Ricardo Baeza-Yates Speaks Out
on CS Research in Latin America, His Multi-continent Commute for Yahoo!, How to Get Real Data in Academia, and Web Mining

by Marianne Winslett



Ricardo Baeza-Yates
http://www.dcc.uchile.cl/~rbaeza/

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 at the ICDE 2007 conference in Istanbul. I have here with me Ricardo Baeza-Yates, who is the vice president of Yahoo! Research in Europe and Latin America. Before that, he was a professor of computer science at the University of Chile for many years. His research interests include information retrieval, databases, algorithms, and user interfaces, and he is a co-author of one of the most widely used books on information retrieval. Ric's PhD is from the University of Waterloo. So, Ric, welcome!*

You have moved from an academic environment in Chile to an industrial lab in Barcelona---isn't that a double culture shock?

Not really. Barcelona is part of Spain, and Spain was the original power that conquered Chile, so I think culturally it is very similar. People do speak a different language in Barcelona, Catalan, which I can understand but I haven't mastered. The academic to industry transition was not so difficult, because we were already doing almost the same research as we are doing now at Yahoo!.

I have heard of telecommuting, but not between continents*. How do you handle the extreme distances between your Yahoo! outposts?*

I travel about three or four times every year to Chile, and I am there for around two months each year. Almost all the people we have there are people who have worked with me before, so I really can work with them remotely. Currently, it is a nice challenge to have a real distributed lab, because it does not matter from where you are connected.

So you view your part of Yahoo! as a single lab that is distributed, rather than separate ones on each continent?

Right, I view it as a single lab. In some projects we have only one location, but in other projects we have both locations working on it.

Will you grow in both locations?

Yes, I think so. That is my goal.

What is it like to do IR and DB research in South America?

When I finished my PhD, I wanted to go back and try to make a difference in my country. There was a small group of PhDs trying to build a good computer science department, and I helped to contribute to that. If you look at South America, after Brazil, Chile is the main research power there. In Chile we cannot compare with Brazil in the size and quantity of computer science research, but I think we can try to compare in quality. And now we have a PhD program; we have students coming from most neighboring countries, and even Mexico, so I think we are doing well.

It is not easy because we have to find many sources of funding, because we don't have big funding agencies. But five years ago, the Chilean government started a big program where you can receive a reasonable size of money per year. Then I started the Center for Web Research, which I think is well known now in the world. But of course, everything is more difficult than in the developed countries.

Does most of your money for your university lab come from industry, then?

No, most of the money comes from public funding from the government. Industry doesn't do research in Chile, and even in some developed countries, like Spain, industry doesn't do too much research.

Do you think that is likely to change?

I don't know. I would like that to change. There is a dilemma between buying technology and producing technology. In some fields it is really very expensive to produce technology. In our field, where it is basically ideas and software, it is not so expensive. But sometimes politicians don't know that potential.

What do you see as the future of database research in Latin America?

Right now I think it is confined to what I mentioned earlier. Brazil has a long tradition of database research, and many groups in that area. They have a very strong program for PhDs inside the country, so now they have more than a thousand researchers in computer science. They have a very strong database conference (SBBD), which is international.

In Chile we are trying to build a very good group. We have a small group at the Center for Web Research, people that studied in, say, Toronto, in the US, and so on. In the rest of South America, the problem is not that there are no good people, but that there are no opportunities; so they usually leave and work elsewhere. If you consider all of Latin America, Mexico is the second power in computer science after Brazil. They are doing well, and they also have a very nice tradition in database research. And the next ICDE will be in Cancun!

Yahoo!'s web site says your TodoCL ('All of Chile') search engine had 2 million users a month, "a rich source of data for Baeza-Yates on how people use a search engine. 'But at Yahoo!, I'll have access to far more data,' Baeza-Yates says. 'You go from gigabytes to terabytes of data.' " While I am happy for you, I see this as a real problem for academia. How can academia retain its leading database researchers? Aren't they all going to be sucked into industry by the siren song of unlimited access to real data?

That's a possibility, but I don't think that will happen. Not all people want to work in industry. For me, it was a very tough decision to move to industry, but I don't regret it now. It is much better than I expected.

Let me tell you about the example of TodoCL. I built a search engine that had a lot of people using it. I did that with very low funding. It was not easy, but having a search engine basically means having a few PCs in a data center, with good software that grinds along. From that search engine we got data that no one else had, apart from the big search engines. That data allowed us to do things that other people couldn't do.

So I don't really believe that academics cannot get access to real data. I believe that academics can be more entrepreneurial in their work, by setting up some services that become used a lot. The main examples in our field are Citeseer and DBLP, which are services that a lot of people use. And DBLP has basically one person behind it, Michael Ley! So I think that if you have the leadership, and you have the will to do it, you can create interesting services and gather a lot of data on how people use them.

My summary of what you just said is that academicians should build their own interesting little programs and collect their own data. Do people still use TodoCL today, or is there a Google version that is specific to Chile that has become popular?

Google was popular even when I started. People use both Google and TodoCL because they give different results. Vertical search engines have the advantage that you can crawl the country much deeper, have more pages, and also have local rankings. You get different search results using Google, Yahoo!, and TodoCL, and all of them are good.

And now the reverse question: won't academic research become increasingly irrelevant for the web, since researchers outside of industrial labs have little to no access to large-scale information about how people behave on the web?

I think I answered that partially. At Yahoo!, we want to do open research, to share our results. We have been sharing data with some people, such as through a special Yahoo! Research Alliance program where you can get access to data. So, we want to have the help of the academic community, and we want to contribute back to the academic community. We don't believe in "black boxes" where you don't know the level of your impact. You have to go to conferences to know where you are. We also don't believe that we have all the answers, that we know all the trends.

I am not familiar with Yahoo!'s data sharing program, but I know the one at Microsoft is very small. Maybe 10-20 researchers get access to Microsoft's Live Labs data. What's the scale of the data sharing program at Yahoo!?

I think right now it is a similar size, maybe a bit larger. But we are trying to increase that.

You have said that the goal of your new Yahoo! research lab is "to shape the future of the internet". Aren't a lot of other people trying to do the same thing right now?

I think that is a goal of all of Yahoo! Research, not only my lab. Yes, Google and Microsoft and other companies are trying to shape the future of the internet, but I think the potential for the internet is so large that it has space for everyone. Even in the market space, there is space for everyone.

What do you think of the current confluence of IR and database research? Did you see it coming?

I think I saw it coming. I did work on structured IR before XML became popular. I have one of the seminal papers on that topic, and that was written in 1995, which was early with respect to the web. We are at a point of confluence on the integration of databases and IR. We will be able to understand each other better if we do more of what I did yesterday at ICDE, i.e., having invited talks by IR people in database conferences, and the other way around. We have different views of the same problem, and both are valid. Today, there is much more unstructured information than structured information. On the other hand, with structured information you know much more, so you can do more complex things. There are both pros and cons on every side.

How should we rank results retrieved from different kinds of sources, such as blogs, shopping centers, structured records, images, and ordinary web pages?

I think that is one of the main open problems in the web: how to do good ranking, especially when you have different sources of data and evidence of quality---such as how people use it, how people created it, how people comment on it, and so on.

For example, you mentioned blogs. A blog is usually written by one person, and commented on by other people. One problem behind the Web 2.0 is how to rank *people* rather than data; if we can know that the author is good and you can trust his or her opinion, we can do better ranking for every kind of data. In ranking, we are working to be able to combine information on people, data, and usage. We call that *community systems*, and at Yahoo! that work is being led by a well known person in the database community, Raghu Ramakrishnan, who was at the University of Wisconsin before coming to Yahoo!.

How will you come up with a ranking system that can't be fooled?

That is very tough. I think today we have to live with a lot of spamming on the web: content spam, link spam, usage spam. We have some answers, but this is a permanent fight. We improve our techniques to detect spammers, and they improve their techniques to deceive us. I hope eventually we'll end with something closer to a semantic web, where we know whether we can or cannot trust a source of information.

At Yahoo!, you have been seeing scale-up issues for ad hoc queries on thousands of processors, even though the workload is almost embarrassingly parallel. What is the cause of the scale-up problems?

The query processing part is really a parallel problem, so that's not an issue. The indexing does not parallelize well. You either subdivide the document collection into pieces with separate indexes, or you build one single index and subdivide it, which would take a lot of time to build and maintain. So you have to use some kind of divide and conquer approach in the collection.

But still that means that you need to have very large memory caches just to hold the index, and that is not necessarily only a parallel problem.

What kind of solution do you foresee for that?

I see a real distributed search engine where you have collections divided by say, language, culture, geography. You can also use the network topology, and the query distribution of the region, and so on. You can use a lot of levels of caching, to try to come up with solutions that will amortize the network latency on the internet.

From your experience as a member of the Chilean Academy of Sciences, can you tell us how computer science is viewed by scientists from other disciplines?

I can tell you how other disciplines view us in Chile. I think that view is shared by many places.

Sometimes people believe that computer science is more like a technology that you can use, more like a tool. They don't believe it is a science; they think that it is closer to engineering, but not so important. They think that you can buy a computer, use it, and that is it; they don't see the complexity behind the technology that created that computer and its software. The few who understand us may be people from electrical engineering and mathematics, who work in areas that are closer to our problems.

I think that in time, people will start to recognize that computer science is at the interface of engineering and science. We are not a pure science like physics, and we are not pure engineering like, say, mechanical engineering. That's what makes computer science problems so difficult.

I wrote a long manifesto on this topic in Spanish, and included a joke that says, "The name *computer science* has two problems. The first one is that the name includes *computer*, which is a machine, an object, not an abstract concept. The second problem is that the name includes *science*; if we are a science, we don't need to say that we are a science." It is as though we are too young, and we have to use this name to be sure of our personality or our ideas.

Like library science *or* political science.

Right, but in this case we are closer to *car science*. With *political science*, at least *political* is not an object, it is an abstraction. The equivalent name in our field would be *computing science*, and some departments do use that name. Even within the field there is disagreement; some departments are called *Computing Science and Software Engineering*. I thought that software engineering was part of computer science, but maybe some people don't.

I also like the quote that you had in that article that said, "In theory there is no difference between theory and practice, but in practice, there is a difference between them."

That quote is from Donald Knuth in his 1999 invited talk at the IFIP World Congress. Many times theoreticians don't turn their ideas into programs, so implementing their ideas looks easy to them. But when you turn a theoretical idea into a program, you may find that the theory was correct but the constant factor in the running time analysis is, say, 10,000, and the approach is not competitive. I think the best theories are inspired by practice and the best practice is inspired by theory, and that is another quote by Don Knuth.

One of my former students had a T-shirt that said, "0 and 1: How Hard Can It Be?" In that vein, please tell us about Al-Khorezm, el matemático olvidado.

That means 'the forgotten mathematician.' He was a mathematician from the Caliphate of Baghdad in the Arabic empire. He was so famous that he was called Al-Khorezm, which means the person from Khorezm, the city where he was born. This person is so important to us because he invented the first algorithm, in some sense. (There were previous algorithms by Euclid, Eratosthenes, and other Greeks, but they did not realize what they were doing.) Al-Khorezm was really thinking in logical steps, like a computer algorithm.

Al-Khorezm did so many things. He brought the concept of zero and decimal notation from India. Before that, when there was a zero people wrote nothing, so it was very hard to recognize that the "nothing" was there. Al-Khorezm also wrote the first book on algebra; his book was later translated in Toledo and then reached the rest of Europe. He was the first person who tried to write mathematics for everybody. Al-Khorezm also collected all the works of the Greeks. Thanks to him, those works were not lost in the Middle Ages in Europe.

I think we owe Al-Khorezm almost everything, starting with a very important word in computer science, *algorithm*. The word *Al-Khorezm* is the root of *algorithm* and *logarithm*.

I understand that maps are your hobby. What does that mean?

I like historical maps, so I collect old maps. I like geography, so I like to know places, I like to go to new countries, to meet people, to learn. I like to learn every time I am researching, and not only in computer science. If I see something new in food, I like to try it. I am not afraid of learning new things. Maybe computer science research is one way to realize my hobby, as I have the opportunity to meet many different people, know many different cultures, and visit many different countries. I like what I do. I am thankful for that.

Do you have any words of advice for fledgling or midcareer database researchers or practitioners?

I am not a real database researcher, but I think that there are many interesting problems in the web. From the web we can capture a lot of different kinds of relationships: between content, between structure, between links, between usages of the data. I am sure there is a large potential to do data mining there. So data mining should be one of the main research topics today, especially web mining. And that implies new interesting fields like graph mining, where I think a lot of things can be done. For example, how can we scale up graph mining techniques to mine a graph of one billion nodes? I think that is a very interesting problem. We can use mining to find interesting patterns, things that may imply, say, a new service or something else that people want, or maybe detecting spam. There are many possible applications there, so I think the web has a lot of room for new research problems. That is why the focus of the Barcelona/Santiago lab is Web Mining, which in the end is the Wisdom of the Crowds.

Among all your past research, what is your favorite piece of work?

It is difficult to pick a single one, but I think I have to mention my book with Berthier Ribeiro-Neto, *Modern Information Retrieval,* because it is used all over the world. It is has been translated to two other languages. I have been surprised that in many places they know it, and even in small places they use it. It is really amazing how you can reach so many people.

On the research side, I think my initial work on algorithms started a couple of different subfields. I wrote a paper about how to use the parallelism at the bit level in the CPU to speed up string matching with possible mismatches ("A New Approach to Text Searching", *Communications of the ACM,* October 1992, with G. H. Gonnet). One of my first papers was about a cow trying to cross a fence to go to the other side to eat more grass ("Searching in the Plane", *Information and Computation,* October 1993, with J. Culberson and G. Rawlins). At the time we did the research (1987), we didn't know that we were working on one of the first online algorithms for computational geometry. That paper inspired a lot of work on how to find things without having the whole picture, without having the map, so to speak. I think that is a very interesting problem. Those two papers are cited by hundreds of other papers.

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

I think it would be to stay more with the people I love. I sometimes feel that I don't have enough time to be with them and also to do the work I like to do. It is very hard when you do work that you like, because it takes time from other things.

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

That is a very hard question. I think, if I could change something, I would like to be better at words. I am not a person who talks too much.

But when you write, you write very eloquently.

Yes, but I am talking about *talking*. Not about *writing*. For me it is much easier to write or to listen than to talk. I think I would like to improve my oratory skills. Some people can stand in front of others and talk about almost anything without having to think about it, and I would like to have that ability.

Thank you very much for talking with us today.

Thank you for the invitation.