Curriculum Vitae — AnHai Doan

University of Washington
Department of Computer Science & Engineering
Mail stop 352350
Seattle, WA 98195-2350

Office: (206) 616-1842
Home: (206) 634-9425
anhai@cs.washington.edu
http://www.cs.washington.edu/homes/anhai

RESEARCH INTERESTS

Databases and Artificial Intelligence, with an emphasis on applying and extending machine learning techniques to address data integration over the Internet and across enterprises. In particular: schema matching, object identification across multiple sources, schema evolution, user interaction, learning with structured data, and text mining.

EDUCATION

Ph.D	Computer Science, **University of Washington** (expected) Dissertation: Learning to Translate between Structured Representations of Data Advisors: Professors Alon Halevy & Pedro Domingos	June 2002
M.S.	Computer Science, **University of Wisconsin-Milwaukee** Dissertation: An Abstraction-based Approach to Decision-Theoretic Planning Advisor: Professor Peter Haddawy	1996
B.S.	Computer Science *(summa cum laude)*, **Kossuth Lajos University**, Hungary	1993

AWARDS AND HONORS

• Graduate School Fellowship, University of Wisconsin	1995-1996
• University Fellowship, Kossuth Lajos University	1991-1993
• America-Hungary Exchange Program Scholarship, Kossuth Lajos University	1993
• Red Diploma (equivalent to *summa cum laude*), Kossuth Lajos University	1991
• Government Scholarship for Undergraduate Studies in Hungary	1987
• Member of the six-person team representing Vietnam at the 27th Int. Math. Olympiad	1986

DISSERTATION: Learning to Translate between Structured Representations of Data

Finding semantic mappings between the schemas of two disparate data sources is a fundamental problem in many data management applications. My thesis presents LSD, a system that applies machine learning techniques to semi-automatically create such mappings. To find the mappings, LSD employs a multi-strategy learning approach: it applies multiple learners, then combines the learners' predictions using a meta-learner. As a result, LSD is extensible and can be easily customized to work on a particular application. Furthermore, the thesis describes GLUE, a system that builds on LSD to learn mappings between ontologies on the Semantic Web. The thesis also makes several contributions to the field of machine learning. To address learning problems brought about by schema and ontology mapping, the thesis presents a novel technique to classify semi-structured data, and an efficient method that employs relaxation labeling to classify interrelated entities.

RESEARCH EXPERIENCE

Research Assistant, University of Washington 2000-present

Advisors: Professors Alon Halevy and Pedro Domingos.

Ongoing research in schema matching, a critical step in many data management applications. Designed, implemented, and experimented with LSD, a system that employs and extends machine learning techniques to match the schemas of disparate data sources. Developed a novel learning method to classify semi-structured (e.g., XML) data.

Designed and experimented with GLUE, a system that builds on LSD to learn semantic mappings between ontologies on the Semantic Web. Developed an efficient learning method that employs relaxation labeling to classify interrelated ontology elements.

Research Assistant, University of Washington 1999

Advisor: Professor Alon Halevy.

Conducted research on query optimization for data integration. Developed and experimented with techniques to efficiently find the best query plans for a broad variety of plan utility classes.

Research Assistant, University of Washington 1997-1998

Advisor: Professor Steve Hanks.

Built a probabilistic AI planner that uses goal regression techniques to efficiently find the best plans in goal-oriented Markov Decision Process settings.

Research Intern, Rockwell Science Laboratory, Palo Alto, CA 1996

Mentor: Dr. Denice Draper.

Implemented and experimented with exact and approximate methods for performing inference on large Bayesian networks.

Intern, Frontier Technologies Corp., Mequon, WI 1996

Studied and designed encryption protocols for data transfer across networks. The internship resulted in a permanent job offer in the research and design division.

Research Assistant, University of Wisconsin, Milwaukee, WI 1993-1996

Advisor: Professor Peter Haddawy.

Designed, implemented, and experimented with DRIPS, a decision-theoretic AI planner that uses abstraction methods to quickly find the best plans. Applied DRIPS to clinical decision analysis. Developed theoretical frameworks for abstracting probabilistic actions and reasoning with probabilistic intervals.

Research Intern, Institute of Nuclear Research, Hungarian Academy of Sciences 1991

Developed a statistics and graphical toolkit to process, visualize, and find interesting patterns in large amount of data obtained from experiments in nuclear physics.

TEACHING AND MENTORING EXPERIENCE

Teaching Assistant, University of Washington Fall 1996

CSE 373, Data Structure and Algorithms, 70 students. Graded assignments and projects, helped students in office hours, assisted in preparing and grading exams, and gave several lectures.

Mentor, University of Washington Fall 2001
Jayant Madhavan, graduate student. Supervised a project on learning semantic mappings between ontologies.

Mentor, University of Washington Spring 2000
Leonid Tsybert, undergraduate student. Supervised an honors class project on applying decision tree techniques to the schema matching problem.

INVITED TALKS

Generic and Extensible Schema Matching with LSD.
IBM Almaden Research, San Jose, CA, July 2001.

Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach.
WatchMark Corp. (data mining for wireless services), Bellevue, WA, July 2001.

PROFESSIONAL ACTIVITIES AND SERVICES

Referee for INFORM Journal of Computing, 2001.

External referee for SIGMOD 2001, VLDB Journal 2001, WebDB 2001, WWW 2002, WISE 2001, AAAI 1996, and UAI 1995-1996.

Creator & developer of the UW online repository of benchmarks and data for schema and ontology matching.

Organized the Departmental weekly reading group on statistics and machine learning, 2000.

Volunteer, SIGMOD 1998 Conference, Seattle, WA.

Member of ACM, SIGMOD, AAAI, and IEEE.

REFERENCES

Prof. Alon Y. Halevy (advisor)
Dept. of Computer Science & Engineering
University of Washington, Box 352350
Seattle, WA 98195-2350
(206) 543-8099
alon@cs.washington.edu

Prof. Pedro M. Domingos (advisor)
Dept. of Computer Science & Engineering
University of Washington, Box 352350
Seattle, WA 98195-2350
(206) 543-4229
pedrod@cs.washington.edu

Dr. Philip A. Bernstein
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
(425) 706-2838
philbe@microsoft.com

Prof. Peter Haddawy
CSIM Program
Asian Institute of Technology
P.O. Box 4, Klong Luang
Pathumthani, 12120, Thailand
66 2 524-5705
haddawy@ait.ac.th

Dr. Steven J. Hanks
Amazon.com
Home address: 616 NW 70th Street
Seattle, WA 98117
hanks@pobox.com

Prof. Dan Suciu
Dept. of Computer Science & Engineering
University of Washington, Box 352350
Seattle, WA 98195-2350
(206) 685-1934
suciu@cs.washington.edu

PUBLICATIONS

PAPERS SUBMITTED OR IN PROGRESS

1. "Learning to Map between Ontologies on the Semantic Web," A. Doan, J. Madhavan, P. Domingos, and A. Halevy, submitted to the *World-Wide Web Conference (WWW)*, 2002.

2. "Learning Complex Mappings between Database Schemas," A. Doan, P. Domingos, and A. Halevy, to be submitted to the *Conference on Very Large Databases (VLDB)*, 2002.

PAPERS IN REFEREED JOURNALS

3. "Geometric Foundations for Interval-Based Probabilities", V. Ha, A. Doan, V. Vu, and P. Haddawy, in *Annals of Mathematics and Artificial Intelligence*, 24 (1-4), 1998.

4. "Decision-Theoretic Refinement Planning in Medical Decision Making: Management of Acute Deep Venous Thrombosis", P. Haddawy, A. Doan, and C. Kahn, in *Journal of Medical Decision Making*, 1996.

INVITED PAPERS

5. "Data Integration: A 'Killer App' for Multi-Strategy Learning", A. Doan, P. Domingos, and A. Levy, in the *Proc. of the 5th Int. Workshop on Multi-Strategy Learning (MSL)*, 2000.

PAPERS IN REFEREED CONFERENCES & WORKSHOPS

6. "Efficiently Ordering Query Plans for Data Integration," A. Doan and A. Halevy, to appear in the *Proc. of the 18th IEEE Int. Conference on Data Engineering (ICDE)*, 2002.

7. "Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach," A. Doan, P. Domingos, and A. Halevy, in the *Proc. of the ACM Conference on Management of Data (SIGMOD)*, 2001.

8. "Learning Source Descriptions for Data Integration," A. Doan, P. Domingos, and A. Levy, in the *Proc. of the Third Int. Workshop on the Web and Databases (WebDB)*, 2000.

9. "Learning Mappings between Data Schemas", A. Doan, P. Domingos, and A. Levy, in the *Proc. of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, 2000.

10. "Efficiently Ordering Query Plans for Data Integration", A. Doan and A. Levy, in the *Proc. of the IJCAI-99 Workshop on Intelligent Information Integration*,1999.

11. "Sound Abstraction of Probabilistic Actions in the Constraint Mass Assignment Framework", A. Doan and P. Haddawy, in the *Proc. of the 12th Nat. Conference on Uncertainty in AI (UAI)*, 1996.

12. "Modeling Probabilistic Actions for Practical Decision-Theoretic Planning", A. Doan, in the *Proc. of the 3rd Int. Conference on AI Planning Systems (AIPS)*, 1996.

13. "Decision-Theoretic Planning for Clinical Decision Analysis", A. Doan, P. Haddawy, and C. Kahn, in the *Proc. of the Annual AI in Medicine Spring Symposium*, 1996.

14. "Efficient Decision-Theoretic Planning: Techniques and Empirical Analysis", P. Haddawy, A. Doan, and R. Goodwin, in the *Proc. of the 11th Nat. Conference on Uncertainty in AI (UAI)*, 1995.

15. “Decision-Theoretic Refinement Planning: A New Method for Clinical Decision Analysis”, A. Doan, P. Haddawy, and C. Kahn, in the *Proc. of the 19th AMIA Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 1995.

16. “Abstracting Probabilistic Actions”, P. Haddawy and A. Doan, in the *Proc. of the 10th Nat. Conference on Uncertainty in AI (UAI)*, 1994.

OTHER PUBLICATIONS

17. “Generating Macro Operators”, A. Doan and P. Haddawy, in the *Proc. of the AAAI Spring Symposium on Extended Theories of Action Representation*, 1995.

18. “Management of Acute Deep Venous Thrombosis of the Lower Extremities (abstract)”, C. Kahn, A. Doan. and P. Haddawy, in *American Roentgen Ray Society Meeting*, 1996.

19. “An Abstraction-Based Approach to Decision-Theoretic Planning”, A. Doan, *Masters Thesis, Technical Report TR-95-12-01*, Dept. of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee.