational Linguistics: 2014—2019
- Member of Editorial Board, Foundations and Trends in Information Retrieval: 2014—2022
- Associate Editor, IEEE Transactions on Big Data: 2015—2016
- Member of Editorial Board, Information Retrieval Journal: 2008—2010
- Guest Editor, Special Issue of the Journal of Parallel and Distributed Computing on Cloud Computing (with Gregory Chockler, Eliezer Dekel, and Joseph JaJa): 2009

13.2 Conference Chair Positions

13.2.1 Program Committees

- Program Co-chair, BigData Innovators Gathering (BIG) 2015
- Program Co-chair, IEEE BigData 2014: The 2014 IEEE International Conference on Big Data
- Program Co-chair, CIKM 2009: The 18th ACM Conference on Information and Knowledge Management

13.2.2 Other

- Data Chair, ICWSM 2015: The 9th International Conference on Weblogs and Social Media
- Practice and Experience Talks Chair, WSDM 2014: The Seventh International Conference on Web Search and Data Mining
- Tutorial Co-chair, NAACL/HLT 2013: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
- Poster Co-chair, SIGIR 2009: The 32th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval
- Workshop Co-chair, ACL/IJCNLP 2009: The 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on NLP
- Area Chair for Question Answering and Applications, EMNLP 2008: The 2008 Conference on Empirical Methods in Natural Language Processing
- Demonstrations Chair, ACL 2008: HLT: The 46th Annual Meeting of the Association for Computational Linguistics

13.3 Reviewing: Journals

Past editorial reviewing for the following journals:

- ACM Transactions in Asian Language Information Processing
- ACM Transactions on Information Systems
- ACM Transactions on Intelligent Systems and Technology
- ACM Transactions on Internet Technologies
- ACM Transactions on the Web
- The Annals of the American Academy of Political and Social Science
- BMC Bioinformatics
- Computational Linguistics
- Concurrency and Computation
- IBM Systems Journal
- Foundations and Trends in Information Retrieval
- IEEE Internet Computing
- IEEE Transactions on Parallel and Distributed Systems
- IEEE Transactions on Knowledge and Data Engineering
- Information Processing and Management
- Information Retrieval
- Information Systems
- Journal of Artificial Intelligence Research
- Journal of the American Medical Informatics Association
- Journal of Biomedical Informatics
- Journal of Natural Language Engineering
- Nucleic Acids Research
- TAL (Traitement automatique du langage naturel)
- Transactions of the Association for Computational Linguistics

Cloudera Exhibit 1006 - Page 81 of 86

13.4 Reviewing: Conferences

Includes roles such as program committee member (reviewer), senior program committee member (meta-reviewer), area chair (meta-reviewer), senior area chair (manages area chairs; for some large conferences), and related duties.

- 2023: AAAI 2023 (Senior Meta-Reviewer), TheWebConf (Senior PC), SIGIR (Senior PC)
- 2022: WSDM (Best Paper Awards Committee), TheWebConf (Senior PC), ECIR (PC), ACL (Senior Area Chair), SIGIR (PC), CLEF (PC), COLING (Senior Area Chair)
- 2021: WSDM (Senior PC), TheWebConf (Senior PC), ECIR (PC), AAAI (Senior PC), NAACL (Area Chair), ACL (Senior Area Chair), CLEF (PC), EMNLP (Area Chair)
- 2020: WSDM (Senior PC), ICWSM (Associate Editor), TheWebConf (Senior PC), JCDL (PC), ACL (Area Chair), KDD (Senior PC), EMNLP (Area Chair)
- 2019: WSDM (Senior PC), NAACL (PC), ICWSM (Associate Editor), SIGIR (Senior PC), KDD (Senior PC), EMNLP (Senior Area Chair), CIKM (Senior PC)
- 2018: VLDB (Associate Editor), WSDM (PC), ICWSM (Senior PC), NAACL (PC), ACL (PC), SIGIR (PC), KDD (Senior PC)
- 2017: WSDM (Senior PC), WWW (PC), ICWSM (Senior PC), SIGIR (PC)
- 2016: VLDB (PC), WSDM (Senior PC), ICWSM (Senior PC)
- 2015: WSDM (PC), ICWSM (Senior PC), SIGIR (Senior PC)
- 2014: ECIR (Senior PC), WWW (PC), ICWSM (Senior PC), SIGIR (PC), KDD (Senior PC), CIKM (Senior PC)
- 2013: NAACL (PC), SIGIR (PC), KDD (PC), ICWSM (Senior PC)

(Note: I was on an extended sabbatical from 2010–2012)

- 2010: SIGIR, NAACL
- 2009: SIGIR, NAACL, EMNLP, EACL
- 2008: SIGIR, ACL, EMNLP, PSB
- 2007: SIGIR, ACL, NAACL, InfoVis
- 2006: SIGIR, ACL, NAACL, EACL
- 2005: SIGIR, ACL, IJCAI
- 2004: SIGIR, COLING
- 2002: ACL

13.5 Reviewing: Workshops

- EMNLP 2020, Workshop on Scholarly Document Processing
- ACL 2020, Workshop on NLP for COVID-19
- WWW 2017, BigData Innovators Gathering
- WSDM 2017, Workshop on Mining Actionable Insights from Social Networks
- WWW 2016, BigData Innovators Gathering
- WWW 2014, BigData Innovators Gathering
- COLING 2008, Workshop on Information Retrieval for Question Answering (IR4QA)
- SIGIR 2007, Workshop on Web Information-Seeking and Interaction
- SIGIR 2004, Workshop on Information Retrieval for Question Answering (IR4QA)
- ACL 2004, The 4th Workshop on NLP and XML (NLPXML 2004)

- ACL 2004, Workshop on Question Answering in Restricted Domains
- EACL 2003, The 3rd Workshop on NLP and XML (NLPXML 2003)
- ACL 2001, Student Research Workshop

13.6 Reviewing: Funding Agencies

- Reviewer, Swiss National Science Foundation: 2021
- Reviewer, The Netherlands Organisation for Scientific Research (NWO): 2014
- Ad hoc reviewer, National Science Foundation (NSF), Directorate for Computer and Information Science and Engineering (CISE), Division of Information and Intelligent Systems (IIS): 2009, 2009, 2010, 2010, 2011
- Review panelist, National Science Foundation (NSF), Directorate for Computer and Information Science and Engineering (CISE), Division of Information and Intelligent Systems (IIS): 2007, 2008, 2008, 2009, 2011, 2013, 2014
- Ad hoc reviewer, National Science Foundation (NSF), Office of Cyberinfrastructure (OCI): 2013
- Reviewer, Faculty Research Committee, National University of Singapore: 9/2004

13.7 Workshop Organization

- SIGIR 2017, Lucene for Information Access and Retrieval Research (LIARR) Workshop
- SIGIR 2015, Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR)

13.8 TREC Organization

The Text Retrieval Conferences (TRECs) are community-wide evaluations organized by the National Institute of Standards and Technology (NIST). Each evaluation is divided into "tracks" that focus on a particular aspect of information retrieval. Track organizers formulate the problem to be tackled, define evaluation metrics, and coordinate the evaluation in conjunction with NIST.

- TREC 2023 Authoring Tools for Multimedia Content (AToMiC) Track
- TREC 2023 Deep Learning Track
- TREC 2022 Deep Learning Track
- TREC 2021 Deep Learning Track
- TREC 2018 Real-Time Summarization Track
- TREC 2017 Real-Time Summarization Track
- TREC 2016 Real-Time Summarization Track
- TREC 2015 Microblog Track
- TREC 2014 Microblog Track
- TREC 2013 Microblog Track
- TREC 2012 Microblog Track
- TREC 2011 Microblog Track
- TREC 2007 complex interactive Question Answering Task
- TREC 2006 complex interactive Question Answering Task

13.9 Other

- Member, Semantic Scholar API Advisory Board: 2023—present
- Member, Advisory Board, Institute of Information Science (IIS), Academia Sinica, Taiwan: 2020—present
- Member, Vector Institute Technical Advisor Board Meeting: 2018—2022
- SIGIR representative, CIKM Steering Committee: 2016—2017
- Member of Steering Committee, NewsREEL News Recommendation Evaluation Lab: 2013–2017
- Member, the ACM Special Interest Group Governing Board Replication Taskforce: 2016
- Invited expert discussant, Bureau of Labor Statistics Technical Advisory Committee (BLSTAC): 6/20/2014
- Jury member, Norvig Web Data Science Award (The Netherlands): 2014
- Jury member, Norvig Web Data Science Award (The Netherlands): 2013
- Organizer, 2009 PI Meeting for NSF's Cluster Exploratory (CLuE) program (which supports cloud computing research), in collaboration with Google and IBM: 2009
- Member, Advisory Board, Web task of the Cross Language Evaluation Forum (WebCLEF): 2008
- Member, MINDS Committee on Future Directions for Human Language Technology: 2006
- Member, Executive Board, Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL): 2004—2007

14 Service: University of Waterloo

14.1 School

- Organizer, Distinguished Lecturer Series: 2017—2020
- Member, Performance Evaluation Committee: 2018—2020
- Chair, Awards Committee: 2018—2019
- Member, School Advisory Committee on Appointments (SACA): 2017—2018
- Member, Planning committee for Masters in Data Science: 2016—2017
- Member, School Advisory Committee on Appointments (SACA): 2016—2017
- Member, School Advisory Committee on Appointments (SACA): 2015—2016

14.2 Faculty

- Member, MDSAI and MMath in Data Science Advisory Board: 2022—present
- Member, NSERC Strategic Plan Consultation Working Group: 2021

14.3 Campus

- Member, Waterloo Innovation Ecosystem Council: 2022—
- Lead for Technological Futures, Futures Roundtable: 2022 ("to support development of... a long-term vision for the university... to address where the university aspires to make global impacts by its 100th birthday, 35 years from now.")

15 Service: University of Maryland

15.1 Department

- Member, UMIACS Committee on Appointment, Promotions, and Tenure: Fall 2014—Spring 2015
- Organizer, CLIP Colloquium Series: Fall 2005—Spring 2008, Fall 2012—Spring 2015

15.2 College: Standing Committees

- Member, Undergraduate Program Committee: Fall 2014—Spring 2015
- Member, Nominating Committee: Spring 2015
- Member, Master of Information Management Program Committee: Fall 2013—Spring 2014
- Chair, Research Committee: Fall 2012—Spring 2013
- Member, Master of Human-Computer Interaction Program Committee: Fall 2012—Spring 2013
- Chair, Research Committee: Fall 2009—Spring 2010
- Assistant Chair, Doctoral Program Committee: Fall 2008—Spring 2009
- Member, Doctoral Program Committee: Fall 2007—Spring 2008
- Member, Master's Program Committee: Fall 2007—Spring 2008
- Member, Doctoral Committee: Fall 2006—Spring 2007
- Member, Doctoral Committee: Fall 2005—Spring 2006
- Member, Research and Teaching Support Committee: Fall 2004—Spring 2005

15.3 College: Ad Hoc Committees

- Member, Task Force, Review of how we teach research study design, data collection & analysis methods across all programs: Fall 2014
- Chair, Record Preparation Committee for the tenure case of Niklas Elmqvist: Spring 2013
- Member, Technology Coordinator Search Committee: Spring 2010
- Chair, Faculty Search Committee: Fall 2009—Spring 2010
- Member, Merit Pay Committee: Spring, 2009
- Member, Technology Advisory Group: Fall 2008
- Member, IT Specialist Search Committee: Spring 2006
- Member, Faculty Search Committee: Fall 2006—Spring 2007
- Member, Faculty Search Committee: Fall 2005

15.4 College: Other

- Faculty senator representing the iSchool: December 2006—Spring 2010

15.5 University

- Chair, Division of IT Infrastructure Working Group: Spring 2015
- Member, Advanced Computing Lecture Series Planning Committee: Fall 2014—Spring 2015
- Member, Search Committee for the Dean of the iSchool: Fall 2014—Spring 2015
- Member, Search Committee for Vice President and Chief Information Officer: Spring 2014
- Chair, Division of IT Research Enablement Working Group: Fall 2013—Fall 2014
- Member, Intellectual Property Committee: Spring 2013—Spring 2015

- Member, IT Council: Fall 2012—Spring 2015
- Member, Research Development Council: Fall 2009—Spring 2010, Fall 2012—Spring 2015
- Member, campus committee to review dean of the iSchool: Fall 2009—Spring 2010

Cloudera Exhibit 1006 - Page 86 of 86