May 2, 2000

Tobun Dorbin Ng

School of Computer Science
Carnegie Mellon University
Wean Hall 3414
Pittsburgh, PA 15213

Office: 1-412-268-4499
Fax: 1-412-268-5576
dorbin.ng@cs.cmu.edu
http://www.cs.cmu.edu/~tng

Education

Doctor of Philosophy in Management, 2000
The University of Arizona, Tucson, AZ
Major: Management Information Systems
Minor: Computer Science
Major Advisor: Dr. Hsinchun Chen
Major Committee Members: Dr. Jay F. Nunamaker Jr., Dr. Olivia R. Liu Sheng
Minor Committee Members: Dr. Richard T. Snodgrass, Dr. John H. Hartman
Dissertation Title: A Concept Space Approach to Semantic Exchange

Master of Science in Management Information Systems, 1993
The University of Arizona, Tucson, AZ

Bachelor of Science in Business Administration, 1990
The University of Arizona, Tucson, AZ
Majors: Management Information Systems, Finance

Academic and Research Experience

Informedia Digital Video Library (1999-present):

Systems Scientist for "Informedia-II: Digital Video Library" project which is one of the Digital Libraries Initiative Phrase 2 projects funded by National Science Foundation (NSF). This project aims to transform the paradigm for accessing digital video libraries through meaningful, manipulable overviews of video document sets, multimodal queries, and adaptive summarizations of very large amounts of video from heterogeneous distributed sources. Video information collages are the key technology in Informedia-II and will be built by advancing information visualization research to effectively deal with multiple video documents. A video information collage is a presentation of text, images, audio, and video derived from multiple video sources in order to summarize, provide context, and communicate aspects of the content for the originating set of sources. I am responsible for research and system development system for this Informedia-II Project. (Principal investigators: H. Wactlar)

GeoWorlds, Technology Transfer & Integration (1998-1999):

Project coordinator and technical point-of-contact for "GeoWorlds: Integrated Digital Libraries and Geographic Information Systems" project funded by the Defense Advanced Research Projects Agency (DARPA). This project is to integrate Geographical Information Systems and Digital Library technologies into a single system. This system will support a strategic alliance of

technology developers and transfer the resulting technology to multiple military partners designated by DARPA; the partners are concerned with applying the technology in disaster relief operations. The integration is a collaborative effort from multiple sources (including the USC/ISI DASHER Project, the University of Arizona, the University of Illinois Urbana-Champaign Digital Library Initiative and NCSA, the University of California at Berkeley, and the University of California at Santa Barbara). The resulting DASHER services will be tested in military environments using disaster relief scenarios developed jointly with organizations such as SPAWAR System Center. I am responsible for coordinating, designing, and implementing application program interfaces (APIs) of Arizona's technological components. I am also responsible for delivering the APIs to and coordinating technology integration with USC/ISI. (Principal investigators: R. Neches, H. Chen)

Multimedia and Multi-dimensional Interactive Systems (1997-1999):

System designer and software engineer for "The Interspace Prototype: An Analysis Environment based on Scalable Semantics" project funded by the Defense Advanced Research Projects Agency (DARPA). This project is to build a complete prototype environment for semantic association of multimedia information and to evaluate its utility on real collections. The semantic association relies on statistical clustering for concepts and categories. Interactive navigation using the semantic association enables information retrieval at a deeper level than previously possible for diverse large collections. I am responsible for design and develop a three-dimensional prototype demonstrating the semantic interoperability between places on a three-dimensional geographic landscape and topics on a three-dimensional semantic landscape. (Principal investigators: B. Schatz, H. Chen)

COPLINK, Database & Intranet Integration (1997-1999):

System designer and software engineer for "COPLINK: Database Integration and Access for a Law Enforcement Intranet" project funded by the National Institute of Justice (NIJ). This project is intended to be a beginning and a catalyst to solve law enforcement information technology problems through a synergy between the research in Artificial Intelligence (AI) Lab at the University of Arizona and the application in the Tucson Police Department. Innovative and functional technologies will expand the uses of COPLINK from database integration, secured Intranet access, and mobile computing, to criminal intelligence and case analysis, mug-shot face recognition, and intelligent agents. I am responsible for designing an algorithm to correlate crime-related information such as people, locations, incidents, vehicles, and weapons. I am also responsible for designing and developing an Intranet prototype. (Principal investigator: H. Chen; Collaborator: D. Smith)

Geographic Information Systems (1996-1998):

Principal system architect for an NSF CISE-funded "Supplement to Alexandria DLI Project: A Semantic Interoperability Experiment for Spatially-Oriented Multimedia Data" project. This research aims to examine semantic interoperability issues related to spatially-oriented, multimedia geographic information access. Based on the concept space approach developed by the Illinois Digital Library Initiative (DLI) project and the Alexandria (University of California at Santa Barbara) geo-referenced collections, this research proposes to develop knowledge representations and structures to capture concepts of relevance to spatial and multimedia information (natural language phrases and geo-related textures). Selected machine learning techniques and general Artificial Intelligence (AI) graph traversal algorithms will also be adopted to assist in semantic, concept-based spreading activation in integrated knowledge networks. I am responsible for designing a system architecture to integrate various semantic components, defining and analyzing

interoperable and scalable semantic components in different media, and developing a web accessible system prototype. (Principal investigators: H. Chen, T. Smith)

Medical Informatics (1996-1999):

Principal system designer and software engineer for "Information Analysis and Visualization for Cancer Literature" project funded mainly by the National Cancer Institute (NCI). This project aims to develop concept spaces (networks of vocabularies) and category maps for cancer-related literature and documents. Based on selected automatic indexing, linguistic parsing, cluster analysis, neural networks, and information visualization techniques, the system-generated concept spaces and category maps will be used to enhance cancer-related information retrieval and knowledge sharing. I am responsible for designing and developing scalable algorithms and software to analyze automatically a million CancerLit records and a common gateway interface (CGI) server for semantic (concept-based) retrieval. I am also responsible for managing both HTML and Java interface development to provide Internet access to CancerLit server. (Principal investigators: B. Schatz, H. Chen, S. Hubbard)

Digital Libraries Initiative (1994-1999):

Principal system designer and software engineer for an NSF/NASA/ARPA-funded "Digital Library Initiative" project. The project goal is to develop a large-scale testbed for building next-generation digital libraries for the National Information Infrastructure (NII). The testbed collections are mainly in the engineering domains, to be contributed by major engineering societies and publishers (e.g., IEEE, John Wiley \& Sons, etc.). I am responsible for developing scalable analysis tools to analyze automatically a gigabyte of textual information for semantic (concept-based) retrieval. I am also responsible for developing both web server and interface for semantic retrieval. (Principal investigators: B. Schatz, H. Chen)

Supercomputing for Knowledge Discovery (1994-present):

Principal investigator for two and principal software engineer for eight National Center for Supercomputing Applications (NCSA) High-performance Computing Resources Grants:

1. "High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management" (1999-2000)
2. "Semantic Analysis on Large Image Collection" (1998-1999)
3. "Parallel Computation for a Semantic Interoperability Environment" (1997-1998)
4. "Medical Information Analysis and Knowledge Discovery" (1997)
5. "Parallel Semantic Analysis for Spatially-Oriented Multimedia GIS Data" (1996-1997)
6. "Information Analysis and Knowledge Discovery for Digital Libraries" (1995-1997)
7. "Terabyte Information Analysis and Knowledge Discovery" (1994-1995)
8. "Building the Interspace: Digital Library Infrastructure for a University Engineering Enviroment" (1994-1995)

These projects aim to generate concept spaces for domains in engineering, medical informatics, and Internet. Concept spaces then will be used to support concept-based retrieval and cross-domain vocabulary switching in scientific information retrieval. I am responsible for scaling various analysis algorithms to utilize the supercomputing resources - SGI/Cray Origin2000 (128 nodes), SGI's Power Challenge Array (148 nodes), Convex's Exemplar (64 nodes), and Thinking Machine's CM-5 (512 nodes) – which have been provided by NCSA. (Principal investigator: H. Chen, T. D. Ng; Collaborators: B. Schatz, L. Smarr)

Worm Community System (1993-1994):

Software engineer for an NSF-funded "National Collaboratory" project. The project goal is to design concept-based information retrieval and information sharing software for molecular biologists whose work is related to the Human Genome Initiative. A (nematode) worm concept space and a fly (Drosophila) concept space which can assist in cross-domain concept exploration and term suggestion during information retrieval have been created and are in use by worm biologists. I was responsible for developing a text parser to automatically index term phrases including scientific names like chemical compound and gene names from fly abstracts. In addition, I was responsible for creating the fly concept space. (Principal investigators: B. Schatz, H. Chen, S. Ward; Collaborators: T. Yim, D. Fye, J. Martinez, K. Powell, E. Grossman, T. Friedman, J. Calley)

Intelligence Analysis on International Information Technologies (1990-1994):

System designer and software engineer for an intelligence analysis and retrieval system, which supports intelligence analysts who study information technology policy, manufacturing, and proliferation in the (former) USSR countries. A content-based intelligence analysis and retrieval system was developed in ANSI C and runs on VAX/VMS and a DECStation (UNIX based). I was responsible for developing the content-based retrieval system component using branch-and-bound algorithm and for designing the system user interface. (Collaborators: H. Chen, S. Goodman, W. McHenry, P. Wolcott, K. Lynch, A. Himler, R. Orwig, K. Basu)

Neural Networks for Pharmaceutical Applications (1992-1994):

Principal system designer and software engineer for an NIH-funded project investigating a neural network approach to pharmaceutical applications. Sample applications include drug solubility prediction and non-linear pharmacokinetics functions approximation. I was responsible for developing a prediction system, which was based on a neural network model using a Backpropagation algorithm to estimate solubility of certain chemical compounds. I derived a systematic approach to train the Backpropagation network. The trained Backpropagation network out-performed regression model which was employed by the same application. (Principal investigators: H. Chen, H. Chow, S. Yakowski; Collaborator: P. Myrdal)

Teaching Experience

Instructor (1996), The University of Arizona, "Data Structures and Algorithms," (undergraduate course), using Pascal. Student instructor-evluations on *overall rating of instructor's effectiveness* based on the following 5-point scale: 5 for "almost always," 4 for "more than 1/2 of the time," 3 for "about 1/2 of the time," 2 for "less than 1/2 of the time," and 1 for "almost never."

- Spring 1996, rating: 4.43 (out of 5.00) **top 10%**
- Summer 1996 (10-week session), rating: 4.47 (out of 5.00) **top 10%**

Teaching Assistant (1990), The University of Arizona, "Data Structures and Algorithms" (Undergraduate course)

Teaching Assistant (1990-1993), The University of Arizona, "Introduction to Federal Taxation" (Undergraduate course)

Computing Experience

Application Developer and Web Engineer & Administrator (1994-present):

World Wide Web (WWW) Common Gateway Interface (CGI) server designer and engineer for Artificial Intelligence (AI) Lab (http://ai.bpa.arizona.edu) at the University of Arizona. Web services include:

- WormSpace server, http://bpaosf.bpa.arizona.edu:8000/cgi-bin/BioQuest, automatic indexing, thesaurus generation, file system organization, and WAIS indexing. (1994)
- CSQuest server, http://ai.bpa.arizona.edu/html/csquest, automatic indexing, thesaurus generation, file system organization, and Simple Web Indexing System for Humans (SWISH) indexing and server. (1995)
- Et-map server, http://ai.bpa.arizona.edu/ent, automatic indexing, thesaurus generation, Thesaurus Indexing & Propagation System (tips) server, HTML interface. (1996)
- GisSpace server, http://ai.bpa.arizona.edu/cgi-bin/tng/GisSpace, automatic indexing, multi-thesaurus generation, Thesaurus Indexing & Propagation System (tips) server, HTML interface, interface to Java applets. (1996)
- CancerSpace server, http://ai.bpa.arizona.edu/go/medical/cancerspace.html, automatic indexing, thesaurus generation, Thesaurus Indexing & Propagation System (tips) server, HTML interface, interface to Java applets. (1997-1999)
- Informedia: Contextual Search Interface, http://processc.inf.cs.cmu.edu/tng/inf, automatic phrase formation and thesaurus generation from CNN broadcasting news video. (1999-present)

System Administrator (1992-1999):

Artificial Intelligent (AI) Lab in the Department of Management Information Systems at the University of Arizona. Systems include:

- Two high-performance supercomputers: SGI Origin2000 (8 R10000 processors, 1 GB Memory) and DEC Alpha 4100 (Dual 466Mhz Alpha processors, 2 GB Memory) (1998-1999)
- Ten Unix servers: 2 Alpha servers, 4 HP servers, 3 SGI servers, and 1 Linux server (1992-1999)
- Seventeen Windows NT workstations (1996-1999)

System Programmer (1989-present):

- Supercomputing facilities: SGI/Cray Origin2000 with Irix, SGI Power Challenge Array with Irix, Convex Exemplar with HP-UX, and Thinking Machine CM-5 with Sun OS. Parallel programming using C language. (Funded by NSF/NASA/ARPA “Digital Library Initiative” project and six NCSA High-performance Computing Resources Grants) (1994-2000)
- UNIX workstations: SGI workstation with Irix, HP workstation with HP-UX, DEC Alpha workstation with OSF1 and Digital UNIX, DECStation with Ultrix, Sun workstation with Sun OS and Solaris, IBM workstation with AIX, and Linux. C, Java, Perl, shell scripts, and SAS/STAT (1991-present)
- DEC VAX/VMS. C, Pascal, Minitab, SAS/STAT, INGRES (1989-1996)

Software Engineer (1989-present):

- Parallel programming in C. Programs are used in NSF/NASA/ARPA-funded and NCI-funded projects. (1994-1999)
- C programming. Programs are used in NSF-funded, NIH-funded, and NSF/NASA/ARPA-funded projects. (1989-present)
- Java programming. (1996-present)
- Perl programming. (1997-present)
- Pascal programming. (1989-1991)
- Shell scripts (C and Bourne shell). Scripts are used in concept space generation projects and World Wide Web server development. (1994-present)

Research Interests

Intelligent information retrieval (IR), concept space generation, automatic thesaurus browsing and traversal, machine learning for IR.

Semantic interoperability for information analysis environment, Internet resource discovery, digital libraries, IR for large-scale multimedia and scientific databases.

Knowledge discovery in multimedia databases, machine learning, knowledge-base systems, knowledge management, neural networks computing.

Software engineering, parallel computing, group support systems, collaborative computing, telecommunication, distributed database systems.

Professional Associations and Activities

Member of the Institute of Electrical and Electronics Engineers (IEEE), IEEE Computer Society, the Association of Computing Machinery (ACM).

Reviewer for IEEE Computer, IEEE Expert, Journal of the American Society for Information Science (JASIS), Hawaii International Conference on System Sciences.

Recognition and Awards

Recipient of a National Center for Supercomputing Applications (NCSA) High-performance Computing Resources Grant, “High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management.” (Principal investigators: H. Chen, T. D. Ng) (1998-1999)

Recipient of a National Center for Supercomputing Applications (NCSA) High-performance Computing Resources Grant, “Parallel Computation for a Semantic Interoperability Environment.” (Principal investigators: H. Chen, T. D. Ng) (1997-1998)

Recipient of Graduate Research Assistantship, NSF/NASA/ARPA-funded “Digital Library Initiative” project (IRI9411318), Department of Management Information Systems, College of Business and Public Administration, the University of Arizona. (1994-1999)

Recipient of Graduate Research Assistantship, NSF-funded "National Collaboratory" project (IRI9211418), Department of Management Information Systems, College of Business and Public Administration, the University of Arizona. (1993-1994)

Recipient of Graduate Tuition Scholarship, Department of Management Information Systems, College of Business and Public Administration, the University of Arizona. (1993-1994)

Recipient of Graduate Research Assistantship, NIH-funded project (BRSG S07RR07002), Department of Management Information Systems, College of Business and Public Administration, the University of Arizona. (1992-1993)

Member of the Honors Center, the University of Arizona. (1990)

The University of Arizona, College of Business and Public Administration Dean's Award. (1989-1990)

The University of Arizona, College of Business and Public Administration Dean's List. (1989-1990)

Arizona Western College, Dean's List. (1987-1988)

Refereed Journal Publications

1. A. L. Houston, H. Chen, S. M. Hubbard, B. R. Schatz, **T. D. Ng**, R. R. Sewell, and K. M. Tolle, "Medical Data Mining on the Internet: Research on a Cancer Information System," *Artificial Intelligence Review*, Volume 13, Number 5/6, Pages 437-466, December 1999.
2. A. L. Houston, H. Chen, B. R. Schatz, R. R. Sewell, K. M. Tolle, T. E. Doszkocs, S. M. Hubbard, and **T. D. Ng**, "Exploring the Use of Concept Space, Category Map Techniques, and Natural Language Parsers to Improve Medical Information Retrieval," *Decision Support Systems*, Special Issue on Decision Support for Health Care in a New Information Age, 1999, forthcoming.
3. B. Zhu, M. Ramsey, **T. D. Ng**, H. Chen, and B. R. Schatz, "Creating a Large-Scale Digital Library for Georeferenced Information," *D-Lib Magazine*, Volume 5, Number 7/8, July/August, 1999, http://www.dlib.org/dlib/july99/zhu/07zhu.html.
4. B. R. Schatz, W. Mischo, T. Cole, A. Bishop, S. Harum, E. Johnson, L. Neumann, H. Chen, **T. D. Ng**, "Federated Search of Scientific Literature," *IEEE Computer*, Special Issue on Digital Libraries, Volume 32, Number 2, Pages 51-59, February, 1999.
5. H. Chen, J. Martinez, A. Kirchhoff, **T. D. Ng**, and B. R. Schatz, "Alleviating Search Uncertainty Through Concept Associations: Automatic Indexing, Co-occurrence Analysis, and Parallel Computing," *Journal of the American Society for Information Science*, Special Issue on "Management of Imprecision and Uncertainty in Information Retrieval and Database Management Systems," Volume 49, Number 3, Pages 206-216, 1998.
6. H. Chen, J. Martinez, **T. D. Ng**, and B. Schatz, "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System," *Journal of the American Society for Information Science*, Volume 48, Number 1, Pages 17-31, January, 1997.
7. H. Chen, B, R. Schatz, **T. D. Ng**, J. Martinez, A. Kirchhoff, and C. Lin, "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Special Section on Digital Libraries: Representation and Retrieval, Volume 18, Number 8, Pages 771-782, August, 1996.

8. H. Chow, H. Chen, **T. D. Ng**, P. Myrdal, and S. H. Yalkowsky, "Using Backpropagation Networks for the Estimation of Aqueous Activity Coefficients of Aromatic Organic Compounds," *Journal of Chemical Information and Computer Sciences, American Chemical Society*, Volume 35, Number 4, Pages 723-728, July/August, 1995.
9. H. Chen, and **T. D. Ng**, "An Algorithmic Approach to Concept Exploration in a Large Knowledge Network (Automatic Thesaurus Consultation): Symbolic Branch-and-bound Search vs. Connectionist Hopfield Net Activation," *Journal of the American Society for Information Science*, Volume 46, Number 5, Pages 348-369, June 1995.
10. H. Chen, K. J. Lynch, K. Basu, and **T. D. Ng**, "Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval," *IEEE Expert*, Special Series on Artificial Intelligence in Text-Based Information Systems, Volume 8, Number 2, Pages 25-34, April, 1993.

Presentations

1. Demonstration: H. Chen, B. R. Schatz, and **T. D. Ng**, "The Interspace Prototype" for the University of Illinois at Urbana-Champaign, in the *Information Management (IM) and Intelligent Collaboration & Visualization (IC&V), Principal Investigator Meeting*, Kahuku - Oahu, Hawaii, October 26-29, 1998. Sponsored by DARPA Information Technology Office (ITO).
2. Presentation and Demonstration: **T. D. Ng**, "Semantic Interoperability for Geographic Information Systems" for the University of Illinois at Urbana-Champaign Digital Library Initiative (DLI) project and the University of California at Santa Barbara DLI project, in the *NSF/ARPA/NASA Digital Library Initiative Winter 1998 All-Project Meetin*g, Berkeley, CA, January 5-6, 1998. Sponsored by NSF/ARPA/NASA and the University of California at Berkeley, http://ai.bpa.arizona.edu/tng/pub/DLI98_Berkeley/index.htm.
3. Demonstration: H. Chen, B. R. Schatz, and **T. D. Ng**, "The Interspace Prototype" for the University of Illinois at Urbana-Champaign, in the *Joint Information Collaboration & Visualization (IC&V) and Information Management (IM), Principal Investigator Meeting*, San Diego, CA, October 15-17, 1997. Sponsored by DARPA Information Technology Office (ITO).
4. Poster Session: H. Chen, T. R. Smith, and **T. D. Ng**, "GeoScience Self-organizing Map and Concept Space," in the *2nd ACM International Conference on Digital Libraries '97*, Philadelphia, PA, July 23-26, 1997.
5. Poster Session: H. Chen, B. R. Schatz, A. L. Houston, R. R. Sewell, **T. D. Ng**, and C. Lin, "Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques," in the *2nd ACM International Conference on Digital Libraries '97*, Philadelphia, PA, July 23-26, 1997.
6. Poster Session: H. Chen, B. R. Schatz, S. M. Hubbard, T E. Doszkocs, A. L. Houston, R. R. Sewell, K. M. Tolle, and **T. D. Ng**, "Medical Information Retrieval," in the *2nd ACM International Conference on Digital Libraries '97*, Philadelphia, PA, July 23-26, 1997.

Papers Under Review/Revision

1. R. V. Hauck, R. R. Sewell, **T. D. Ng**, and H. Chen, "Concept-based Searching and Browsing: A Geoscience Experiment," submitted to the *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews*, 1999.
2. K. M. Tolle, H. Chen, and **T. D. Ng**, "Improving Concept Extraction from Text Using Natural language Processing Noun Phrasing Tools: An Experiment in Medical Information Retrieval," submitted to the *Journal of the American Medical Informatics Association*, 1999.
3. A. L. Houston, H. Chen, S. M. Hubbard, B. R. Schatz, **T. D. Ng**, R. R. Sewell, and K. M. Tolle, "Health Care Information Infrastructures: A Critical Component of the NII," submitted to *Journal of the American Society for Information Science*, 1999.