ChengXiang Zhai

Department of Computer Science
University of Illinois at Urbana-Champaign
1304 West Springfield Avenue
Urbana, IL 61801
U.S.A.

Office Phone: (217) 244-4943
Office Fax: (217) 244-6500
Home Phone: (217) 356-1292
Email: czhai@cs.uiuc.edu
http://www-faculty.cs.uiuc.edu/~czhai

Research Interests

I am broadly interested in developing algorithms for text information management, including retrieval, filtering, classification, summarization, and mining of text. I am especially interested in approaches that exploit statistical language modeling and machine learning. My current focus is on the application of statistical language modeling to information retrieval. I also have a strong interest in bioinformatics, including both biological text processing and sequence analysis, with an emphasis on integrating text mining with sequence mining to discover meaningful biological sequence patterns.

Education

Aug. 2002	**Carnegie Mellon University**, Pittsburgh, Pennsylvania Ph.D. in Language and Information Technologies Advisor: John Lafferty
May 1995	**Carnegie Mellon University**, Pittsburgh, Pennsylvania M.S. in Computational Linguistics Advisor: David Evans
Dec. 1990	**Nanjing University**, Nanjing, China Ph.D. in Computer Science Advisor: Jiafu Xu
July 1987	**Nanjing University**, Nanjing, China M.S. in Computer Science Advisor: Guoliang Zheng
July 1984	**Nanjing University**, Nanjing, China B.S. in Computer Science

Employment

Aug. 2002 – Present	**University of Illinois at Urbana Champaign**, Urbana, Illinois Assistant Professor
Nov. 2000 – Aug. 2002	**Carnegie Mellon University**, Pittsburgh, Pennsylvania Senior Research Programmer, Lemur project
Jan. 1997 – Nov. 2000	**Claritech (now Clairvoyance) Corporation**, Pittsburgh, Pennsylvania Research Scientist, Senior Research Scientist (June 2000 – Nov. 2000)
Sep. 1990 – July 1993	**State Key Laboratory for Novel Software Technology**, Research Associate, Nanjing University, Nanjing, China

Research Experience

Nov. 2000 – Aug. 2002 **Carnegie Mellon University**, Pittsburgh, Pennsylvania
Dissertation Research, Advisor: John Lafferty

My dissertation is titled "Risk Minimization and Language Modeling in Text Retrieval". It presents a new text retrieval framework based on Bayesian decision theory. The framework unifies many existing retrieval models within one general probabilistic framework, and facilitates the development of new principled approaches to text retrieval that use statistical language models. The dissertation explores several special cases of this framework, and develops a number of new language model-based retrieval methods. Through these specific models, the thesis demonstrates the generality of the proposed retrieval framework and the advantage of using language models in text retrieval, such as being able to optimize retrieval parameters automatically.

Nov. 2000 – Aug. 2002 **Carnegie Mellon University**, Pittsburgh, Pennsylvania
Senior Research Programmer, Lemur project

I contributed extensively to the design, implementation, and evaluation of the Lemur toolkit for language modeling and information retrieval.

Jan. 1997 – Nov. 2000	**Claritech (now Clairvoyance) Corporation**, Pittsburgh, Pennsylvania Research Scientist, Senior Research Scientist (June 2000 – Nov. 2000) I conducted original and evaluation research on information retrieval and filtering, especially with the CLARIT approach. I led the CLARITECH filtering team to participate in the TREC-7 and TREC-8 filtering tracks and achieved excellent system performance. I designed and implemented a research-oriented C++ toolkit for retrieval and filtering.
Aug. 1993 – Dec. 1996	**Carnegie Mellon University**, Pittsburgh, Pennsylvania Graduate Research Assistant, Helios project I participated in the Helios project, which uses the CLARIT system for indexing and retrieval of Heinz data. I also participated in CLARIT TREC experiments in 1994, 1995, 1996, including ad hoc retrieval, routing, natural language processing track, interactive track, Chinese track, and confusion track. I conducted research on noun phrase analysis and its use in document indexing.
Sep. 1990 – July 1993	**State Key Laboratory for Novel Software Technology**, Research Associate, Nanjing University, Nanjing, China I conducted research on Chinese language understanding, and proposed a conceptual programming language. I participated in the design and implementation of a graphic editing interface using C and OpenLook on a Sun-3 workstation, which allows users to specify object model, dynamic model, and the functional model for object-oriented software design.
Sep. 1984 – Dec 1990	**Nanjing University, China** M.S. and Ph.D. thesis research (both in Computer Science) In my M.S. thesis research, I designed and implemented the kernel of an integrated programming environment for Modula-2 on IBM-PC, which supports incremental compilation and separate debugging. In my Ph.D. thesis research, I proposed and studied an extended categorial grammar for Chinese as well as the use of the grammar for parsing Chinese sentences. I explored the issue of syntactic ambiguity of Chinese in the proposed grammar formalism.

Teaching Experience

Fall 2002	**University Illinois at Urbana Champaign**, Urbana, Illinois Instructor for "Information Retrieval" (graduate course CS497)
Fall 2001	**Carnegie Mellon University**, Pittsburgh, Pennsylvania Teaching assistant for "Machine Learning" (upper-level undergraduate and graduate) Duties: office hours, grading, guest lecture
Spring 1986	**Nanjing University**, Nanjing, China Teaching assistant for "Software Engineering" (upper-level undergraduate) Duties: office hours, grading
Fall 1985	**Nanjing University**, Nanjing, China Teaching assistant for "Pascal Programming" (lower-level undergraduate) Duties: office hours, laboratory supervision, grading

Awards

1993–1996	**Computational Linguistics Research Fellowship**, Carnegie Mellon University, Pittsburgh, Pennsylvania
1993	**Humboldt Research Fellowship**, Alexander von Humboldt Foundation, Bonn, Germany (declined)
1990	**First-Class Guanghua Scholarship**, Nanjing University, Nanjing, China
1985	**Distinguished Graduate Student Award**, Nanjing University, Nanjing, China

Publications

Refereed Publications

1. ChengXiang Zhai and John Lafferty. Two-stage language models for information retrieval. In *Proceedings of SIGIR'2002*, pages 49–56, Aug 2002.

2. Rong Jin, Alex G. Hauptmann, and ChengXiang Zhai. Title language model for information retrieval. In *Proceedings of SIGIR'2002*, pages 42-48, Aug 2002.

3. Rong Jin, Alex G. Hauptmann, and ChengXiang Zhai. A content-based probabilistic correction model for OCR document retrieval In *Proceedings of the SIGIR 2002 Workshop on Information and Retrieval and OCR: From Converting Content to Grasping Meaning*, 2002.

4. Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, pages 403–410, Nov. 2001.

5. Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'2001*, pages 334–342, Sept 2001.

6. John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'2001*, pages 111–119, Sept 2001.

7. Chengxiang Zhai and John Lafferty. The dual role of smoothing in the language modeling approach. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, pages 31–36, 2001.

8. John Lafferty and Chengxiang Zhai. Probabilistic IR models based on query and document generation. In *Proceedings of the Language Modeling and IR workshop*, pages 1–5, May 31 – June 1 2001.

9. Chengxiang Zhai, Peter Jansen, and David A. Evans. Exploration of a heuristic approach to threshold learning in adaptive filtering. In *Proceedings of the SIGIR 2000*, pages 360–362, 2000. Poster presentation.

10. Chengxiang Zhai. Exploiting context to identify lexical atoms – a statistical view of linguistic context. In *Proceedings of 1997 International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97)*, pages 119–129, 1997.

11. Chengxiang Zhai. Fast statistical parsing of noun phrases for document indexing. In *5th Conference on Applied Natural Language Processing (ANLP-97)*, pages 312–319, March 31 – April 3 1997.

12. David A. Evans and Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of ACL 1996*, pages 17–24, 1996.

13. Zhijian Wang, Chengxiang Zhai, and Jiafu Xu. Exploration of the basic concepts in object-oriented programming. *Computer Science Survey*, 19(4):14–17, April 1992. In Chinese.

14. Yanbing Wang, Jiazhong Zhang, and Chengxiang Zhai. Exploration of Chinese syntactic theory for machine translation. *Pattern Recognition and Artificial Intelligence*, 5(1):15–23, March 1992. In Chinese.

15. Chengxiang Zhai, Jiazhong Zhang, and Yanbing Wang. A formal study of syntactic ambiguity in Chinese. In *Proceedings of the 1991 International Conference for Young Computer Scientists (ICYCS'91)*, pages 608–611, July 1991.

16. Chengxiang Zhai, Yanbing Wang, Jiazhong Zhang, and Jiafu Xu. A combinatory type grammar of Chinese. *Journal of Chinese Information Processing*, 5(3):1–6, 1991. In Chinese.

17. Chengxiang Zhai. Preliminary ideas of a conceptual programming language. *ACM SIGPLAN Notices*, 25(12):93–100, December 1990.

18. Guoliang Zheng and Chengxiang Zhai. A Modula-2 integrated environment and its kernel. *Chinese Journal of Computers*, 13(5):367–373, May 1990. In Chinese.

19. Min Dai, Ning Xiao, and Chengxiang Zhai. On the verification of abstract data types. In *1989 Beijing International Symposium for Young Professionals (BISYCP' 89)*, pages 267–270, Aug 1989.

20. Chengxiang Zhai and Guoliang Zheng. Incremental compilation in a Modula-2 integrated environment. *Microcomputer*, (1):24–26, 1989. In Chinese.

21. Guoliang Zheng and Chengxiang Zhai. The kernel of a Modula-2 integrated environment. *ACM Software Engineering Notes*, 13(4):57–60, October 1988.

22. Chengxiang Zhai and Guoliang Zheng. Separate debugging of Modula-2 program. *Computer and Modernization*, (2):5–9, 1988. In Chinese.

23. Guoliang Zheng, Xiofeng Ye, Laiqiang Wang, Qing Tang, and Chengxiang Zhai. FPTS – an automatic translation system from FORTRAN to PASCAL. *Computer and Modernization*, (1-2):1–6, 1987. In Chinese.

24. Ke Chen and Chengxiang Zhai. PT: A syntax-directed editor for high-level programming language. *Journal of Computer Applications*, 10(1):31–39, 1987. In Chinese.

25. Chengxiang Zhai. FMT – a Pascal formatting program. *Mini-Micro computers*, (6):28–32, 1986. In Chinese.

Unrefereed Publications

1. Chengxiang Zhai, Peter Jansen, Norbert Roma, Emilia Stoica, and David A. Evans. Optimization in CLARIT adaptive filtering. In E. Voorhees and D. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 253–258, 2000.

2. Chengxiang Zhai, Peter Jansen, Emilia Stoica, Norbert Grot, and David A. Evans. Threshold calibration in CLARIT adaptive filtering. In E. Voorhees and D. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 149–156, 1999.

3. Chengxiang Zhai, Xiang Tong, Natasa Milic-Frayling, and David A. Evans. Evaluation of phrase-based index – CLARIT TREC-5 NLP track report. In D. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 347–357, 1997.

4. Xiang Tong, Chengxiang Zhai, Natasa Milic-Frayling, and David A. Evans. Experiment on Chinese text indexing — CLARIT TREC-5 Chinese track report. In D. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 335–339, 1997.

5. Xiang Tong, Chengxiang Zhai, Natasa Milic-Frayling, and David A. Evans. OCR correction and query expansion for retrieval on OCR data — CLARIT TREC-5 confusion track report. In D. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 341–345, 1997.

6. Natasa Milic-Frayling, Xiang Tong, Chengxiang Zhai, and David A. Evans. CLARIT compound queries and constraint-controlled feedback in TREC-5 ad hoc experiments. In D. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 315–333, 1997.

7. Natasa Milic-Frayling, Chengxiang Zhai, Xiang Tong, Michael P. Mastroianni, and David A. Evans. CLARIT TREC-4 interactive experiments. In D. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 323–357, 1996.

Professional activities and service

Program Co-chair, ACM CIKM 2004

Program committee member, ACM SIGIR 2003

Program committee member, ACL 2003

Program committee member, ACM CIKM 2003

Program committee member, ICML-2002 Workshop on Text Learning (TextML'2002)

Program committee member, IEEE ICDM-2001 Workshop on Text Mining (TextDM' 2001),

Ad hoc reviewer for major journals of information retrieval (e.g., ACM Transactions on Information Systems, Journal of Information Retrieval, Information Processing and Management).

President, Chinese Graduate Student Organization (CGSO), Carnegie Mellon University, 1994-1995

References

Jamie Callan	School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 (412) 268-4525 callan+@cs.cmu.edu
Jaime Carbonell	School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 (412) 268-7279 jgc@cs.cmu.edu
W. Bruce Croft	Department of Computer Science University of Massachusetts, Amherst Amherst, MA 01003 (413) 545-0463 croft@cs.umass.edu
David A. Evans	Clairvoyance Corp. 5301 5th Avenue Pittsburgh, PA 15232 (412) 621-0570 d.evans@clairvoyancecorp.com
John Lafferty	School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 (412) 268-6791 lafferty@cs.cmu.edu
Roni Rosenfeld	School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213 (412) 268-7678 roni@cs.cmu.edu