AnHai Doan Speaks Out

on His ACM Dissertation Award, Schema Matching, Following Your Passion, Least Publishable Units, and More

by Marianne Winslett

AnHai Doan
http://www.cs.wisc.edu/~anhai/

Welcome to this installment of ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today [Summer 2007] we are talking with AnHai Doan, Assistant Professor of Computer Science at the University of Wisconsin at Madison. His research interests include databases and AI, with an emphasis on data integration, information extraction, mass collaboration, managing text and unstructured data, and Web technology. AnHai received the ACM Dissertation Award in 2003 for his thesis, entitled Learning to Map between Structured Representations of Data. *AnHai received an NSF CAREER Award in 2004, and is currently a Sloan Fellow. His PhD is from the University of Washington. So, AnHai, welcome!*

AnHai, you were the first DB person to win the ACM Dissertation Award. What was your thesis about?

That was five years ago! My thesis was about finding semantic correspondences across different representations of data. For example, if we have two relational tables, then we might want to find out that the *Address* field in one table is semantically the same as the *Location* field in another table. This problem is a fundamental component of many data management applications.

So what was different about your approach to the problem? I mean, people have been working on data integration for perhaps 20 years, so there must have been something special about your approach.

Indeed, people had been working on this problem, which is now commonly known as *schema matching*, for many years before I started working on it. They had made a lot of progress, actually. Researchers had come to understand the problem better and had developed many different solution approaches by the time I started working on schema matching.

I think one of the main contributions of my thesis work is a *multi-model architecture*. Essentially, that architecture allows you to combine many different schema matching techniques in a plug and play fashion, so that given a particular application, you can pick the right kinds of

techniques and combine them in powerful ways to develop the best solution. The other important contribution of my approach was the idea of reusing past matching efforts, using machine learning techniques. Certainly this is something that previous work did not look at.

So even researchers from the AI community hadn't tried using machine learning on data integration?

Data integration is a very broad area. Within the data integration area, people have used machine learning on many different problems. For example, during the mid 90s, it was very popular to apply machine learning techniques to develop wrappers, e.g., programs that extract structured data from web pages. The wrapper construction work first came from the AI community. But for the schema matching problem in particular, the earliest work that I am aware of that used machine learning techniques was done by Chris Clifton, back in the early 90s.

What impact do you think your thesis work has had?

For me personally, the thesis work and the award that came with it was a very big encouragement. Clearly, it gave me a big boost to do research.

For the database community as a whole, I think the award shows that the larger computer science community acknowledges that schema matching is a very important problem, and that the database community has some promising solutions to it. I think that this is the biggest impact, first and foremost. I like to think that my thesis helped contribute to people becoming more aware of the schema matching problem and starting to work on it. And now, schema matching has become a very popular problem that receives a lot of attention. Most specifically, my thesis helped people look at the multi-model composition of solutions, together with the work by Erhard Rahm and his colleagues in Germany. Now this multi-model architecture is the dominant architecture for schema matching solutions.

How did you come to do your undergraduate degree in Hungary?

I was in Vietnam when I finished high school. At that time in Vietnam, if you finished high school and you did very well, you got a scholarship to go to one of the then-Communist countries to study. Every year the Vietnamese government sent perhaps three or four hundred students to study on those scholarships. I thought that I would be going to the Soviet Union, so I was studying Russian quite a bit. Then I was very surprised to learn that I was actually going to Hungary. I asked around and I heard that someone had been sitting on my application folder up until the point where they thought that they had finished processing all the applications. Then they realized that there was one more folder left, and they just threw it over on one of the stacks, and it just happened to be the Hungarian stack. That is how I ended up in Hungary.

Were you disappointed at first?

I was definitely disappointed, but in retrospect, I think it was a lucky choice.

How is your Hungarian today?

Well, I can still understand Hungarian, but I cannot speak well anymore. But with a few months of effort, maybe I could get back to it.

You've almost finished your time as an assistant professor. Do you have any words of advice for new assistant professors?

When I finished my PhD, my ex-advisor told me that I should follow my passion. [*Speaking ironically*:] At the time, I thought that this was really a very operational piece of advice.

In retrospect, as time passes, I realize more and more how correct this advice is. For an assistant professor, I think it is very important to follow your passion, and do what you think is the right thing, something that you are very strongly interested in. There were cases when I was pursuing something and people were saying *what the heck is this?* or *how can this possibly work?* and so on. I just had to continue. So, have passion and have courage in pursuing what you are doing.

Second, I see a lot of assistant professors who get bogged down in the mode of looking for the next least publishable unit. That is a pity, because you can take some time instead to look at the broader development of the field, develop a sense about where the field is going, and decide which direction to push to have the most impact. That is also very important. So try not to get sucked too much into looking for the next least publishable unit.

If you magically had enough extra time at work to do one thing that you are not doing now, what would that be?

At work, I would like to learn more about the fundamentals of the field: what has happened in the relational database management area, what things people have tried, what failed. I want to expand my DB knowledge both in terms of systems and theory, because as I see it, as we expand to cover more nontraditional data management, we actually need to know more about what has happened at the fundamental level. It turns out that a lot of the fundamental issues in relational database management are very relevant to managing nontraditional data. I also wish I had time to do a far better job of educating my students.

If you could change one thing about yourself as a computer science researcher, what would it be?

That is a tough question. As I mentioned above, I would like to learn more about the data management field. I'd also like to learn how to communicate better.

Thank you very much for talking with me today.

Thank you, Marianne.