Curriculum Vitae
Lan Du

Building E6A, Room 371
Department of Computing
Macquarie University
Sydney, NSW, 2019
Australia

Office: (61) 2 9850 9571
Mobile: (61) 433018598
Email: dulan520@gmail.com
Homepage: http://web.science.mq.edu.au/∼ldu/

Research Interests

His research interests focus on statistical modelling and learning for text analysis, which broadly covers statistical machine leaning, natural language processing, data mining, information retrieval and social network analysis. He is particularly interested in the following research areas

Probabilistic topic modelling: He is interested in developing advanced topic models that can take into account diverse information or features exhibited by free language text, for example, topic models exploring various discourse constraints (either semantic or syntactic constraints), topic models considering temporal information, and topic models with authorship and the relations between authors. He is also interested in the evaluation and application of topic models, for example, applying topic models to text segmentation, document summarisation/classification, information retrieval, sentiment analysis and image processing.

Nonparametric Bayesian methods: He has broad interests in the study of nonparametric Bayesian methods, e.g., Dirichlet process, Pitman-Yor process, and Indian Buffet process. He has developed two novel sampling methods for doing posterior inference for both Dirichlet process and Pitman-Yor process. One of the two samplers has recently been accelarated by a group of researchers at Google to handle large scale data sets. He is also interested in extending the existing nonparametric Bayesian methods to deal with more complex structures. For example, the use of a transformed base measure (known as transformed Pitman-Yor process) to handle different word usages across different corpora, and the use of multiple base measures (knowns as the multi-floor Chinese restaurant process) to do statistical adaptation.

Inference/optimisation algorithms: He is particularly interested in Variational Bayesian (VB) inference, MCMC methods and their combination. For example, the online VB, stochastic VB, Gibbs sampling, Slice sampling, and Metropolis Hasting. He is also interested in numerical optimisation algorithms, such as Qusai-Newton methods (e.g., BFGS and LBFGS) and gradient methods.

Matrix factorisation: He is recently interested in matrix factorisation methods for social network analysis, relation learning and learning. For example, probabilistic matrix factorisation, matrix co-factorisation, and high-order tensor factorisation have been widely used in social network analysis and relational learning. He has great interest in collaborative filtering, for example the combination of matrix factorisation techniques and probabilistic generative models to learn both relations and their semantic meaning.

Statistical ranking models: Label ranking is an important task in, for example, preference learning. The existing ranking methods include but not limited to the pair-wised ranking models, distance based ranking models and multistage ranking models. However, if the number of items to be ranked is countably infinite, it is unfeasible for each ranker to rank all the items. Therefore, one has to consider

partial rankings as either top-t orderings or incomplete orderings. Here, he is interested in both partial ranking models and infinite ranking models, and applying them to, for example, topic segmentation.

High performance computing: Scaling up existing algorithms, for example, various inference algorithms for topic models, is also one of his research interests. Large scale data processing can bridge the gap between advance Machine learning techniques and real-word applications. He is interested in parallel techniques, such as MapReduce, and the use of clusters. He has recently parallelised his topic segmentation model published in NAACL13 using multithreading techniques.

Models of Dynamics: Currently, he is involved in a Google project, where time-servers models are needed to model sequences of life events of each individual entity. He is particularly interested in models, like hidden Markov models, factorised hidden Markov model and infinite hidden Markov model, and adapting these models to modelling temporal information that exists in either structured or unstructured data. It is interesting that the model that he is developing for the Google project can also be applied to modelling, for example, the development of disease of patients.

Besides, he is also interested in 1) relation extraction and relational learning for enhancing existing knowledge bases (e.g., freebase), 2) statistical language models (*i.e.*, n-gram model), and 3) language acquisition (*e.g.*, Bayesian word segmentation)

Education

B.S. Information and Communication Technology, Flinders University, Australia, Jul. 2004 – Jul. 2006.

B.S. Information Technology with 1st class honours, The Australian National University, Australia, Feb. 2007 – Dec. 2007.

Thesis title: *Ontoloty-Driven Information Retrieval for Digital Forensics*

Honours project supervisor: Dr. Huidong Jin

Ph.D. Computer Science, The Australian National University, Australia, Aug. 2008 - Dec. 2011.

Thesis title: *Nonparametric Bayesian Methods for Structured Topic Models*

Supervisors: Prof. Wray Buntine and Dr. Huidong Jin

Degree awarded: December, 2012

Employment History

Research fellow, Department of Computing, Macquarie University, Dec. 2011 - now.

Supervisor: Prof. Mark Johnson

Awards

National Natural Science foundation of China, Grant No. 61402312 (AUD 50,000)
Project: "Intention-oriented Trajectory Search and Recommendation System"
Chief investigators: Associate Professor Jiajie Xu (Soochow University, China) and Dr. **Lan Du**

Google Natural Language Understanding focused awards 2013, (USD225,000).
Project: "Generative Models of Life Stories"
Chief investigators: Professor Mark Johnson (Macquarie University, Australia), Dr. **Lan Du**, and Professor Wray Buntine (Monash University, Australia)
Role: postdoctoral Fellowship

Scholarships

ANU-NICTA PhD scholarship, 2008–2011.

NICTA PhD Supplementary scholarship, 2008–2011.

NICTA PhD Assignment scholarship, 2008–2011.

ANU summer research scholarship, 2008.

ANU-NICTA honours scholarship, 2007.

Publications

Journal Articles

1. Dat Quoc Nguyen, Richard Billingsley, **Lan Du**, and Mark Johnson "Improving Topic Models with Latent Feature Word Representations", ***Transactions of the Association for Computational Linguistics***, Volume 3, page 299-313, 2015

2. Changyou Chen, Wray Buntine, Nan Ding, Lexing Xie, and **Lan Du** "Differential Topic Models", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Volume 37, Issue 2, page 230-242, 2015.
(rank A*, impact factor 5.694)

3. Jianping Gou, Yongzhao Zhan, Min Wan, Xiangjun Shen, Jinfu Chen, **Lan Du** "Maximum Neighborhood Margin Discriminant Projection for Classification", ***The Scientific World Journal***, accepted, 2014
(impact factor 1.219)

4. **Lan Du**, Wray Buntine, Huidong Jin, Changyou Chen, "Sequential Latent Dirichlet Allocation", ***Knowledge and Information Systems***, Volume 31, Number 3, 475-503, 2012
(rank B, impact factor 2.639)

5. Jianping Gou, Zhang Yi, **Lan Du**, Taisong Xiong, "A Local Mean-Based K-Nearest Centriod Neighbor Classifier", ***The Computer Journal***, Volume 55, Issue 9, 1058-1071, 2012
(rank A*, impact factor 0.888)

6. Jianping Gou, **Lan Du**, Taisong Xiong, "Weighted K-Nearest Centroid Neighbor Classification", ***Journal of Computational Information Systems*** 8: 2 (2012) 851-860

7. Jianping Gou, **Lan Du**, Yuhong Zhang, Taisong Xiong, "A new distance-weighted k-nearest neighbor classifier", ***Journal of Information and Computational Science***, 9: 6 (2012) 1429-1436

8. **Lan Du**, Wray Buntine, Huidong Jin, "A Segmented Topic Model based on the Two-parameter Poisson-Dirichlet Process", ***Machine Learning Journal***, Volume 81, Number 1, page 5-19, 2010
(rank A*, impact factor 1.689)

Conference papers

1. Zhendong Zhao, **Lan Du**, Benjamin Börschinger, John K Pate, Massimiliano Ciaramita, Mark Steedman and Mark Johnson, "A Computationally Efficient Algorithm for Learning Topical Collocation Models", To appear in **ACL2015** (CORE rank A*)

2. **Lan Du**, John K. Pate and Mark Johnson, "Topic Segmentation with An Ordering-Based Topic Model", page 2232–2238, **AAAI 2015** (CORE rank A*)
3. **Lan Du**, John K Pate, and Mark Johnson "Topic Models with Topic Ordering Regularities for Topic Segmentation", page 803–808, **ICDM 2014** (CORE rank A*)
4. Youliang Zhong, **Lan Du**, Jian Yang "Learning Social Relationship Strength via Matrix Co-Factorization with Multiple Kernels", The 14th International Conference on Web Information System Engineering (**WISE**), page 15–28, 2013 (CORE rank A)
5. **Lan Du**, Wray Buntine, and Mark Johnson, "Topic segmentation with a structured topic model," **NAACL**, pages 190–200, 2013 (CORE rank A)
6. Huidong Jin, Lijiu Zhang, **Lan Du**, "Semantic Title Evaluation and Recommendation Based on Topic Models", The 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (**PAKDD**), pages 402-413, 2013 (CORE rank A)
7. **Lan Du**, Wray Buntine, Huidong Jin, "Modelling Sequential Text with an Adaptive Topic Model", Empirical Methods in Natural Language Processing (**EMNLP**), page 535-545, 2012 (CORE rank A)
8. Changyou Chen, **Lan Du**, Wray Buntine, "Sampling Table Configurations for the hierarchical Poisson-Dirichlet Process", The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (**ECML-PKDD**), page 296-311, 2011 (CORE rank A)
9. **Lan Du**, Wray Buntine, Huidong Jin, "Sequential Latent Dirichlet Allocation: Discover Underlying Topic Structures within a Document", Proceedings of the 2010 IEEE International Conference on Data Mining (**ICDM**, in top 10 papers), page 148-157, 2010. (CORE rank A*)
10. **Lan Du**, Wray Buntine, Huidong Jin, "A Segmented Topic Model based on the Two-parameter Poisson-Dirichlet Process", The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (**ECML-PKDD** 2010, in top 7 papers, published as a journal article). (CORE rank A)
11. Wray Buntine, **Lan Du**, Petteri Nurmi, "Bayesian Networks on Dirichlet Distributed Vectors", Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (**PGM**-2010), page 33-40, 2010.
12. **Lan Du**, Huidong Jin, Olivier Y. de Vel, Nianjun Liu, "A Latent Semantic Indexing and WordNet based Information Retrieval Model for Digital Forensics", Proceedings of IEEE International Conference on Intelligence and Security Informatics (**ISI**), page 70-75, 2008. (CORE rank C)
13. Phan Thien Son, **Lan Du**, Huidong Jin, Olivier Y. de Vel, Nianjun Liu, Terry Caelli, "A Simple WordNet-Ontology Based Email Retrieval System for Digital Forensics", Pacific Asia Workshop on Cybercrime and Computer Forensics (ISI workshop), page 217-228, 2008.
14. Junlei Song, Dianhong Wang, Nianjun Liu, Li Cheng, **Lan Du**, Ke Zhang, "Soil Moisture Prediction with Feature Selection Using a Neural Network", Proceedings of International Conference on Digital Image Computing: Techniques and Applications (**DICTA**), page 130-136, 2008. (CORE rank B)

Google scholar citations

http://scholar.google.com.au/citations?user=HtiTsgwAAAAJ&hl=en

Professional Activities

Conference programme committee:

The Annual Meeting of the Association for Computational Linguistics (ACL) 2014, 2015
The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2013, 2014, 2015
Empirical Methods in Natural Language Processing (EMNLP) 2014, 2015
The 2014 Australasian Language Technology Association Workshop (ALTA) 2014, 2015
The International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, 2015
Neural Information Processing Systems (NIPS), 2015
The Asian Conference on Machine Learning (ACML) 2014, 2015
The International Joint Conference on Artificial Intelligence (IJCAI), 2011, 2013
The International Conference on Artificial Neural Networks (ICANN), 2011, 2013

Journal reviewer:

Journal of Machine Learning Research
IEEE Transactions on Neural Networks and Learning Systems
ACM transactions on Intelligent Systems and Technology
Neural Computing & Applications journal
Information Processing & Management
Neurocomputing

Technical talks delivered

1. "The Pitman-Yor Process and Its Variants", talk at Department of Computing, Macquarie University, Sydney, July 2014
2. "Bilinear Tensor Models", talk at Department of Computing, Macquarie University, April, 2014
3. "Topic Segmentation with a Structured Topic Model", invited talk at Google Zurich centre, in February, 2014
4. "Topic Models with Discourse Constraints", invited talk at Soochow University, China, in November, 2013
5. "Dirichlet Distribution and Dirichlet Tree", talk at Department of Computing, Macquarie University, Sydney, September, 2013
6. "CRP, Stirling Number and Gibbs Sampler", talk at Department of Computing, Macquarie University, November, 2012
7. "Bayesian Unsupervised Learning", talk at Department of Computing, Macquarie University, March, 2012
8. "Introducing Dependencies into Dirichlet Process", talk at Department of Computing, Macquarie University, March, 2012
9. "Segmented Topic Model", invited talk at IBM China research lab, Beijing, China, in August, 2010
10. "Structured Topic Modelling", invited talk at the Machine Intelligence Laboratory, Sichuan University, China, in September, 2010
11. "A network of Dirichlet distributions", invited talk at School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), China, in Oct 2010
12. "Sequential Latent Dirichlet Allocation: Discover Underlying Topic Structures within a Document", conference talk at ICDM 2010, Sydney, Australia, December 2010.