# Chun Jin

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213
1-732-462-9421 (home) 1-732-804-6536 (cell)
http://www.cs.cmu.edu/~cjin/
cjin@cs.cmu.edu

## Research Interests

* Database systems: stream databases, query optimization, and database applications.
* Information retrieval, and topic detection and tracking.
* Natural language processing and text summarization.
* Applied machine learning and data mining.
* Speech and multimedia processing.

## Education

*Carnegie Mellon University, Pittsburgh, PA*
**Ph.D. in Language and Information Technologies** **10/2006**
Thesis: "Optimizing Multiple Continuous Queries"

*Columbia University, New York, NY*
**M.S. in Computer Science** **1999**

*Xidian University, Xi'an, China*
**M.S. in Computer Science** **1996**
**B.S. in Computer Science** **1993**

## Research

*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA*
**Post Doctoral Fellow**

- **Research Grant Proposal Preparation** **2007**
  Preparing, with the team, the AWARD-WINNING research grant proposal RAPID on data uncertainty and predictive modeling for intelligence analysis tasks.
- **Multiple Continuous Query Optimization** **2006 - 2007**
  Extending the ARGUS work, including benchmarking the data and query repository, publishing the source code, and working on distributed query processing.

**Research Assistant**

- **Analyst Resource for Guarding the U.S. (ARGUS)** **2003 - 2006**
  Designed a stream monitoring system atop of a commercial DBMS to support running multiple continuous queries over data streams with shared query evaluation plans (thesis). ARGUS provides intelligence analysis tools to detect and monitor special event patterns from massive structural data, including data streams. As part of the project, my work addresses the challenges posed by efficiently monitoring multiple complex queries. Fast continuous monitoring is realized by the Rete-based incremental evaluation scheme on selections, joins, and aggregations with materialized intermediate results. Various optimization techniques, e.g. transitivity inference and conditional materialization, are integrated into the system. The large-scale problem is addressed with the incremental multiple query optimization (IMQO) approach. The system indexes existing plan computations R, searches the common computations between the new query Q and R, chooses the optimal sharing paths, and expands R to obtain final results for Q. The combination of the techniques is up

to hundreds of times faster than the naïve approach (re-evaluating unshared query plans when new data arrive) on typical intelligence analysis queries. The system has a broad range of applications, and is designed to be easily deployed for monitoring various data streams. We have experimented with Fedwire money transfer transactions and patient medical records for detecting possible frauds and disease outbreaks. We are considering deployment of the system to national geo-spatial databases, and we plan to work on network traffic data.

- **Topic Detection and Tracking (TDT)** **2001-2002**
  Incorporated the named-entities and time-dampening factors into the detection decision making scheme of CMU TDT system, and participated NIST TDT evaluations. The system detects new events in streaming text data with techniques of information retrieval and clustering.
- **Text Summarization** **2001**
  Extended an existing text summarizer to support meaningful summarization on Spanish and Chinese documents, beyond English and Japanese ones, and integrated the summarizer into a Topic Detection and Tracking System in real use by government agencies. The summarizer summarizes a long document or multiple documents into an informative short summary with Maximal Marginal Relevance (MMR) approach.
- **Is this conversation on track?** **2000**
  Examined the features that accurately measure the speech recognition confidence of a dialog utterance with annotated Communicator data logs using neural networks. And the confidence measure is used to improve Communicator dialog management.

*Department of Computer Science, Columbia University, New York, NY*
**Research Assistant**

- **Multimedia Abstract Generation for Intensive Care (MAGIC)** **1999**
  Built tools to extract semantic and syntactic information from the feature descriptions produced by MAGIC text generator and language parsing results to facilitate knowledge base building and retrieval for speech synthesis, with which and other techniques MAGIC provides patient information retrieved from databases to medicare professionals in forms of speech and graphics.

*Department of Computer Science, Xidian University, Xi'an, China*
**Research Assistant**

- **Chinese Text Compression** **1995-1996**
  Developed text compression tools based on the modified dictionary compression algorithms to support Chinese text compression.
- **Intelligent Process Control System** **1992-1993**
  Designed the hardware and drivers of Intelligent Process Control System based on Intel 8031-microcomputer-chip. The system was deployed in a cement manufacturer.

## Working Experience

*Kenosia Corporation, Danbury, CT*
**Software Engineer**

- **dataAlchemy** **2000**
  Improved dataAlchemy Version 2.0, i.e. enhancing data importing, data querying, data analysis, and interactive presentation development. dataAlchemy is Kenosia's flagship software for category management, providing manufactures and retailers with data management and analysis tools to draw actionable conclusions from proprietary retail data to maximize their revenues, and with presentation development tools to convey their results effectively.

*Iconic Data Information Ltd. Beijing, China*
**Software Engineer**

- **Image Processing Application** **1996-1997**
  Extended an image processing application that displays and edits digital images to support various formats, e.g. bmp, gif, jpeg.

*Department of Computer Science, Beijing Information Technology Institute, Beijing, China*
**Instructor**

- **Computer Architecture Lab** *for undergraduate students* **1996-1997**
  Instructed and guided the circuit-building lab for senior students.

**Co-instructor**

- **Computer Architecture** *for undergraduate students* **1996-1997**
  Gave lectures and graded partial homework problems.

## Teaching

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA*
**Teaching Assistant**

- **Web Commerce, Security, and Privacy** *for undergraduate and graduate students* **2005**
  Gave a guest lecture, designed parts of the homework problems, guided a course project group, held office hours, and graded homework and exams.
- **Discovery Technology Project** *for CMU West graduate students* **2003**
  Remotely assisted the project by setting up the project environment, and answering questions by email.
- **Web-based Information Architecture** **2001, 2003**
  Assisted the course for three classes, one for CMU eCommerce graduate students, one for a class of visiting students from Korea, and one for a distance learning class of Pratt & Whitney employees. I held recitations, designed and graded parts of homework and exam problems.

*Department of Computer Science, Columbia University, New York, NY*
**Teaching Assistant** *for undergraduate and graduate students*

- **Programming Language and Translators** **1998**
- **Computer Networks** **1998**
  For both courses, I graded homework and held office hours.

## Publications

**Refereed Publications**

[1] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, and Chun Jin. "Red Opal: Product-Feature Scoring from Reviews". To appear in *Proceedings of the 8th ACM Conference on Electronic Commerce (EC) 2007.* **2007**

[2] Chun Jin and Jaime Carbonell. "ARGUS: Efficient Scalable Continuous Query Optimization for Large-Volume Data Streams". In *Proceedings of the International Database Engineering and Applications Symposium (IDEAS) 2006.* **2006**

[3] Chun Jin and Jaime Carbonell. "Incremental Aggregation on Multiple Continuous Queries". In *Proceedings of the 16th International Symposium on Methodologies for Intelligent Systems (ISMIS),* 2006. **2006**

[4] Chun Jin, Jaime Carbonell, and Phil Hayes. "ARGUS: Rete + DBMS = Efficient Continuous Profile Matching on Large-Volume Data Streams". In *Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems (ISMIS)*, pages 142-151, 2005. **2005**

[5] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. "Topic-conditioned Novelty Detection". In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 688-693, 2002. **2002**

[6] Paul Carpenter, Chun Jin, Daniel Wilson, Rong Zhang, Dan Bohus, and Alexander Rudnicky. "Is This Conversation on Track?". In *Proceedings of EuroSpeech 2001*, pages 2121-2124. **2001**

**Technical Reports**

[7] Chun Jin and Jaime Carbonell. "ARGUS: Efficient Scalable Continuous Query Optimization for Large-Volume Data Streams". Technical Report, CMU-LTI-06-005, 2006. **2006**

[8] Chun Jin and Jaime Carbonell. "ARGUS: Rete + DBMS = Efficient Persistent Profile Matching on Large-Volume Data Streams". Technical Report, CMU-LTI-04-181, 2004. **2004**

**Poster Publications**

[9] J. Carbonell, E. Fink, C. Jin, C. Gazen, J. Mathew, A. Saxena, V. Satish, S. Ananthraman, D. Dietrich, G. Mani, J. Tittle, and P. Durbin. "Scalable Data Exploration and Novelty Detection". *NIMD Grand Finale PI Meeting*, Arlington, VA, 2006 **2006**

[10] Jaime Carbonell, Eugene Fink, Chun Jin, Cenk Gazen, Santosh Ananthraman, Phil Hayes, Ganesh Mani, and Dwight Dietrich. "Exploring Massive Structured Data in ARGUS". *NIMD PI Meeting*, Orlando, FL, 2005. **2005**

[11] Jaime Carbonell, Phil Hayes, Eugene Fink, Chun Jin, and Cenk Gazen. "Approaches to Massive Structured Data in Argus". *NIMD PI Meeting*, Orlando, FL, 2004. **2004**

[12] Cenk Gazen, Jaime Carbonell, Phil Hayes, Chun Jin, and Eugene Fink. "Hypothesis Formation and Tracking in ARGUS". *NIMD PI Meeting*, Orlando, FL, 2004. **2004**

[13] Jaime Carbonell, Chun Jin, and Phil Hayes. "Monitoring Large, Constantly Incrementing Collections of Structured Data for Complex Watch Patterns". *NIMD PI Meeting*, Crystal City, VA, 2004. **2004**

[14] Jaime Carbonell, Cenk Gazen, Chun Jin, Phil Hayes, Aaron Goldstein, Ganesh Mani, and Johny Mathew. "Finding Novel Information in Large, Constantly Incrementing Collections of Structured Data". *NIMD PI Meeting*, San Diego, CA, 2003. **2003**

## Presentations

[1] "Finding Novel Patterns in Large, Constantly Incrementing Collections of Structured Data", by Santosh Ananthraman, Dwight Dietrich, Ganesh Mani, Abhay Saxena, Vini Satish, Chun Jin, and Eugene Fink. NIMD Grand Finale PI Meeting, Arlington, VA, 2006 (poster and demo). **2006**

[2] "Toward Incremental Sharing on Large-Scale Continuous Queries", by Chun Jin. LTI Seminar, CMU. **3/2006**

[3] "Finding Novel Information in Large, Constantly Incrementing Collections of Structured Data", by Santosh Ananthraman, Phil Hayes, Ganesh Mani, and Chun Jin. NIMD PI Meeting, Orlando, FL, 2005 (poster). **11/2005**

[4] "Project ARGUS", by Jaime Carbonell, Phil Hayes, Santosh Ananthraman, Chun Jin, Ganesh Mani, and Dwight Dietrich. ARDA NIMD Site Visit Report on ARGUS, CMU. **10/2005**

[5] "ARGUS: Rete + DBMS = Efficient Persistent Profile Matching on Large-Volume Data Streams", by Chun Jin. LTI Seminar, CMU. **4/2005**

[6] "CMU TDT Report", by Jaime Carbonell, Yiming Yang, Ralf Brown, Chun Jin, and Jian Zhang. TDT Workshop. **2001**

## Software

[1] ARGUS 1.0 monitors structural data stream for multiple continuous queries. It allows incremental query registration (IMQO), and provides efficient continuous query result generation (Rete-based query evaluation, and query optimization). See Research for details. **2007**

## Service

Reviewer for

NAACL-HLT 2007
IEEE Consumer Communications and Networking Conference (CCNC) 2007
COLING/ACL Conference 2006
IEEE International Conference on Multimedia & Expo (ICME) 2006, 2007
AXMEDIS 2006
ACM SIGIR Conference 2006
Journal of Data & Knowledge Engineering (DKE) 2006
IEEE International Conference on Intelligent Computing (ICIC) 2005
Book chapter on data mining 2003

Student Volunteer for

NAACL 2001

## Honors and Awards

Research fellowship, Carnegie Mellon University **2000-2006**
Outstanding Student Fellowship, Xidian University **1990-1993**

## Extracurricular Service

Member of Parent-Teacher Association, Huaxia Chinese School – South Branch, NJ, 2003-2005, received the Honor Certificate for Services.

## Skills

| | |
|---|---|
| Operating Systems: | Windows XP, LINUX/UNIX |
| Programming Languages: | Perl, C++/C, Java |
| Database Systems: | Oracle PL/SQL |
| Special Tools: | Matlab, SPlus/R |

## Status

U.S. Permanent Resident

## References


Allen Newell Professor **Jaime Carbonell** (Advisor)
Director of Language Technologies Institute
NSH 4519, LTI/SCS
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
Phone: 1-412-268-7279
Email: jgc@cs.cmu.edu

Dr. **Phil Hayes**
Project Leader, FirstGov
Vivisimo, Inc.
1710 Murray Ave, Suite 300
Pittsburgh, PA 15217
Phone: 1-412-422-2499
Email: hayes@vivisimo.com

Associate Professor **Jamie Callan**
Language Technologies Institute
NSH 3615, LTI/SCS
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
Phone: 1-412-268-4525
Email: callan@cmu.edu

Associate Professor **Norman M. Sadeh**
Director of Mobile Commerce Lab
Director of e-Supply Chain Management Lab
ISR/SCS, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
Phone: 1-412-268-8144
Email: sadeh@cs.cmu.edu