

Refining the Search Engine

*The vast amount of information on the Internet is growing every day
-- it's enough to gag a Google search. Researcher Ramesh Jain
offers up new strategies for information retrieval.*

Researcher, entrepreneur and teacher Ramesh Jain is currently professor of computer science at the Georgia Institute of Technology. His previous appointments include similar positions at the University of Michigan, Ann Arbor, and the University of California, San Diego. He has founded three companies (PRAJA, Virage, and ImageWare), and his numerous books and articles include "Machine Vision," a textbook used at several universities. He serves on the editorial boards of several magazines in multimedia, business and image and vision processing.

UBIQUITY: Let's talk today about Web searching strategies. Do you think that search technology is now in an evolutionary stage, or is it entering stage better thought of as revolutionary?

JAIN: I think both processes are going to work together: there will be an evolutionary process but I don't think it's going to take us to the solution that we are looking for. There will also be some revolutionary things coming down the line, but in my opinion they're not going to happen in the next few years, because people are not frustrated enough with the current approaches. For example, today I'll use Google and complain about it, but it has not become so frustrating for me that I'd say, "Enough of this, I'm going to do something else!" And when I say this about myself, as someone who's already thinking about search techniques, then the same point could be made even more strongly for the common person, for whom searching has not become all that frustrating yet. But it is going to be soon in my opinion.

UBIQUITY: What do you think are the main characteristics of this frustration?

JAIN: A variety of different things, of course. Number one is that the amount of data is continuously increasing -- the amount of data that Google throws up at me is increasing really fast. Now their ranking algorithms work in very well-defined cases, so that for example if I am looking for a restaurant in Atlanta with a certain kind of food, my Web

search is going to work extremely well. But if I am looking for so-called "lifestyle" restaurants, it's not going to work that well. So if the problem is very well-defined, then the answer will be very well-defined, whereas ill-defined, unstructured problems are of course the hardest to solve.

UBIQUITY: Besides the structure factor (or lack of structure factor), isn't there also an anticipation factor -- where the search problem is compounded by the system failure to anticipate certain kinds of requests? If you anticipate that you are going to need something -- data or objects or old clothes -- then you think about them and you classify them.

JAIN: That's right, so in fact if I understand you right, search engines like Google try to find out what the user's intents are, and in some cases the system is doing a good job; for example, on Google I can put in a complete address and it immediately understands that it is an address. And it will take me to Mapquest or a map of that particular area. But one of the examples that I like to use is this: what about if I am trying to understand whether President Bush's popularity is increasing or decreasing; what do I do? There is no way I can find out this information.

UBIQUITY: And what would be the nature of the project that would give an answer to that?

JAIN: Ah. If I had an answer I would have implemented it. But I do have some speculations on that topic, along the following lines. Current search engines like Google do not give me a "steering wheel" for searching the Internet (the term steering wheel was used by William Woods in one of his articles). The search engines get faster and faster, but they're not giving me any control mechanism. The only control mechanism, which is also a stateless control mechanism, asks the searcher to put in keywords, and if I put in keywords I get this huge monstrous list. I have no idea how to refine this list. The only way is to come up with a completely new keyword list. I also don't know what to do with the 8 million results that Google threw at me. So when I am trying to come up with those keywords, I don't know really where I am. That means I cannot control that list very easily because I don't have a holistic picture of that list. That's very important. When I get these results, how do I get some kind of holistic representation of what these results are, how they are distributed among different dimensions.

UBIQUITY: What would that kind of holistic representation be like?

JAIN: Two common dimensions that I find very useful in many general applications are time and space. If I can be shown how the items are distributed in time and space, I can start controlling what I want to see over this time period or what I want to see in that space.

UBIQUITY: How would you compare the search on Google or other search mechanisms with the process of searching for food in a supermarket?

JAIN: That's a very good analogy. If I went to a Publix, and I wanted a tube of toothpaste, but could only see two tubes, I could ask for assistance -- and that's how the current Google is. But when I go to a Publix I have a holistic picture -- I have a complete distribution of the same things and I can search by going through those things and finding exactly what I require. Google does not give me an option like that. With Google, you have to play a game of 20 Questions -- you ask a question and it gives you an answer. Since you asked me about supermarkets, let me give you a slightly different example, still related to shopping. In some of the old countries -- such as India where I grew up -- when you go to a shop or grocery store you are not allowed to access items in the shop directly. Instead, there is a person outside, and you say to the person, "This is what I want." Say I go to buy clothes. I say "I want fur." He will say "What kind of fur?" I would say, "This is my size and I am looking for a farmer's fur." He will bring three or four furs to you, and you look at those and say, "I don't like this, I don't like that, but I like this one." He will then bring four or five more that he thinks you will like, and this process continues until you find something that you really like. This takes place in shopping all over the world. It is not the free shopping mall that you have here in the US and many other countries, where you can enter and you can touch everything, play with everything, you can do all kinds of things. Here, the responsibility is yours to browse through and search. There, they ask you questions -- and based on those questions they give you things.

UBIQUITY: Sounds like fun actually.

JAIN: Well it can be fun and it can be frustrating. The main reason they did this is so nothing was stolen there. The fact that the nature of shopping is changing even in old

countries and is becoming more like the model of shopping malls suggests that people prefer that mode of shopping. The same thing will happen to information markets.

UBIQUITY: What do you think of online shopping experiences -- Amazon or Macy's or whatever?

JAIN: Well, they're good for people who know reasonably well what they want, and they are a lot more orientated towards branded things. I am one of those people who buy quite a few things on the Internet but my wife will never touch it because she wants to feel the merchandise, she wants to do the comparison -- and right now the comparison is not that easy. People are trying to make the comparison easy but it's still hard.

UBIQUITY: Will your daughter and granddaughter do it?

JAIN: My daughter does it a little bit. Will my granddaughter do it? I think by that time the technology will be advanced. If it's not I will be disappointed.

UBIQUITY: So what are the next steps?

JAIN: There are all kinds of people studying this problem, including ones coming from audio or video perspectives. So far searching has been limited to text but very soon it's going to involve a lot more audio and video. Cameras now are absolutely everywhere -- still cameras and phone cell cameras -- so the problem's become very interesting: how do you combine text, audio and video? In fact, if you look at Yahoo and Google and now MSN also, they are all trying to put together all of these techniques. They are trying to put together media search and image search. Right now these things are, as to be expected, in very early stages, but they are putting a lot of effort in those directions. Now who will actually do it? I don't know, but I believe that none of these big companies will do it, because they are heavily invested in text and in the current technology. That's why in order to beat Google, another big company like Google will rise up from the work of some young people who will come up with these things.

UBIQUITY: Presumably from your group at Georgia Tech, right?

JAIN: That would be wonderful.

UBIQUITY: In your recent work you seem to clearly prefer the word "exploration" to "querying"; why is that?

JAIN: In fact, that is very true, and I am excited to use that term in my new white paper, in which I've started arguing that we will soon be prospecting for data and information, and exploration will become part of prospecting.

UBIQUITY: How does the use of sensors add to the dimensions of the search problem?

JAIN: That's a very interesting development. I believe Yahoo search relies mostly on the *name* of the file for the image -- so that if you type in the word "horses," then if the file name also contains the word "horses" the search will pull in the particular thing. But Google analyzes the text from where the picture is also coming. Google possibly goes slightly beyond that because I thought that all the files that Yahoo brings in at that particular time that I put in the keywords. Google doesn't necessarily have that and, because of that, Google's results are sometimes not as good as Yahoo's. Why does that happen? It's very simple. What I am saying is that they are trying to configure the associated text with the name of the file -- or you take the speech component in the audio and video and somehow transcribe that, then apply the text with this technique into that.

That's what they are trying to do and that's not bad, and that will buy you some time. You can make some progress with that but soon you find out that if that was the only case people would not rely on audio and video so much. Because text alone doesn't capture the emotion and they don't capture the experience we get from all kinds of other things that you get when you are listening to something or when you are watching something and you are seeing all the reactions there.

So how do you get that information? Text is based on well-defined language. We use an alphabet to form words and use those words to express ideas to form the text. Without words, there is no language. We don't have anything equivalent to the alphabet and words in audio, video, images or any other sensor. So we have to develop those things first; there are some efforts like MPEG and things like that, but they are not yet advanced enough. And how will they advance? That's where the interesting results really lie. I don't have the answer to that yet.

UBIQUITY: So some day a system will tell me what's wrong with my grass and why it has yellow spots?

JAIN: Well that's the easy part, they will do that today. There are systems that you can start implementing for these kinds of things. But you'll have a problem if you want to ask whether your lawn is really in excellent condition or if you want to say "I want to see pictures of a nice beach today." The question could mean, "Is the beach beautiful?" or could mean "Is the weather there good today?" or could mean "Are the people there interesting?" The question could mean any or all of those things, and there is no specific visual language right now to distinguish those kinds of meanings in images and video.

UBIQUITY: Would I be wrong in saying that the essence of what you're doing is trying to get beyond language?

JAIN: That's very correct. I think one of my favorite books, which I love and quote a lot, is a famous book by S.I. Hayakawa, "Languages in Thought and Action," in which he makes the point that it's amazing how some arbitrary noises and some scribbles on paper started to make meaning to us. It's amazing that we all agree that some noise is going to be presenting some particular thing or some particular concept. And similarly it's amazing that we agree that such-and-such particular scribbles are going to be representing this and that real thing. Another person who had great insights on these issues is Carl Popper. Popper started talking about word one, word two, and word three. Where word one is the real word, word two is what concepts and what mechanisms you learn and have in your head, and word three is the model you build using word two about word one. And those things become very exciting because what we don't see -- which is word two (what we have in our head) -- is a lot more complex and sophisticated than our language allows. And that's how we sometimes fail to find the words to represent what we want to say. Language allows us to represent some of the things that become a lot more explicit and a lot clearer. Language is a knowledge representation language.

UBIQUITY: I'm interested in knowing whether you think that your personal history -- coming from India and speaking different languages -- has influenced much of your thinking on these topics.

JAIN: Oh boy. Now you are asking a difficult question because all of us are products of what we have experienced, so I am sure that it has influenced me. But I am not an expert to tell how it has influenced me that way. I do know that because of that background, I like to experiment with different things, but then I also know that many of my friends from India are completely disinterested in experimenting with anything. So, how did I become different from them? I don't know that. These are the complex problems.

UBIQUITY: Let's move from words to numbers. You've had some thoughts about statistics from baseball and other sports; share them with us, along with related thoughts you've had on the analysis of stock market performance.

JAIN: My basic point is that there is a big difference between information and insights. If I ask how your stock is doing, and you give me one year's worth of numbers, it would take me a long time to go through those numbers and understand what that stock has been doing. But if you show me a chart of stock prices over the last one-year period, it takes me just a second to understand how the stock is doing. And the same thing applies to a baseball player's performance: how the player is performing. The point is that just churning out data doesn't really help too much. The difference between the data and the context is what makes it useful. This is exactly the same as when we were initially talking about the results of Google and I was talking about the holistic picture and trying to put together along the time and the space. The data warehousing people realized this, too, and when they began dealing with large volumes of data using visualization techniques they started calling their work "business insights" or "business insight applications."

UBIQUITY: Insights seem so elusive. Take as a purported insight the anti-war slogan "War is not the answer." Without worrying about the truth or falsity of that slogan, how would any search mechanism deal with it?

JAIN: No search engine can deal with that, and in fact that's very interesting. Let me give you a very interesting example along the same lines. In playing with the various search engines, I tried to determine the popularity of George Bush in India. So I used many different keywords, but there was no way that I could find anything about the popularity of George Bush in India. All it would tell me or give me is articles that had the term "India" and "Bush" in them, and had some terms related to "popular." But it simply did not talk

about the popularity of George Bush in India. In the same way, if you write the query "War AND Answer" all you will get is that self-same slogan.

UBIQUITY: Have you looked at AskJeeves?

JAIN: Actually AskJeeves started going in the natural language direction, but they soon realized that natural language understanding is not ready for this. So they now have a combination of people sitting and analyzing and then trying to put the answer. It's a very interesting combination of natural language understanding with case-based reasoning and human-based organization.

UBIQUITY: What's the standing of this field today in academia?

JAIN: It is becoming very interesting and respectable in academia now, and people are paying attention to this. At one time, information retrieval was not as respected a field as it's slowly becoming. If you went to ten years ago, I don't think that computer academics and computer science departments would have considered this as respected a field as they would consider it now.

UBIQUITY: And that changed because of -- what?

JAIN: Two things. First it is widely used now so its applicability has been proven. And because it is now becoming a field where you can write a lot of papers. Faculty can get tenure, get promoted based on publications in this area.

UBIQUITY: Do you get back to India much?

JAIN: I go at least once or twice a year.

UBIQUITY: India is in the news a lot these days because of its great success in attracting work outsourced from US and European companies. What's your opinion of the political disagreements about outsourcing?

JAIN: Let me just say that when, for example, IBM Global Services decides to do a big project in India -- for the government of India or for Indian Airlines -- it is also outsourcing

from India to America. So outsourcing goes on not just from America to India. In fact the real debate should be, "What is the business-related policy necessary to accomplish this commercial exchange?" That's what the policy makers should be doing and that's what the real debate should be. The debate is not whether or not to ban outsourcing. It should be about facts, not emotions.

UBIQUITY: As we've discussed, much of your recent work has been focused on search engines. What has been most surprising about what you've found?

JAIN: One thing I've found particularly surprising is that you can find the same thing on a hundred different places on a single site in a response to the same query. I have no idea why they can't very quickly go through the results list and remove those duplicates. So that's certainly been a surprise to me. But there are lots of other puzzles as well. Yet that's the fun part of it -- to complain about all of this. I will be furious when those things are no longer there to complain about.

[See <http://jain.faculty.gatech.edu/>]

Source: Ubiquity, Volume 5, Issue 29, Sept. 15 - 21, 2004, <http://www.acm.org/ubiquity/>