Rakesh Agrawal Speaks Out

on Where the Data Mining Field Is Going, Where It Came From, How to Choose Problems and Open Up New Fields, Our Responsibilities to Society as Technologists, What Industry Owes Academia, and More

by Marianne Winslett

Rakesh Agrawal
http://www.almaden.ibm.com/u/ragrawal/

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 we are in San Diego, home of the 2003 SIGMOD and PODS conferences. I have here with me Rakesh Agrawal, who is a member of the research staff at IBM Almaden Research Center. Rakesh is well known for his work on data mining, for which he has received the SIGMOD Innovations Award and the SIGKDD Innovations Award. Rakesh is an IBM Fellow and an IEEE Fellow, and his PhD is from the University of Wisconsin. So, Rakesh, welcome!

Thanks, Marianne.

Rakesh, a paper of yours with Arun Swami and Tomasz Imielinski on association rule mining just won the Test of Time Award at this year's SIGMOD conference, as the paper that has had the most impact among those that appeared in the SIGMOD conference ten years ago. Since that pioneering paper appeared, what has surprised you most about the development of the field of data mining?

When the three of us did the association rules paper, the truth is that we were a little skeptical about even sending the paper to SIGMOD. We thought the ideas were simple and the reviewers might reject the paper, saying, "there is not enough depth in the paper". What finally convinced us to send it in was the fact that the paper was solving a real problem, and the solution was a fairly general abstraction into which a lot of problems could be cast. I never imagined the different ways people have used association rules and the various extensions others have made to our paper. I read new papers on association rules and say, "Boy, I wish I could have done some of them myself!"

David DeWitt once said that Jim Gray wrote a paper on data cubes and a few years later we had 300 papers on data cubes. Have too many people rushed to jump on the data mining bandwagon, or is the situation different here?

There is always a bit of bandwagon phenomenon in research. True, there have been a lot of data mining papers, but data mining is fulfilling an extremely important need. We have a huge amount of data coming in, and the amount is increasing, and there is a desire to use this data to make important decisions. Data mining asks, "Can we increase the querying power of the data management systems? Can we ask richer questions about the data we have?" The database field has so far focused on how to *manage* the data well, but extracting the *value* from data will be more important in future, and that is what data mining can do for you.

I never imagined the different ways people have used association rules and the various extensions others have made to our paper.

Now the ideas from data mining are being applied to text mining and multi-media mining, and people are thinking of applying data mining in very interesting ways to the combination of various modalities. So I think there is going to be more interesting work in the data mining area.

So, compared to data cubes, data mining is addressing a broader class of problems?

Gray's data cube paper is a very, very nice paper; some of my own work has been inspired by this paper. Data mining is trying to address a large set of problems, and there is no dearth of the kinds of things people can do with that.

I have heard it said that association rule mining tends to produce too many rules. Has association rule mining been applied successfully in industry? Do we have success stories, or is there more work that needs to be done?

When we started doing data mining, we were concerned that we were generating too many rules, but the companies we worked with said, "this is great, this is what exactly what we want!"

Association rule mining does produce lots of rules, no doubt about it.

When we started doing data mining, we were concerned that we were generating too many rules, but the companies we worked with said, "this is great, this is what exactly what we want!" The prevailing mode of decision making was that somebody would make a hypothesis, test if the hypothesis was correct, and repeat the process. Once they had data mining tools, the decision-making process changed. Now they could use a data mining algorithm to generate all rules, and then debate which of them were valuable.

If you are running a business and you don't know what the most powerful rule is, then you will be out of business in no time …

This approach is different from the way people in AI or statistics had approached the decision-making problem. They were interested in finding the most valuable or most interesting rule. If you are running a business and you don't know what the most powerful rule is, then you will be out of business in no time. The valuable rules fall in the middle, where you don't have the strongest support or strongest confidence for a rule. People feel if they can extract those rules, they can apply business judgment on them, doing the pruning themselves. Lots of successful applications for association rules fall in this category.

Having said that, we need work to bring in some notion of "here is my idea of what is interesting," and pruning the generated rules based on that input. Then the user could hone in and say, "these are the rules that I think are interesting; give me more rules like these." We would eventually like to be able to handle the following request: "Here's my data; tell me something that is interesting, something that is of value to me." But that work remains to be done.

Is there a notion of similarity of rules, based on some sort of business concepts or statistical concepts or ...?

The past work has mostly used a statistical notion of what is interesting, what in some sense is unexpected. But the unexpected might not be what is valuable. So, we have to introduce a notion of what is interesting to a particular user or what is valuable to a particular business, and use that notion as a similarity measure to do the focusing.

Are there fundamental concepts in the field of data mining, or is it just a collection of application-specific techniques?

I'm a database person, so my view of data mining has been that it is essentially a richer form of querying. We want to be able to ask richer questions than we could conveniently ask earlier.

Codd's relational algebra had a few basic operations, which did not capture every computation but did capture a large class of them. We were trying to do something similar. If we could figure out a set of basic data mining operations, and if these operations were composable, then the new applications could be built by combining them. That is how we thought about operations like association rules, sequential patterns, classification, and clustering in the Quest project. Some of these we invented ourselves, and some of these we borrowed and adapted for our use from statistics and AI literature.

So we have the operations now. We have fast, scalable algorithms for the operations and we can parallelize them. The part still missing is a nice algebra for composing the operations. We should also be able to combine data mining operations with traditional database operations. So those are the central concepts around which we can build data mining. Data mining is still in its infancy; a lot more needs to happen.

With the composition part, is it that we don't have a clue what the composition operators should be, or it is just that we have not gotten around to writing them down yet?

The applications I know of in industry do very simple compositions, involving a few operations. Currently this composition is done on a fairly ad hoc basis, at the user interface level.

Composition is not yet automatic and has not yet been pushed down into the database system. People have thought about it, but it is not yet a solved problem.

Would your work have had more impact if you had gone into academia?

Whatever I have done, I have been inspired by seeing that it can be used or applied, and I have tried to stay honest to that

Yes and no.

A big advantage of being in academia is that you have students who challenge you and keep you on your toes and force you to think. You can also have lots of students working on different kinds of problems. Sometimes I have interest in a certain topic but I just don't have time to work on it, because I don't have graduate students to help me develop the ideas.

On the other hand, being in industry, particularly in IBM, the advantage is that once you have done something interesting, the company can invest quite a lot in developing and commercializing that particular technology. Doing Intelligent Miner in a university would have been hard. So it cuts both ways; there are pluses and minuses of being in industry.

If we are solving a general problem, then people will find applications for it. You don't have to go and tell them, "this is how it is to be used."

In your career, several times you have published papers that attracted many followers to a new area: first in association rule mining, then in high dimensional clustering, and also in scalable classification algorithms. It is not often that we see researchers who have demonstrated an ability to repeatedly open up new areas for research. What is the secret of your success?

I guess I am a lucky person! I am generally curious and like to explore different ideas. Whatever I have done, I have been inspired by seeing that it can be used or applied, and I have tried to stay honest to that. So most of the data mining work grew out of talking to real customers, understanding their problems, and thinking whether we can abstract these problems. That is, can we create abstractions around which we can build software? If these abstractions are useful to us, then perhaps other people will find them useful too.

But wait--- if you were in academia, how could you have talked to IBM customers and gotten all these ideas?

You don't have to go to customers to learn what is important. For example, some of my recent privacy work grew out of talking to my brother, who is a doctor, not an IBM customer, and telling him, "Look, I am appalled with what is happening," and learning what they do in medicine.

Relating back to your first question of the interview ("Did you find it surprising?"), the growth of data mining was not surprising in retrospective because if we solve a general problem, then people find applications for it. You don't have to tell them, "this is how it is to be used." We thought data mining would be used for retailing, but it is getting used in all kinds of other applications. The important part is that when we do abstraction, we stay honest to the problem.

I noticed that you mentioned curiosity as an underlying factor in your success. We were just talking about Jim Gray opening up a new area of research in data cubes, and he is certainly insatiably curious, so maybe that is the common thread with people who open up new areas.

There has been a public outcry over the implications of data mining as envisaged in the U.S. government's proposal for catching terrorists via Total Information Awareness (TIA). Do you think that the new work on privacy preserving data mining can solve these problems? And what about things like cameras in public areas that can identify individuals by the way they walk?

> If we are creating powerful technology, it is going to have social implications. If we believe that … we have to be responsible as technologists in whatever we create, then social problems … will be addressed much better [in our creations].

This is something which is dear to my heart, and I will try to answer it more broadly than just as the TIA issue. TIA has crystallized the debate, but the debate was there before TIA.

The fundamental issue is the following: as technologists, are we responsible for the technology we create? Over the years, I have become convinced that we cannot distance ourselves from the implications of our creations. We have got to understand that if we create powerful technology, it is going to have social implications. We have to factor in those implications, and understand the potential negatives. If we think hard about these issues, we will put in technology safeguards to avoid problems down the road.

Coming back to privacy preserving data mining, a person who influenced my thinking is Dr. Ann Cavoukian, the privacy commissioner of Ontario in Canada. She wrote a paper in 1998 in which she showed how the current uses of data mining were in conflict with a number of principles of the fair information practice. If we had her insight, we would have tried to build the concept of preserving privacy into the mining algorithms right from the beginning. Data mining is about working at the aggregate level, and privacy is at the individual record level. For a lot of data mining tasks, you do not need access to accurate individual records to build good models. It is a question of the mindset: if we believe that we have to be responsible as technologists in whatever we create, then social problems like this will be addressed much better.

> We have no choice--- we have to own the responsibility, we have to consider the societal impacts of what we create.

At the recent Almaden Institute on Privacy that you organized, I was surprised to hear highly emotional comments about privacy from members of the database community. Are we ready to tackle, on a technical level, the challenges to individuals' privacy which in some sense we ourselves have created, or is there more work needed before we can move from emotion to reason?

We have no choice. If you think about any complex problem, the solution is a combination of technology, the laws, societal norms, and the pull of the market. By inventing technology and changing what is technically feasible, we can change the mix of these ingredients. You can eliminate the need for certain regulations, and in the process you can improve the overall quality

of the solution. That is what we need to do as technologists: create the technology, and tell the policy makers and legislators, "Look, you don't need those laws here, because by enacting them you might do more harm than good." An example is the HIPAA (Health Insurance Portability and Accountability Act) legislation. At the Almaden Institute on Privacy, Dr. Harry Guess (of Merck) stated that HIPAA was hindering epidemiologic research. The lawmakers did not understand the implications of this particular law, and did not understand that by using

I think a lot of industry has had a free ride many times… I am not sure whether industry has responsibly repaid what they have gained from academia.

technology such as privacy preserving data mining, you eliminate the need for the kind of restrictions put into HIPAA. This is an example of where by creating appropriate technologies, you can make unnecessary some laws that can do more damage than good in the long run. So we have no choice--- we have to own the responsibility, we have to consider the societal impacts of what we create.

Do you think that industry should fund more research in academia?

Absolutely. I think a lot of industry has had a free ride many times. I look at David DeWitt and other professors, and I see that industry has benefited tremendously from their research, and from the students they have produced. I am not sure whether industry has responsibly repaid what they have gained from academia.

Do you think industry does not recognize the contributions from academia? Or does industry recognize the contribution intellectually, but cannot put the dollars into it due to pressure to improve the bottom line?

I don't know. I hope the level of responsibility in businesses will---should---increase. It is clear that the contributions made by academic research in databases are phenomenal, but I don't think industry has reciprocated adequately. I am really worried when I look at the number of research labs that are still there and doing research in databases and---

It's shrinking.

Yes, exactly, and that is a very unhealthy situation. The database industry must make sure the research continues and is nurtured and advances.

You are an officer of SIGKDD, which has grown to be the twelfth largest SIG in ACM in just a few years of existence. The SIGMOD organization is the fifth largest SIG, but on the other hand, the KDD conference is now much larger than the SIGMOD/PODS conference. Has KDD's rapid growth surprised you?

Yes! When we collocated KDD with VLDB in New York in 1998, we thought we would get a boost in attendance because of VLDB. But surprisingly for a diehard database person like me, KDD ended up having a much larger registration than VLDB did.

Then again, there are KDD attendees with different kinds of backgrounds, like statistics, machine learning, databases, performance, and people who are interested in discovering rich patterns. The

thinking is that data mining can solve some very, very interesting problems, and that is the reason there is such a large interest in data mining.

I understand that your daughter just finished her undergraduate degree in computer science. We hear a lot nowadays about the factors that lead young women to choose or not to choose computer science as their field. Do you have any suggestions for readers who might want to encourage their own daughters to follow in their footsteps in terms of their choice of field?

My suggestion would be to show the applications, the usefulness of sciences, to our kids. The sense I get from talking to kids is that they are not very enthusiastic about abstract ideas. But if they can see the applications and how technology can really be useful--- that besides making money, technology can help a lot of developing countries or it can help with big societal problems--- if they can make that important connection and they can see the meaningful contributions they can make, then that is going to be much more valuable to them than saying,

When kids think about medicine or law, they can relate it more directly to society, whereas when we talk about engineering or science, sometimes it comes across as too abstract.

"here is the abstract technology you are going to work on." I think that is the biggest problem with recruiting kids to computer science. When kids think about medicine or law, they can relate it more directly to society, whereas when we talk about engineering or science, sometimes it comes across as too abstract.

India has become very hot in information technology. If you were finishing your undergraduate degree in computer science in India today, what would your next step be? Would you still head off to Wisconsin?

Yes, because by heading off to Wisconsin I got a great advisor (David DeWitt) who taught me what it means to do research. I think India is an interesting case, where there is a very strong pool of technical talent but the opportunities for doing cutting-edge research are still limited. As this situation changes, which is happening, you will see a lot more people coming here, then going back, maybe working there for some time, and then coming here for a sabbatical.

Are you envisaging the cutting edge activities taking place at the universities, or at company labs in India?

Both. IIT Bombay has a very strong database department, one that has a chance of becoming a world class department. Other IITs, Indian Institute of Sciences, and some other institutions also have very strong departments. So clearly they can provide the nucleus from the academic side. But the good research happens when academia and industry come together. Before I came to Wisconsin I worked for some time in India, and I talked to people in industry. There are phenomenal problems that can be attacked with very specific solutions designed for the circumstances they are in. Industry has a chance to invest in solutions to those problems in partnership with the academia.

There are phenomenal problems [in India] that can be attacked with very specific solutions … [Indian] industry has a chance to invest in solutions to those problems in partnership with the academia.

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

It has become such a cliché these days to talk about medicine, but that's something I would like to think more about: how we can apply information technology to problems in medicine?

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

I would love to have more interaction with students, and a much richer interaction with universities.

Almaden is a great place, but the industrial labs tend not have as rich an environment as is present in universities, in terms of variety of topics being pursued. The other thing we lack is graduate students. So if I could, I would love to have more interaction with students, and a much richer interaction with universities.

Did you get the students for summer interns, though?

Yes, we do every summer, but ...

So maybe that is the medium of connection.

Yes, but they come for just two or three months. We have had students who stayed longer, but we try to avoid that now. We want students to go back to their universities to complete their PhD work. But that is one thing I would like to do, get more connected with universities than I am at this stage.

Thank you very much.

Thanks, Marianne.